Why Every Fab Should be Using AI

Jonathan L. Herlocker, PhD[1]

[1]*Tignis, Inc. jon@tignis.com. +1 (206) 531-0433*

Keywords: Artificial Intelligence, Machine Learning, Semiconductor Manufacturing, AI, ML

Abstract

This talk will provide compound semiconductor device maker professionals with a core understanding and a set of tools to help them effectively participate in and lead conversations within their organizations and with vendors around using artificial intelligence to achieve non-trivial business improvements in throughput, cost, agility, and resiliency.

Introduction

It seems that almost everywhere we look, we see claims of how AI is going to disrupt our world. If you take all those claims at face value, you might believe that *Artificial Intelligence* (AI) was a kind of magic, able to solve almost any of humanity's issues. Or you may remember that these claims have all happened before – we were told that AI was going to change the world, and then the reality did not live up to the claims.

My goal in this paper is to help filter out the marketing noise and surface the valuable information that you need to have as leaders of today's semiconductor manufacturing ecosystem. To provide you with the high-level understanding needed for you to start to have intelligent conversations about AI with your team, including specifics on how AI can and will deliver practical advances to your semiconductor manufacturing today.

Let's start with the basics. AI is not magic. Successful AI does require a level of creativity. I think it is useful to think of the successful usage of AI as part math, part science, and part innovation.

Most important, today's AI will move the business needle – it will increase your wafer throughput, it will reduce your costs, and it will reduce your inventory. It will increase your agility and your resiliency. McKinsey argues that AL/ML contributes $5-8B annually to earnings at semiconductor companies today and they expect that contribution to increase 10x over the next 2-3 years![1]

This time AI is not a passing fad. While previous waves of AI enthusiasm failed to yield broad market benefits, today's and tomorrow's AI will become a critical part of the compound semiconductor manufacturing world. Those that successfully adopt early will have competitive advantages. Those without an appreciation for AI will find their market value reduced.

AI vs ML vs DL vs LLM, Oh My

AI vs ML. There is often confusion as to the relationship between AI and *Machine Learning* (ML). As early as the 90s in Computer Science academia, not all AI researchers were ML researchers. Today, every AI researcher is also a ML researcher as well. Also, marketing has added noise to the conversation. I like to define AI as "computers doing stuff previously only doable by humans."

Within computer science, ML is a subset of AI which I define as 'computers learning to do useful stuff from data.' What belongs in AI but not in ML is a topic for computer science academics. *For semiconductor manufacturing, we can treat AI and ML as equivalent.*

ML is a toolbox, within which there are many tools, with new tools being added regularly. Fig. 1 shows a Venn diagram of the most popular ML tools.

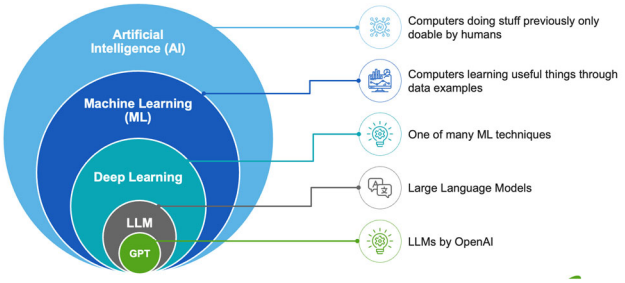

Fig. 1. AI Terminology Cheat Sheet.

What is Machine Learning Anyway?

Part of your job in the next decade will be deciding where and when to invest in ML. You probably will not have time to become an ML expert. In this section I will try to give you a high-level mental model that should help you participate meaningfully in discussions and decisions around ML.

Within the context of manufacturing, it is useful to think of ML as part modeling and part optimization.

Let's start with *modeling*. One way to think of ML is as function approximation. You are given a set of data that includes the results of a set of experiments with a mathematical function – the inputs that were provided to the function and the resulting outputs. ML can then reverse-

engineer a new function that provides outputs that closely match the experimental results. With this capability, we can now take physical or even human processes, map the inputs and outputs into mathematical notation, and use ML to reverse-engineer the function. The result: a model of the process. This model can then provide some transformative capabilities: such as predicting outcomes for new sets of inputs. We can model the physics in a process chamber, the lifespan of a part, the throughput of a fab, and much more. Human innovation is needed to map real world problems to mathematical representations. One way to identify places where ML could be useful is to think about where predictive models would be useful.

Optimization is also a key part of ML. When ML is 'learning' a model that mimics the process that created the provided data, that mimicry will not be exact. When a human learns, they are shown examples of how to do something, and they generalize from those examples. ML is no different. And just as two different humans can learn different things from the same life examples, the same is true for ML. There may be two different recipes for a process tool that achieves the same result. But one recipe may take less time and use less energy. The beauty of ML is that users can control what tradeoffs they would like to optimize based on the requirements for the device.

Let's go a level deeper. In the ML process, a model structure is first selected – in the simplest cases a linear model, in the more complex cases a deep neural network. Then the parameters of the model must be selected – the linear coefficients or the weights on the neural network. The 'optimal' parameters are the parameters that allow the model to generate results that most closely match the data. Part of the evolution of ML is that it has become exceptionally good at methods for efficiently finding approximately optimal parameters – in other words, optimization!

What does optimal even mean? If we are building a virtual metrology model to predict a critical dimension (CD) without having to measure it, we intuitively recognize that we want to minimize the magnitude of error between predicted CD, and actual CD. But consider a model that perfectly predicts most of the time, then occasionally makes errors of very large magnitude. That may have a low mean error. It may be preferable to have a model that makes very small errors all the time. And perhaps false positives are not as bad as false negatives. Thus, different kinds of errors may have a different weight of impact. In ML, we need another function that maps events to costs (or rewards). In optimization theory, this is called a loss function. Thus, one of the key innovation contributions is selecting the right loss function appropriate for the process that you are modeling.

I specifically call out optimization as a strength of ML because it helps to explain why ML can be so impactful across the semiconductor manufacturing ecosystem. Think of all the physical or logistical processes across your operations where a bit of optimization can go a long way. Decreasing the time wafers are waiting, increasing the time a tool is up, reducing the time it takes engineers to find the root cause of an equipment problem, etc. By applying ML, we can optimize almost every function of a semiconductor manufacturing organization.

AI/ML in Compound Semiconductor Manufacturing

Leading edge device makers like Intel, TSMC, and Samsung are well-advanced on the journey of deploying AI. Their revenue and market valuation have enabled them to justify billions of dollars of investment into custom-built AI solutions for their semiconductor manufacturing environments.

Just in the last couple of years, AI technology optimized for semiconductor manufacturing is becoming available on the market through specialized vendors. Now small to midsize device manufacturing facilities can acquire semiconductor-manufacturing optimized AI/ML solutions equivalent or superior to that used by the leading-edge vendors at a tiny fraction of the cost those solutions would have been five years ago.

Compound semiconductor manufacturing facilities have much to gain from AI/ML. There is non-trivial variability in CS manufacturing and AI software can significantly improve without having to spend capital on new hardware.

But Seriously, Didn't we Hear this all Before?

Skepticism is good science. But there really are some things that are different this time around.

1. Hardware advancements. High performance CPUs, cheap memory, low-cost storage, and GPUs.
2. Data availability. So many more sensors, combined with hardware means so much more data.
3. Algorithmic advancements in ML. Particularly around deep learning, transformers, and large language models.
4. Greater ecosystem. Particularly well-documented open-source software & models, and many data scientists trained in the art.

But don't believe all the hype. AI/ML is not magic, and certain areas of ML are currently overhyped. For example, Gartner lists Generative AI as being at the peak of their hype cycle.

Generative AI / Large Language Models (LLMs)

Generative AI is not entirely hype. Entire industries are right now being restructured due to generative AI – particularly those responsible for content generation such as marketing, sales, computer games, and art. Even software development. That same level of disruption is not going to happen immediately to semiconductor manufacturing, but there are immediate opportunities, and with the speed of

development of AI today, it will likely be less than ten years before semiconductor manufacturing is similarly disrupted.

To be fully armed for the conversation, let's define the key terms. Technically, Generative AI is a form of ML capable of generating useful text, images or other data using generative models. Colloquially it is used to encompass the set of recent ML innovations based on generative models with massive number of parameters, such as Large Language Models.

Large Language Models (LLMs) are the most relevant generative AI capability today to the semiconductor manufacturing space. LLMs are chatbots trained to predict (aka generate) the sequence of words that are the most likely response to a prompt in a chat. Because of the size of these models, they have been described as emergent intelligent properties. While most researchers agree that LLMs are not yet "general intelligence," they do currently provide some transformative capability that you should not ignore.

The key value that LLMs provide today is technology for a human-accessible unified query interface across a combination of unstructured and structured data. Today I can go to an LLM and ask it to "plan a tourist itinerary for Seville." All the necessary information to answer this request exists on the web. The LLM handles the interpretation of intent, identifies information relevant to the request, and handles the synthesis and summarization of potentially thousands of sources.

One of the biggest downsides of the current generation of LLMs is that, just like humans, they can make mistakes, but unlike most humans, they appear supremely confident even when they are wrong (although we can all probably remember working with a human that was similar – also not good).

A simple framework for understanding where LLMs will help you is to identify areas where you would like to greatly increase the automation of data retention, research and synthesis. Over the next year or two, think of a LLM chatbot as a tireless intern. You can ask them to synthesize an answer across a wide range of data sources. They will happily generate endless information for you, but their priorities and conclusions may be faulty without proper guidance. Be sure to implement proper checks and balances.

Turning AI into Business Outcomes in the Fab

Over the past decade there have been many documented instances where AI/ML projects failed to meet expectations and deliver meaningful value. I am making the argument that AI/ML is quite different today and is much more likely to succeed and meet expectations. However, avoid the pitfall of applying AI/ML to use cases that are easy to deploy but will not move the needle for your company.

One easy to appreciate formulation of business value in a manufacturing facility comes from Goldratt[2]. There are only three things that matter – Throughput, Cost, and Inventory.

1. Throughput – more sellable product out the door.
2. Cost – money spent to produce the product.
3. Inventory – the assets that are committed to the production of that product that could otherwise be redeployed.

Goldratt further argues that throughput is the most important value proposition, as improvements in throughput usually result in ripple-through benefits in cost and inventory in addition to other derivative benefits.

As you think about deploying AI/ML, it is also important to think about a large enough deployment so that you can see a movement in throughput or cost. The easiest way to achieve this goal is to tackle your bottlenecks.

In addition to the core business properties described above, I believe it is worth introducing two additional considerations: agility and resiliency. The two are very closely related. *Agility* is the ability of your business to adapt to changes more quickly. It could be market changes – like we are seeing with electric vehicles and Silicon Carbide and Gallium Nitride. It could also be customer changes. *Resiliency* is the ability of your business to survive or even thrive in the face of losses – losses of key personnel, equipment, customers, facilities. The modeling and optimization capabilities of ML will improve both your business agility and resiliency. Modeling will help you to understand the impacts of change. Optimization will enable your team to adjust your manufacturing more quickly and optimally to that change.

Specific Opportunities for AI/ML in the Fab

In this section I will provide a different lens that can be applied to your operations to identify opportunities for AI/ML.

Today, almost all implementations of AI in the fab can be reduced to one of Automation, Optimization, or Exception Management.

Automation is the reduction of human intervention in a process. Human intervention usually happens because of a need a decision or the need for human dexterity. AI/ML is a great enabler of increasing automation. Look for automation where you have humans spending non-trivial time on repetitive tasks, from tweaking recipes all the way to updating spreadsheets.

Optimization is about making the best possible decision given the available data. It becomes a true superpower when paired with automation, where your fab can continuously adjust operational parameters to remain optimal.

Exception Management is a core foundation of a more autonomous fab. Automation and optimization are about creating increasingly predictable processes. But disturbances and exceptions happen. Equipment breaks, software has bugs, operators do unexpected things, material sometimes does not meet specifications. As you automate more and more of your fab, how do you even know when something is wrong or

about to go wrong? ML is fantastic at combing tirelessly through terabytes of information looking for signals that suggest an exception has or will occur.

Here are some areas where AI is and will continue to add differentiated value to the fab.

1. **Robotics and computer vision.** Automating wafer moves and visual inspection. Already happened.
2. **Design**. The next big disruption will likely happen here. Today semiconductor devices are designed with a large amount of manual effort and human heuristics. In the near-term future, you will describe the specifications of your desired output product, and a chip will be automatically designed[3].
3. **Scheduling** – of product starts, of wafer moves. The ML algorithms for optimization, the data to support those mean that we can optimize many more complex loss functions faster and better. Infineon reduced critically delayed lots from 14% to 2% using AI[4].
4. **New product introduction.** Auto-characterization and tuning of recipes. Micron ramped its alpha1 chip 30% faster thanks to AI[5].
5. **Process control.** Faster and more automated root cause analysis, more sophisticated statistical process control, significantly improved advanced process control. At Tignis customers we have seen reductions in process variability from 15% to 80% through the application of AI and process chamber lifespans increased by up to 5%.
6. **Maintenance & supply chain.** What preventative maintenance should be done when, given the product mix that was deployed? Which suspect parts should be replaced during a maintenance event? When to order long-lead parts? Applying AI to maintenance will lead to significant decreases in cost and increases in availability[6].
7. **Exception Detection & Diagnosis.** FDC systems today have limited capabilities compared to the state of the art in computer science and ML. Scale of number of signals that can be monitored, raw traces vs summaries, the complexity of interrelationships that can be monitored, the length of history over which predictive models can be built, etc. Over the next decade, you can expect to replace your FDC systems with more advanced manufacturing intelligence systems that more holistically model not just your trace data, but your statistical process control data, your maintenance data, and your yield data. Not just identifying exceptions but diagnosing the root cause as well.
8. **Knowledge Capture and Transfer.** LLMs bring a huge new opportunity to effectively capture knowledge AND ALSO effectively transfer it to other key personnel in the fab, such as engineers and technicians. LLMs make stored knowledge more discoverable and accessible and LLMs chatbots are well aligned to guide technicians or engineers.

How these opportunities lead to Business Value

More automation means less human intervention – lower cost, reduced time, more agility/resiliency. Optimization of throughput and costs are a direct business value. Exception management ensures that the manufacturing process remains available (throughput) and minimizes exception cost. These are all direct revenue or cost benefits.

Perhaps even more important than improving the financials of a fully depreciated fab running an existing product is agility and resiliency (although harder to model financially). How quickly can you respond to changes in raw materials due to obsolescence? A change in product mix due to a market shift? A dramatic change in workforce availability? Investing in AI/ML will increase this agility and resiliency.

Conclusions

In closing, AI is not magic, it is part math, part innovation, and part science. AI is not a distraction. It will move the business needle. It will increase your wafer flow, it will reduce your costs, and it will reduce your inventory. It will also increase your fab's agility and resiliency. And this time, AI is not a passing fad. While previous waves of AI enthusiasm failed to yield broad market benefits, today's and tomorrow's AI will become a critical part of the compound semiconductor manufacturing world. Those that successfully adopt early will have competitive advantages. Those without an appreciation for AI will find their market value reduced.

References

[1] S. Göke et al. *Scaling AI in the sector that enables it: Lessons for semiconductor-device makers*. McKinsey & Company. Captured February 2024.

[2] E. Goldratt. *The Goal: A Process of Ongoing Improvement*. North River Press. 2014.

[3] A. Mirhoseini et al. *A graph placement methodology for fast chip design.* Nature. Vol. 594, pp. 207–212. 2021

[4] B. Waschneck et al. *Deep Reinforcement Learning for Semiconductor Production Scheduling*. ASMC 2023.

[5] S. Gatzemeier. *US Fabs of the Future*. Semi.org. International Strategy Symposium. 2023.

[6] R. Fletcher et al. *Need to boost semiconductor fab efficiency? Look to maintenance.* McKinsey and Company. Captured February 2024.

Acronyms

CD: Critical Dimension
FDC: Fault Detection & Classification
AI: Artificial Intelligence
LLM: Large Language Model
ML: Machine Learning