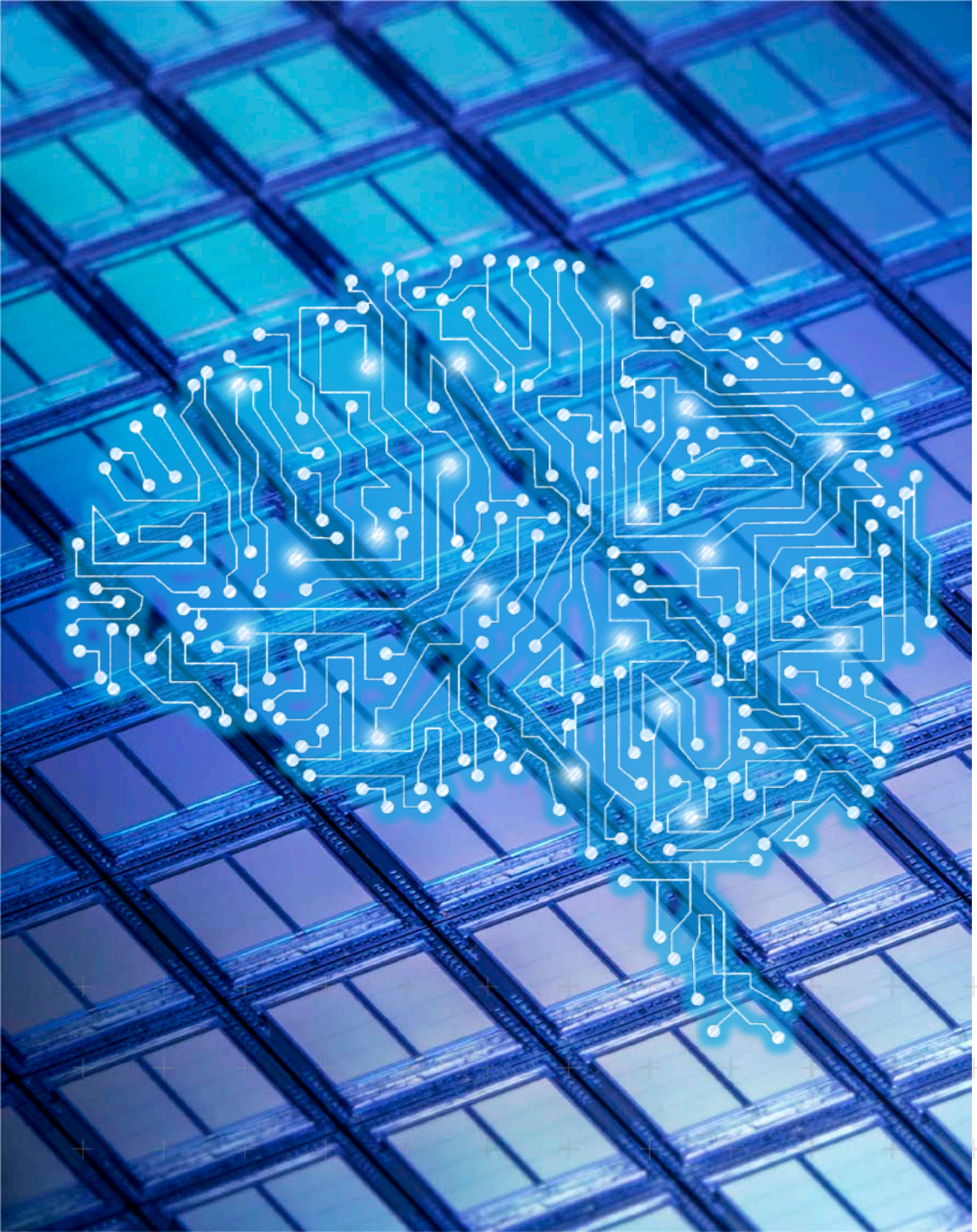


Challenges in Adopting ML in Manufacturing

Jacob George

Sr. AI Engineer

Analysis group

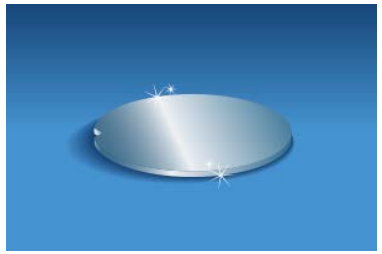
28 September 2021

Agenda

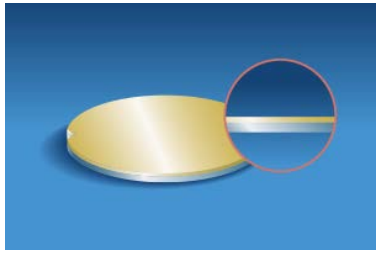
- Wafer Inspection
- AI/ML applications
- Challenges in Adopting AI/ML
  - Limitations from optical physics
  - Data challenges
  - Throughput challenges

Wafer Inspection

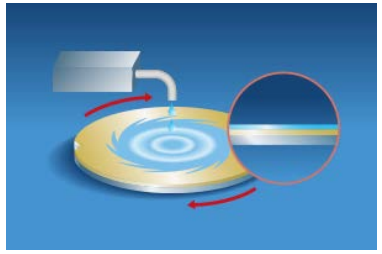
Semiconductor Manufacturing Process



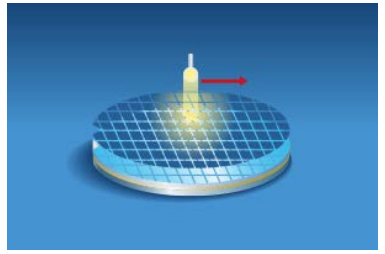
1. Cleaning



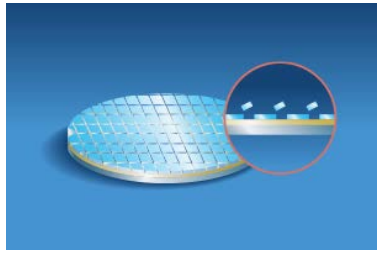
2. Film Deposition



3. Resist Coating

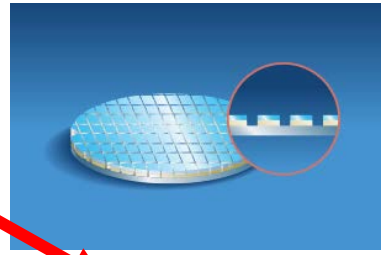


4. Exposure

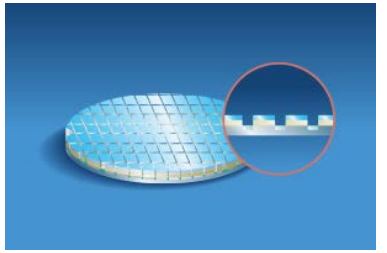


5. Development

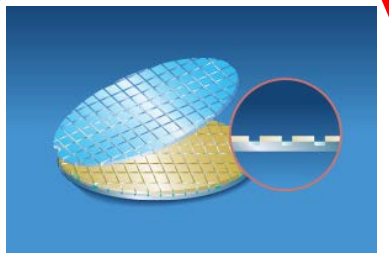
6. Etching



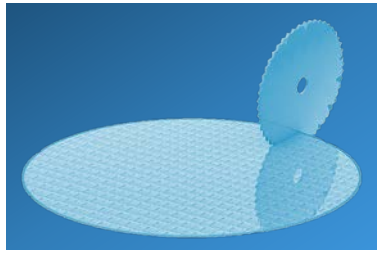
7. Implantation of Impurities



8. Resist Stripping



9. Dicing



9. Testing and Assembly

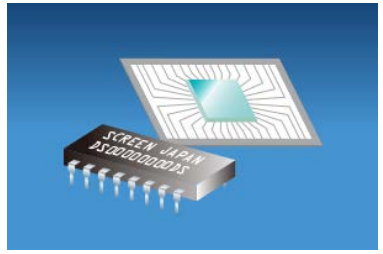
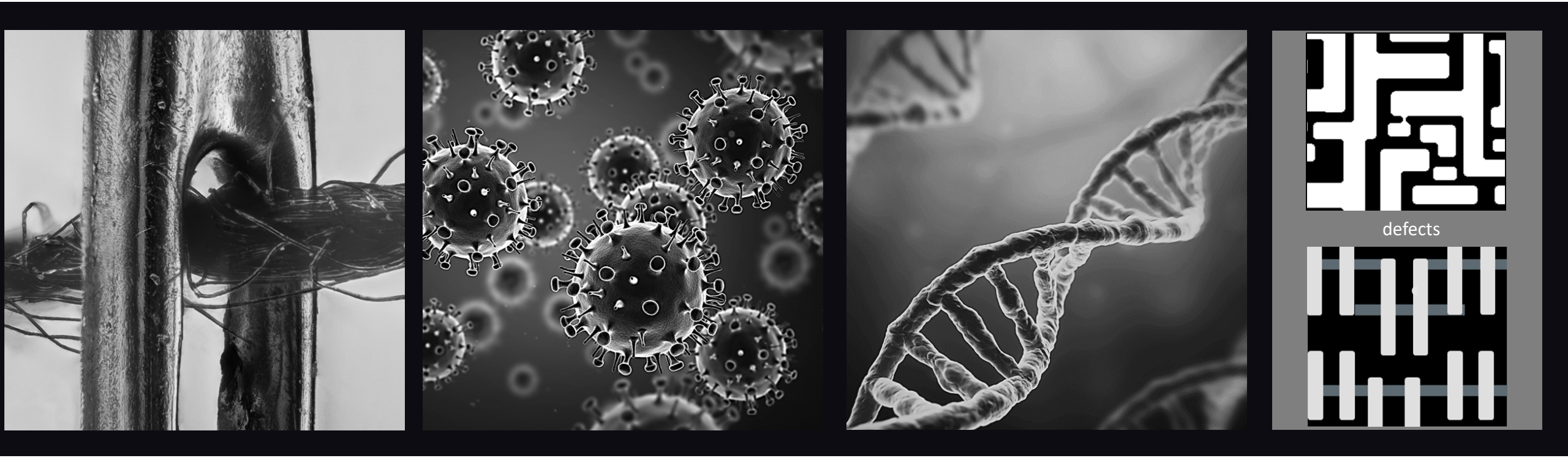


Image source: https://www.screen.co.jp/spe/technical/process

We Must Find and Classify Really Small Defects



Eye of Needle

2,000,000nm

Flu Virus

100nm

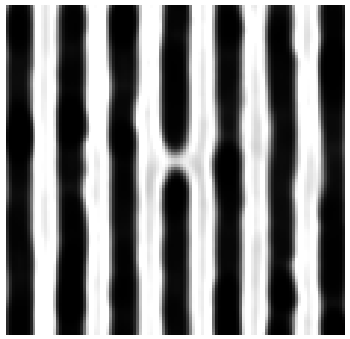
DNA Strand

2nm (width)

Semiconductor

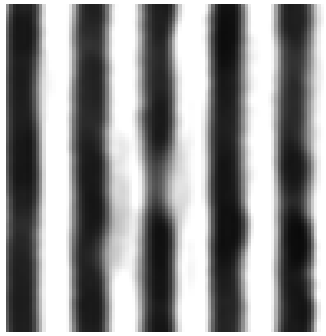
<10nm defect size

Defect Types

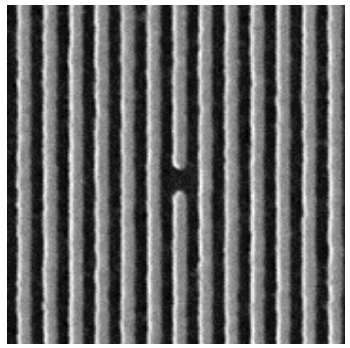


SPIE Advanced Lithography, 1161128, 2021

Bridge



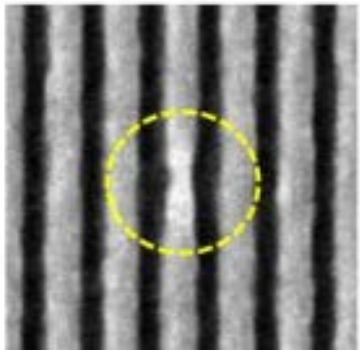
protrusion



SPIE Photomask Technology, 104510L

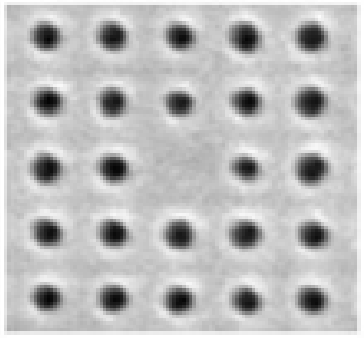
break

Shrink



SPIE Advanced Lithography, 113231J ,2020

Missing Hole



Proc. SPIE 10809, International Conference on Extreme Ultraviolet Lithography 2018

Electron-beam Review System

Scanning Electron Microscope (SEM) Technology

Broadband Plasma Optical Patterned Wafer Inspectors

DUV | UV | Vis

392x

295x

Optical-Based Inspection (photons)

Wavelength	: DUV - Visible
Optical Physics	: Diffraction of light
Throughput	: 1000x SEM Inspection

Broadband Plasma Optical Inspection

Electron Beam

Ebeam Wavelength: 2.5 pm

Defect size < 20 nm

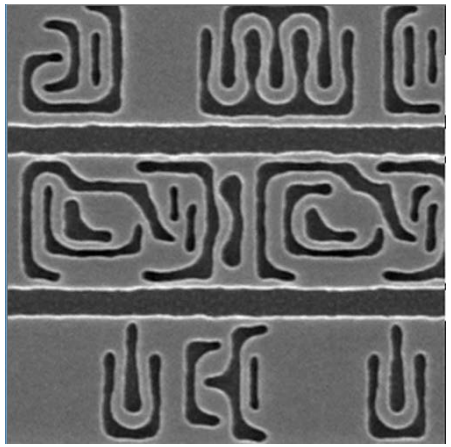


Image: KLA

Photons

Wavelength >> Defect size

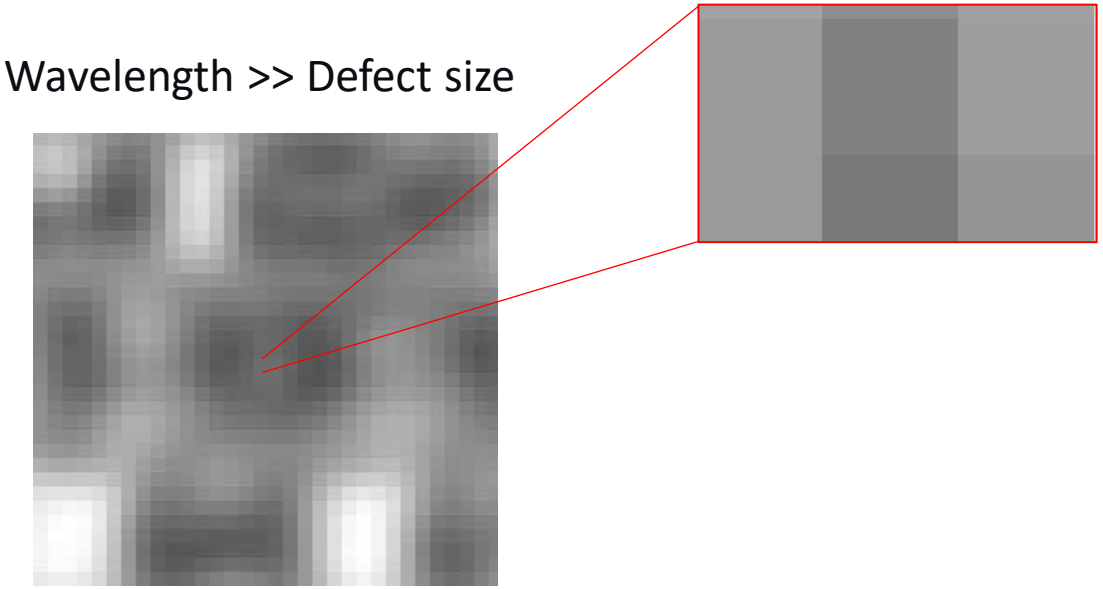
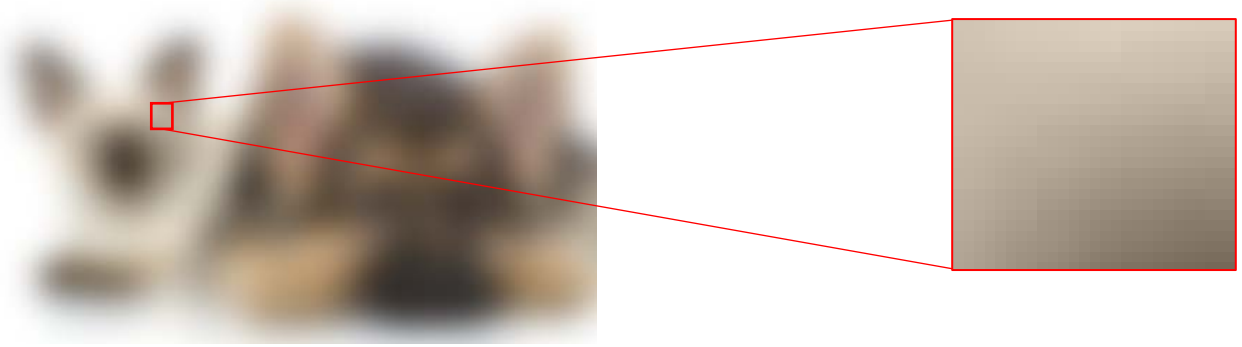


Image: KLA



Is the cat blind?

Cat or Dog?

Image Grabbing
- Sensors grab the images from wafer

Detection
- Report events
- Events include DOI and nuisance

Feature Extraction
- Calculate statistical attributes

Classification
- Suppress Nuisance
- Classify to DOI groups

Sensitivity

Capture more Defects of Interest at lower nuisance rate

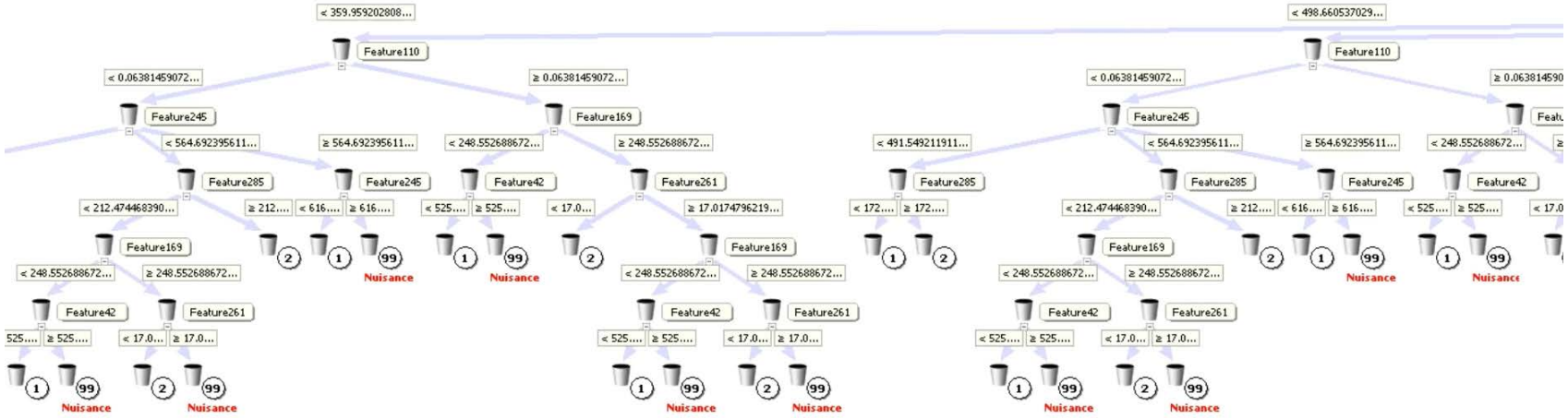
Throughput

Capture them all at the lower time rate

Its not only how good you classify, but how fast you can do that!

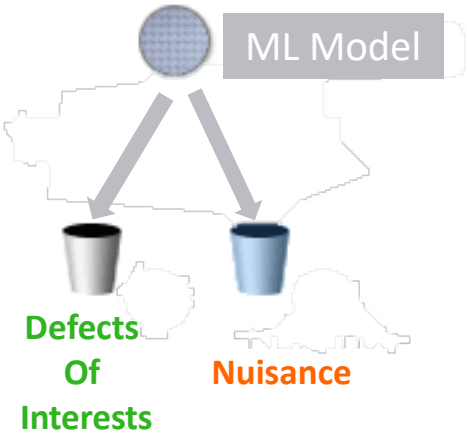
Classification prior to ML

Defect Attributes (several hundreds)



Classification with ML

Defect Attributes (several hundreds)



Challenges in Adopting AI/ML

Limitations from Optical Physics

Layer differences

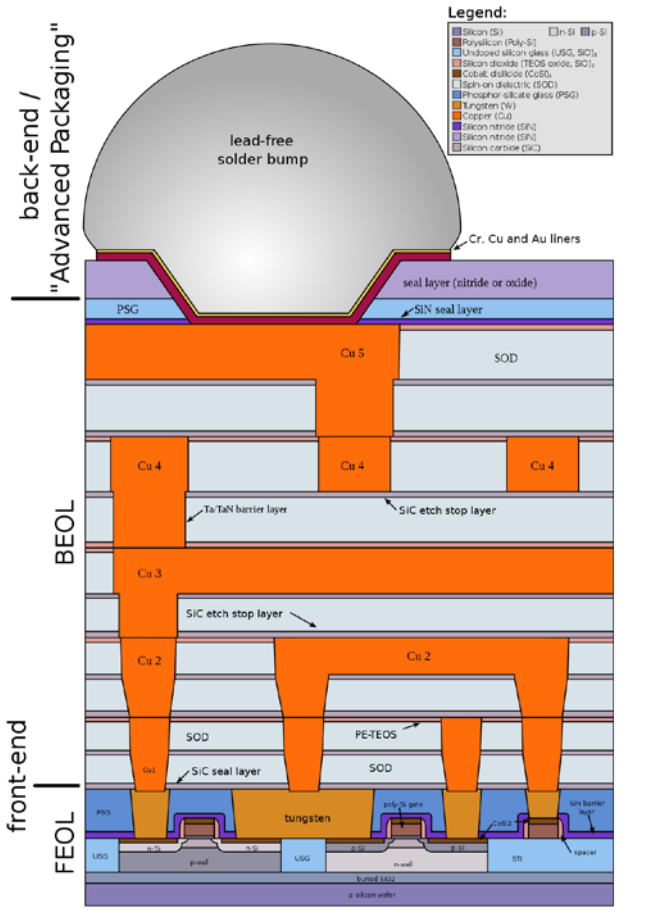
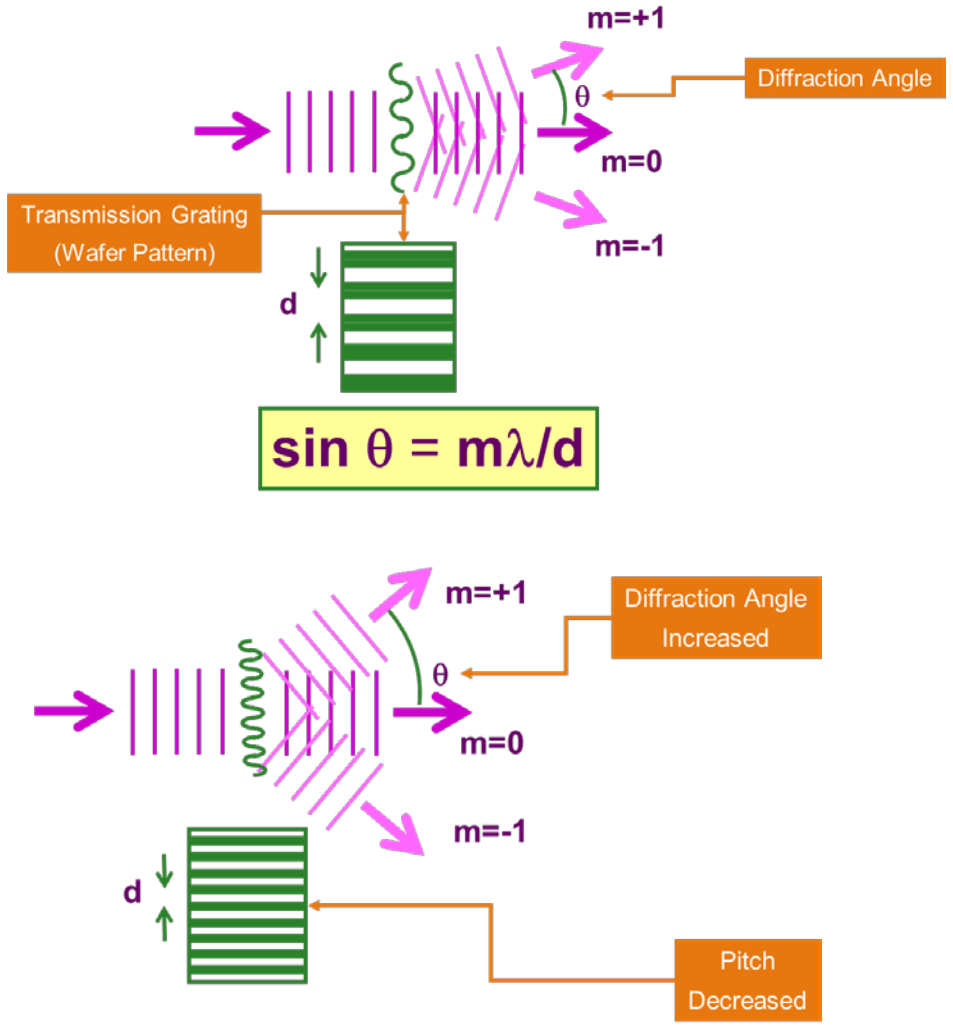
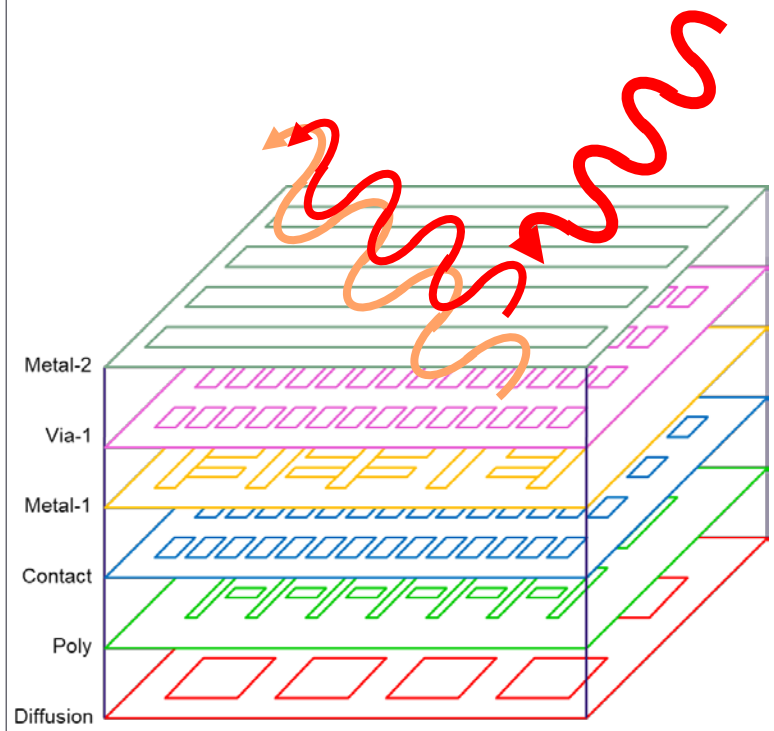


Image source: https://en.wikipedia.org/wiki/Front_end_of_line

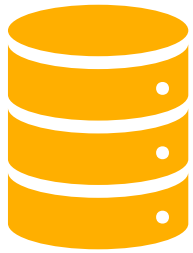
Design differences



Interferences from underneath layers



Data Challenges

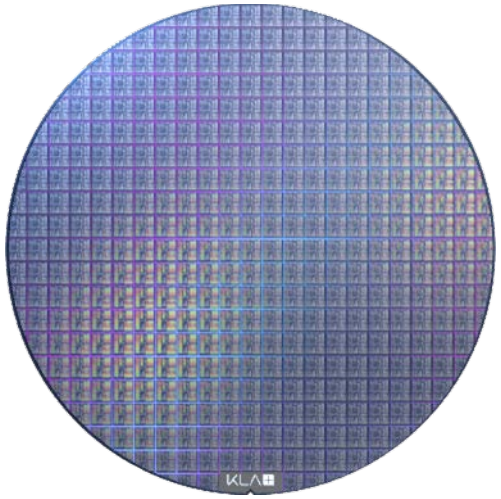


Obtaining Labelled Data

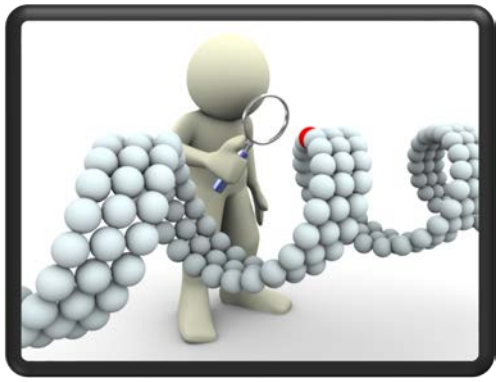
Process Variation

Model Maintenance

Obtaining Labelled Data



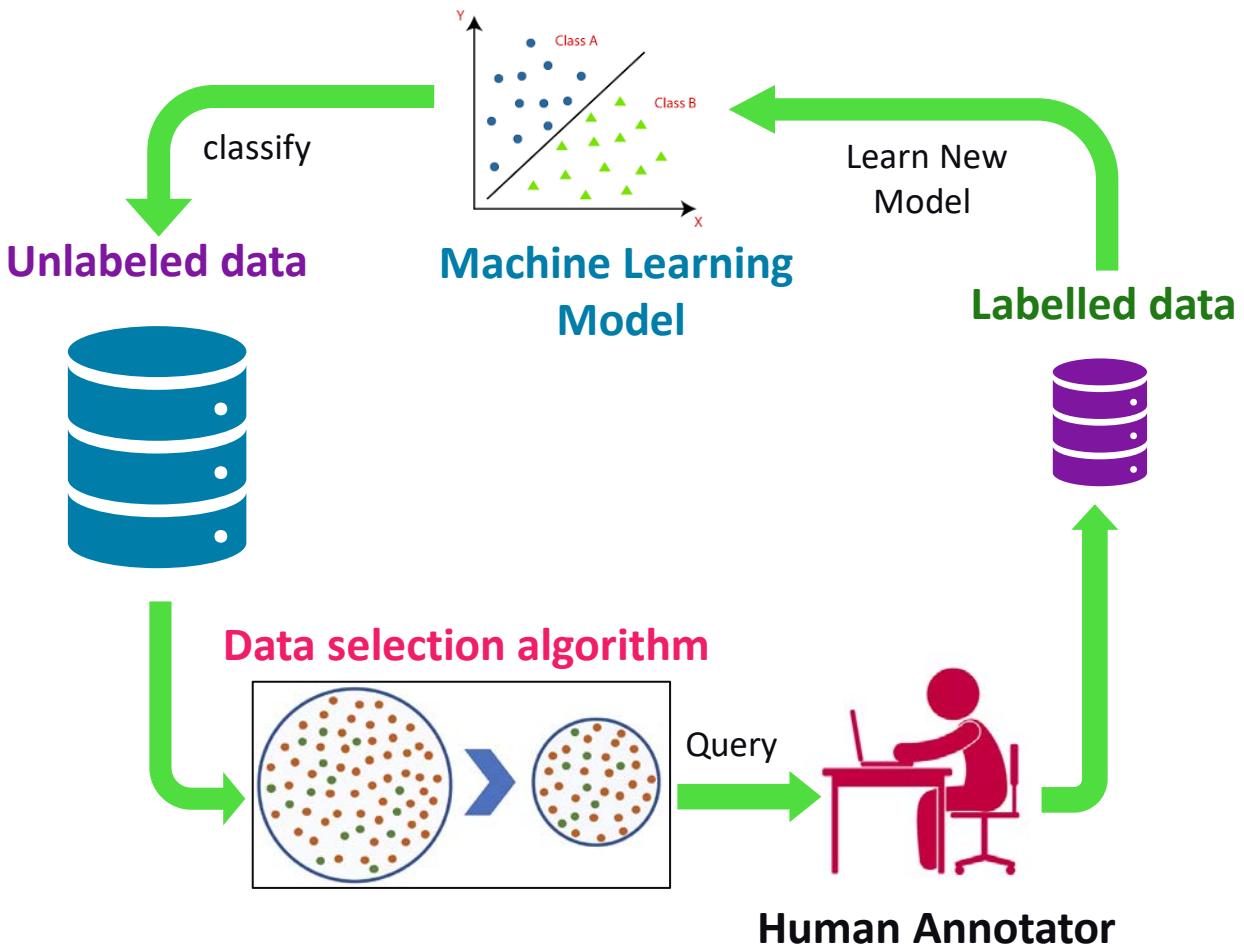
Obtaining 1st defects to train



Human Labour

SEM Review

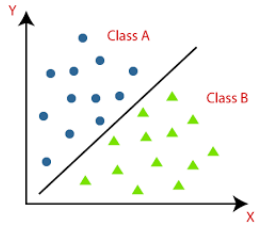
Active Learning: Learning from Limited Labels



Goal: Machine automatically and adaptively selects most informative data for labeling

Settles, Burr. "Active learning literature survey." (2009).

Active Learning: Learning from Limited Labels



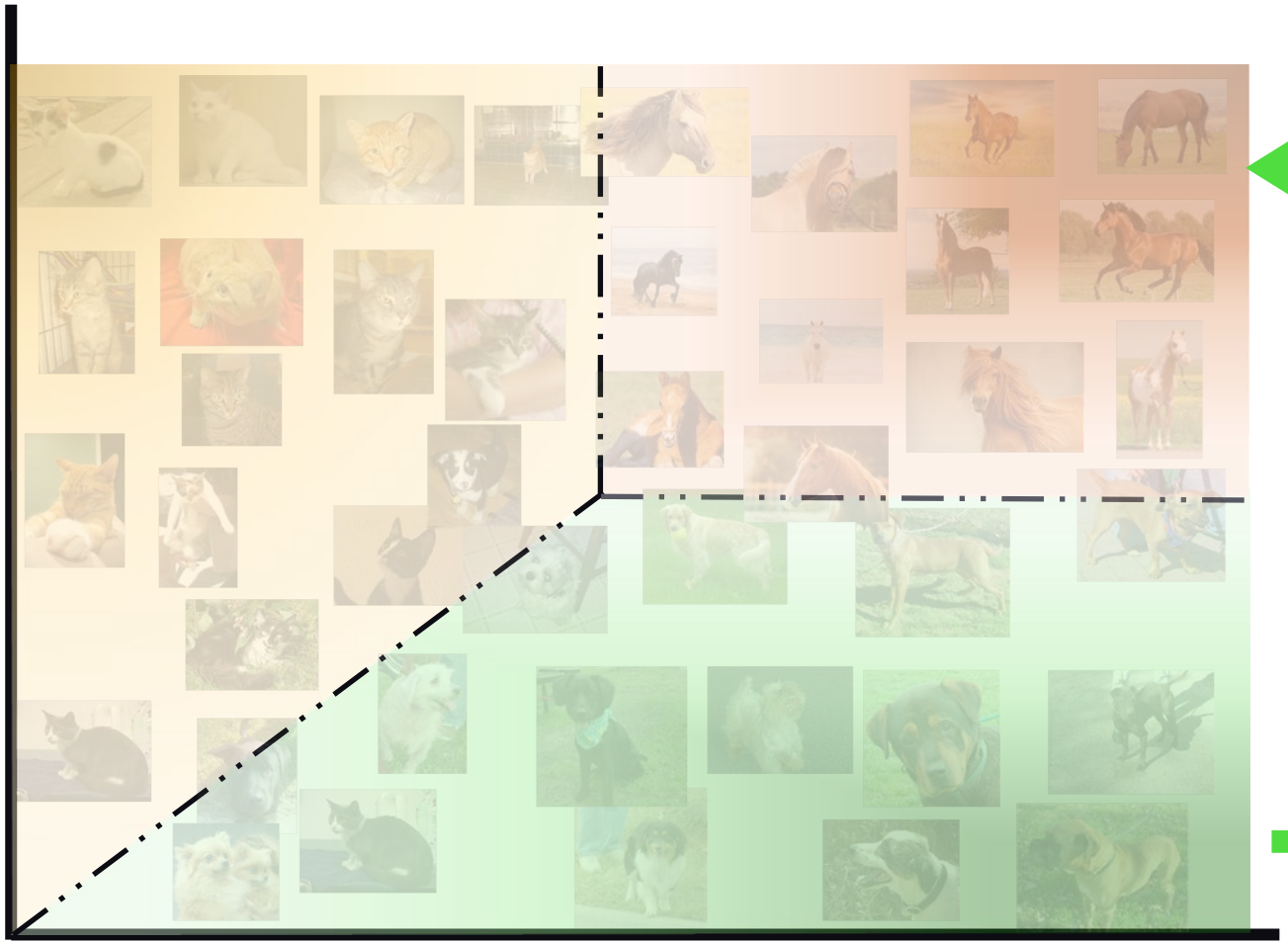
Learn New Model

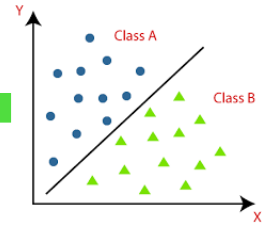
Labelled data

$5/image

Query

Active Learning: Learning from Limited Labels





classify

Learn New Model

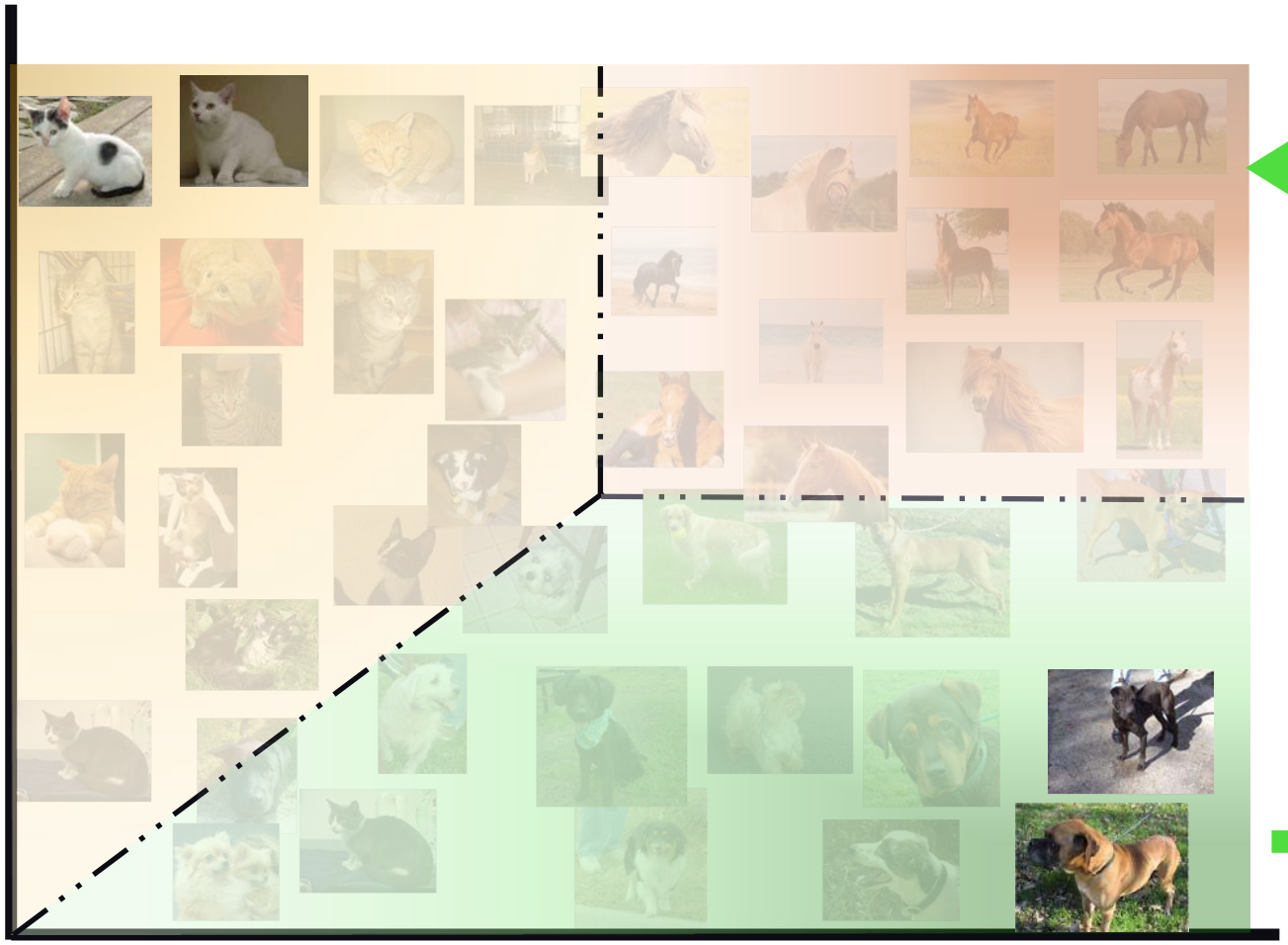
Labelled data

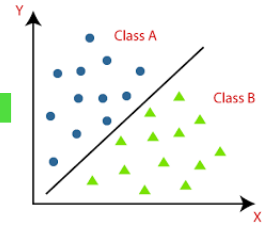
$5/image

Query

Human Annotator

Active Learning: Learning from Limited Labels





classify

Learn New Model

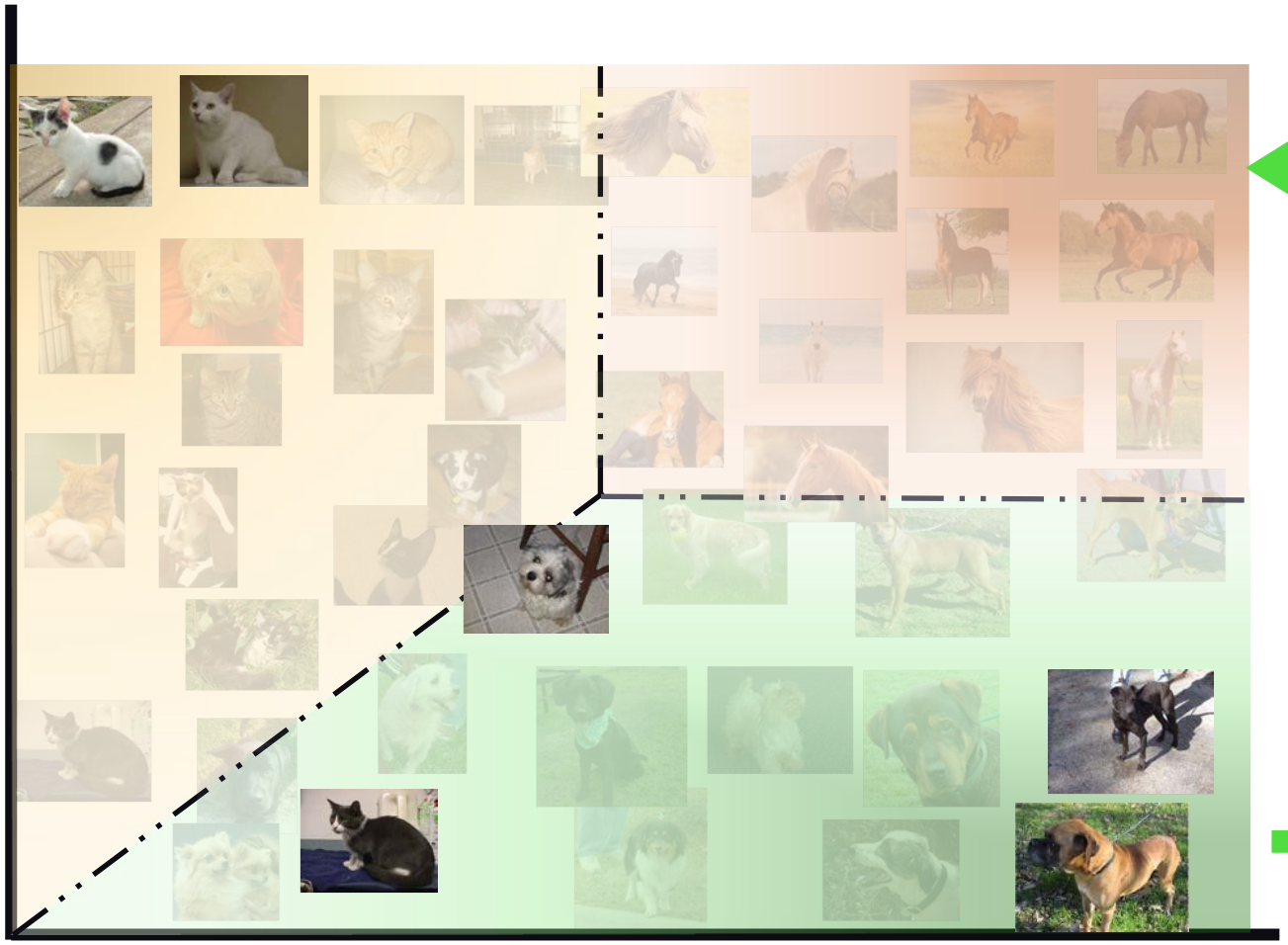
Labelled data

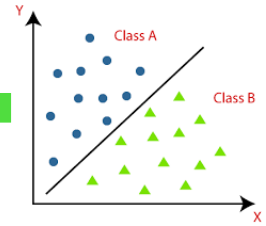
Iteration 0

$5/image

Query

Active Learning: Learning from Limited Labels





classify

Learn New Model

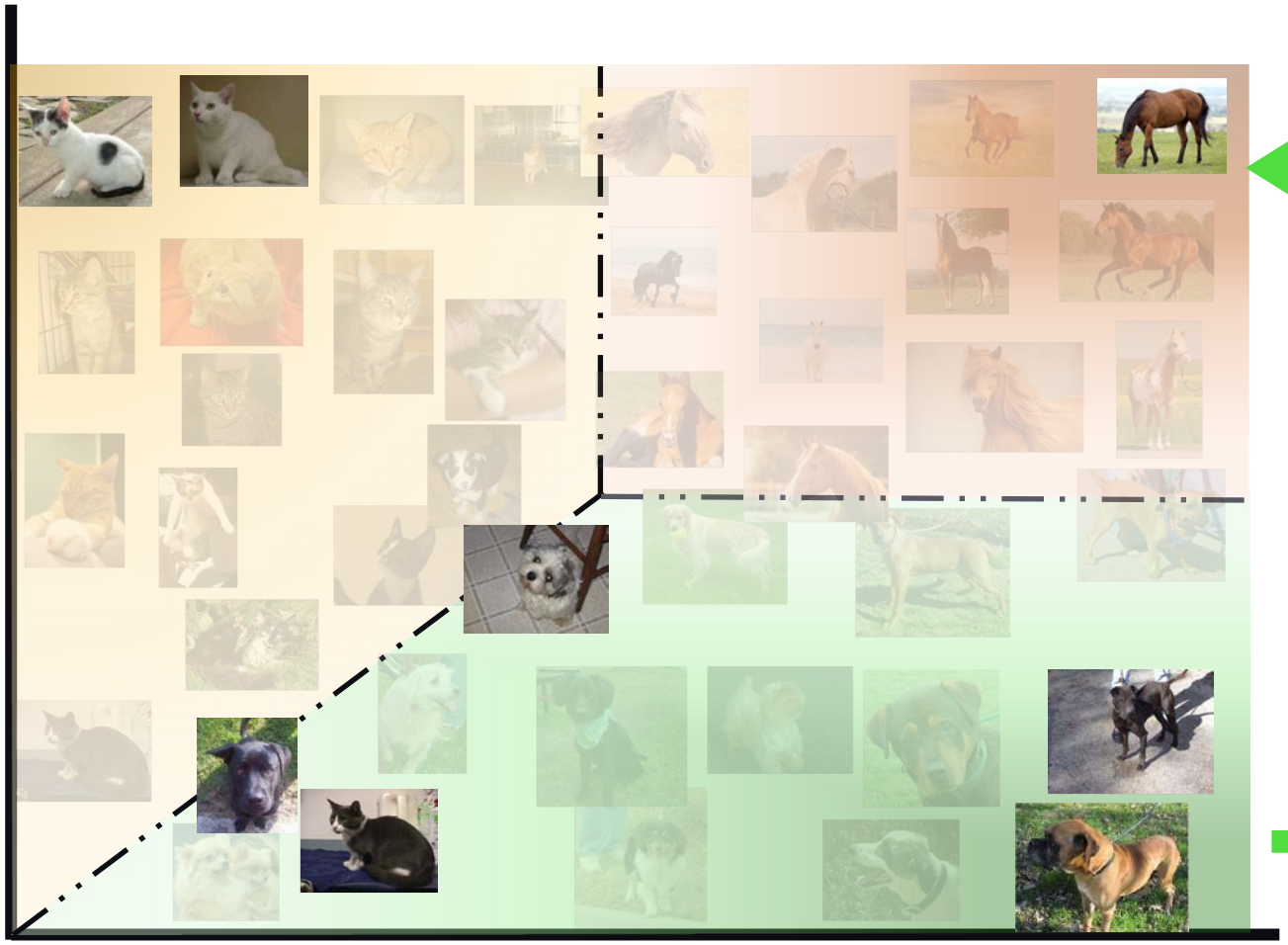
Labelled data

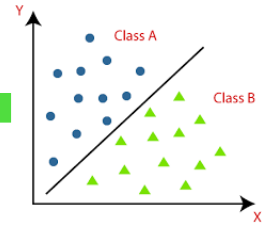
Iteration 1

$5/image

Query

Active Learning: Learning from Limited Labels





classify

Learn New Model

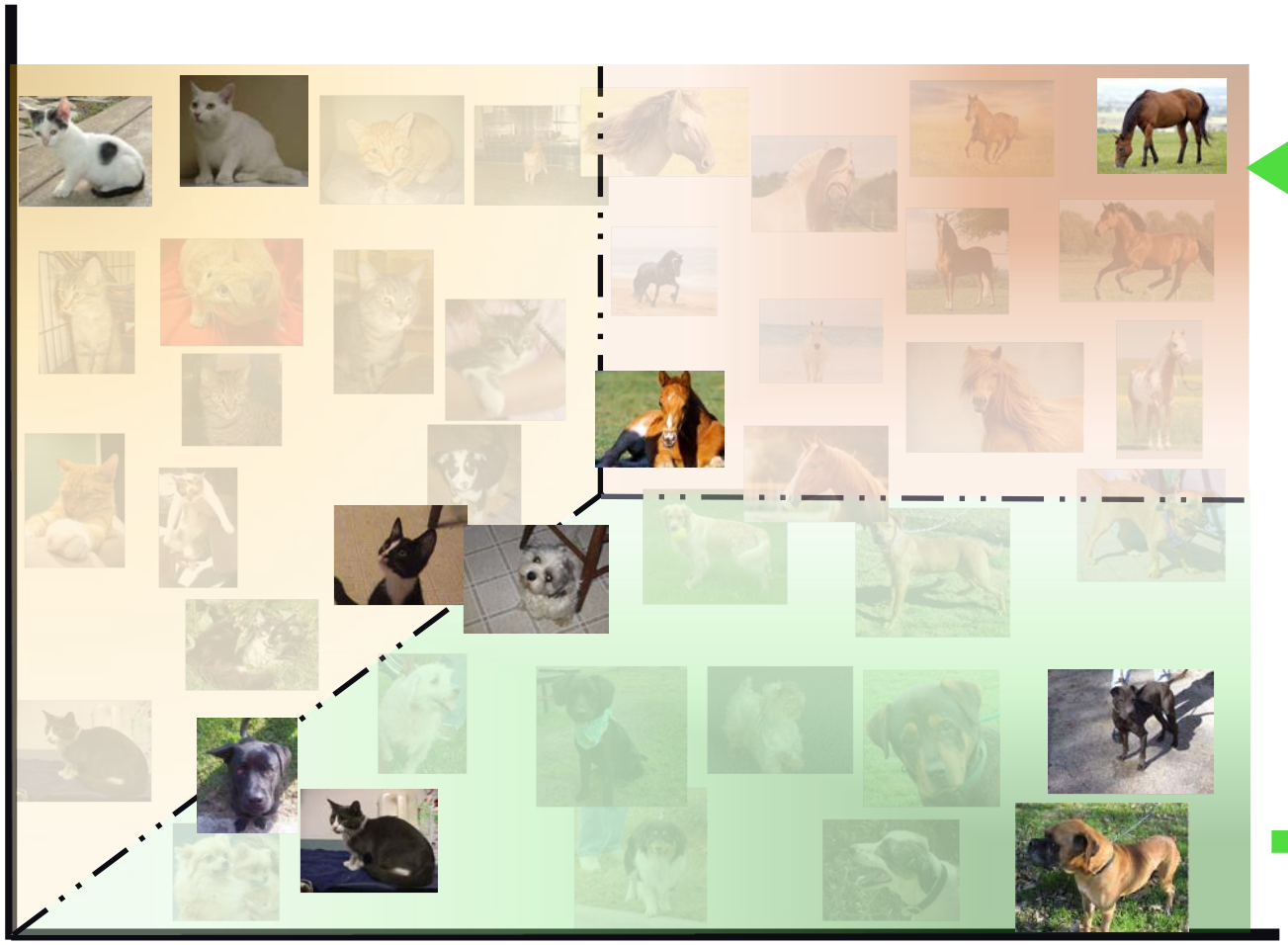
Labelled data

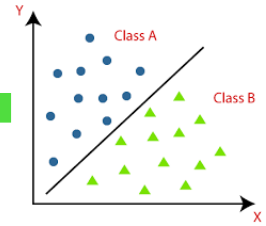
Iteration 2

$5/image

Query

Active Learning: Learning from Limited Labels





classify

Learn New Model

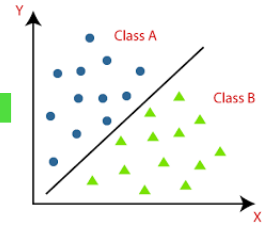
Labelled data

Iteration 3

$5/image

Query

Active Learning: Learning from limited labels



classify

Learn New Model

Labelled data

Iteration N

$5/image

Query

Active Learning: Learning from Limited Labels

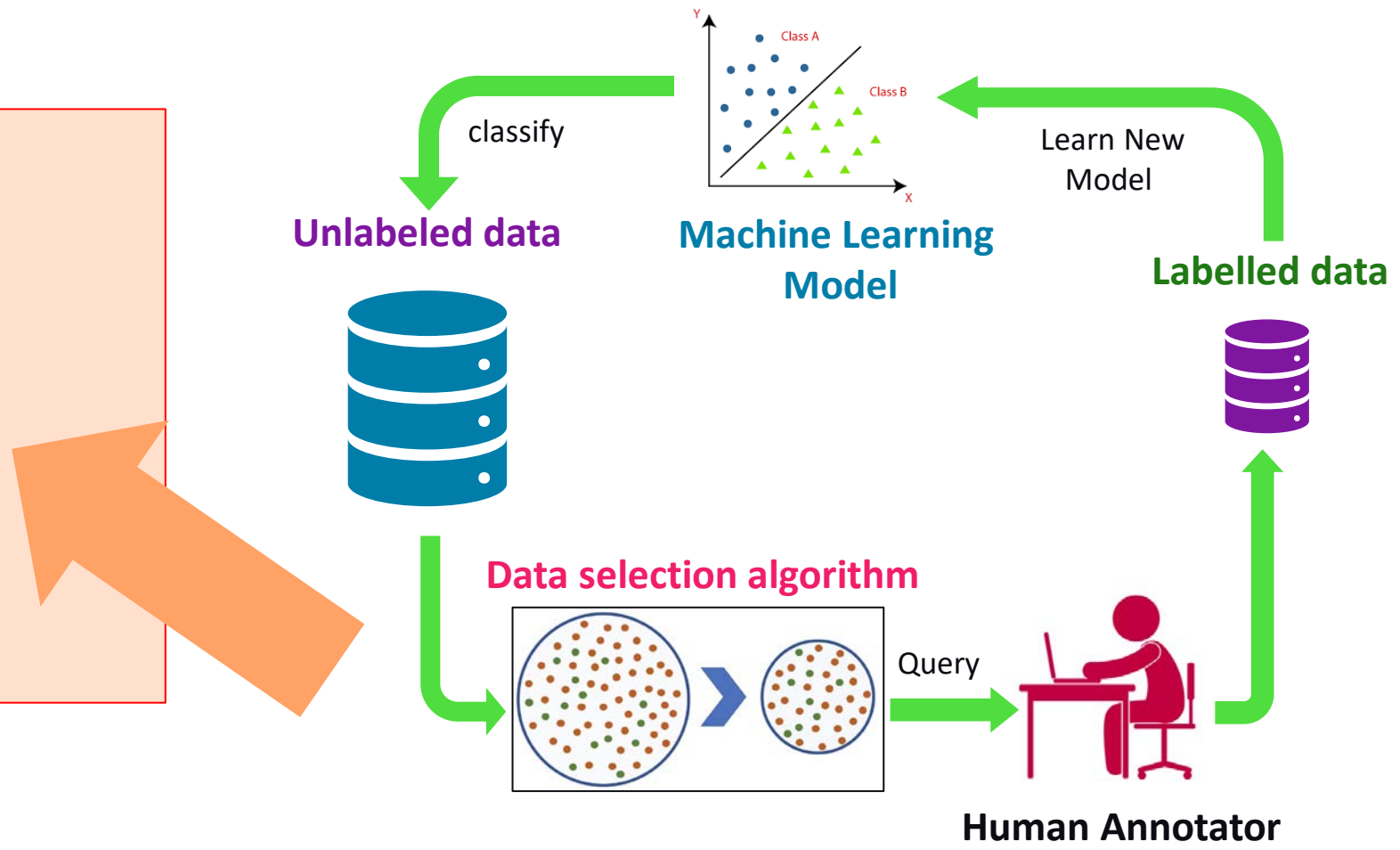
Uncertainty Sampling
- Least-Confident
- Margin
- Entropy

Query-By-Committee
- Measure level of disagreement
- E.g. vote entropy, KL divergence

Expected Model Change

Expected Error Change

Output Variance Reduction

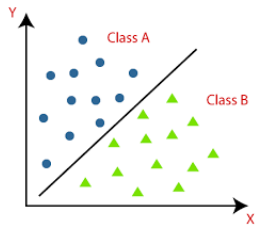


Active Learning: Learning from Limited Labels

BBP Optical Inspection System

High Productivity: wafer-level data

classify



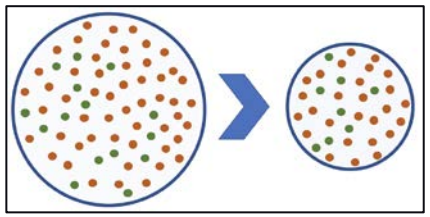
Machine Learning Model

Learn New Model

Unlabeled data

Labelled data

Data selection algorithm



Query

Human Annotator

High Resolution

e-beam Review System

Active Learning: Learning from limited labels

Better Sensitivity

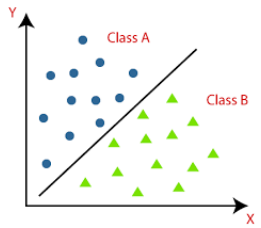
Production worthy ML model

classify

BBP Optical Inspection System

High Productivity: wafer-level data

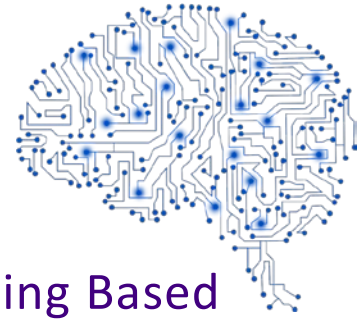
Unlabeled data



Machine Learning Model

Learn New Model

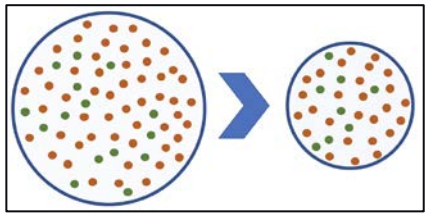
Labelled data



Deep Learning Based defect classification

High Resolution

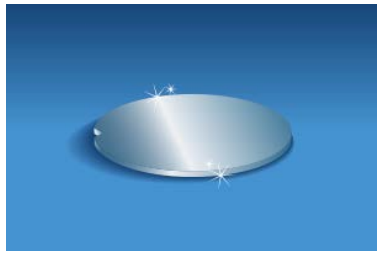
Data selection algorithm



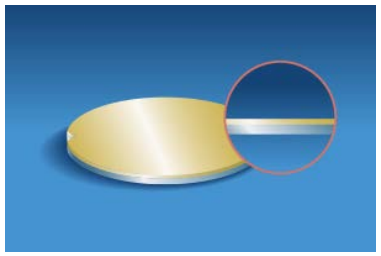
Query

e-beam Review System

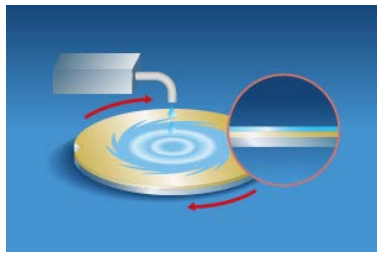
Process Variation



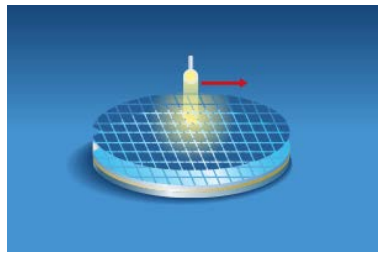
1. Cleaning



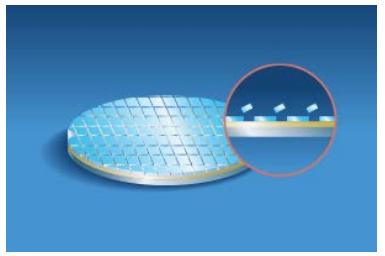
2. Film Deposition



3. Resist Coating



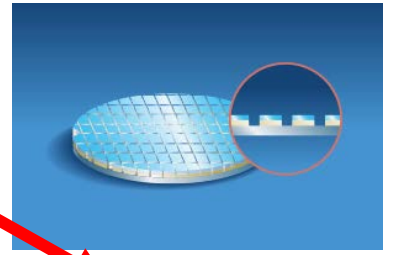
4. Exposure



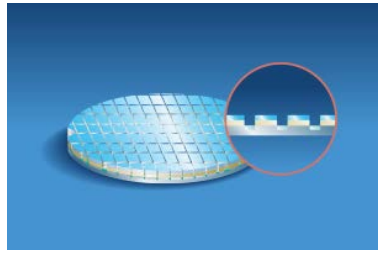
5. Development

- Material Changes
- Design variations
- Focus/ Exposure changes in Photolithography

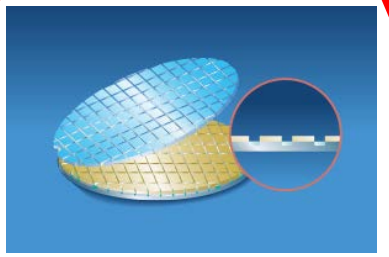
6. Etching



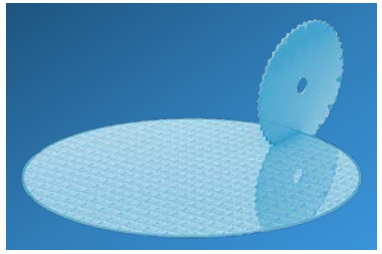
7. Implantation of Impurities



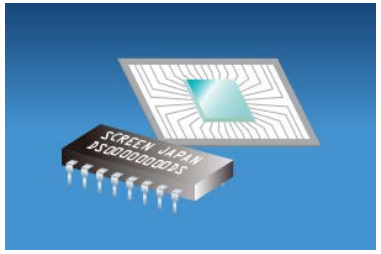
8. Resist Stripping



9. Dicing



9. Testing and Assembly



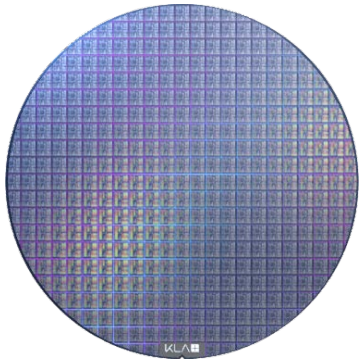


Image source: https://www.screen.co.jp/spe/technical/process

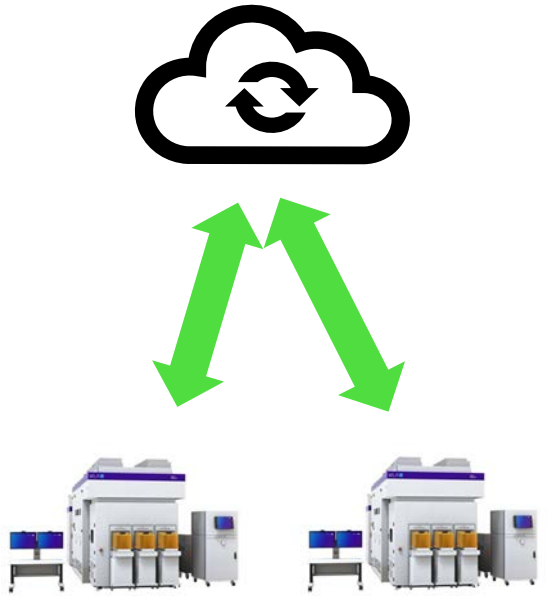
Adaptive Models

Models that periodically updates itself.

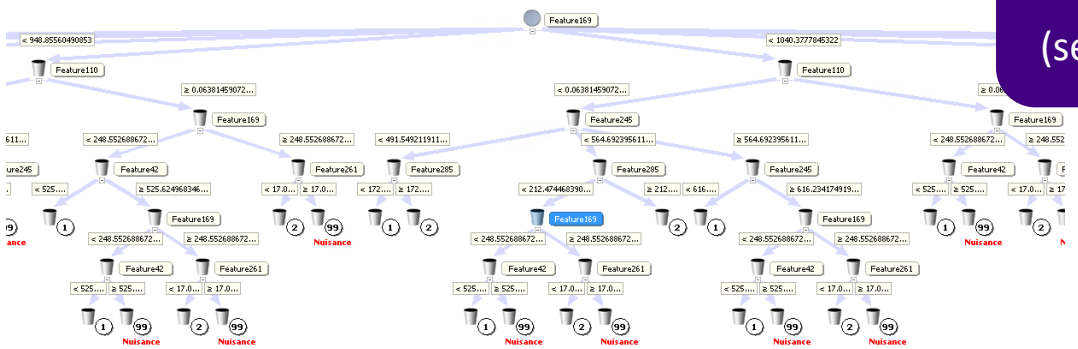
Tracking Model Performance

Deriving statistics to alert process drift

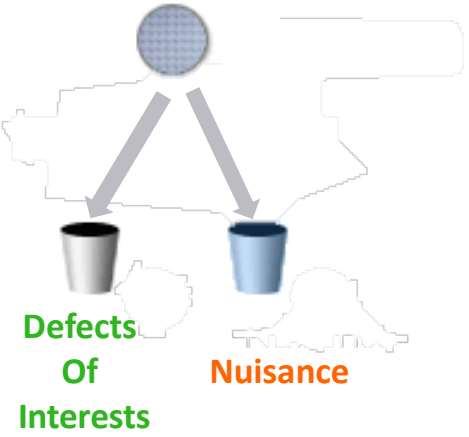
Software solutions and infrastructure to track model performance



Model Maintenance



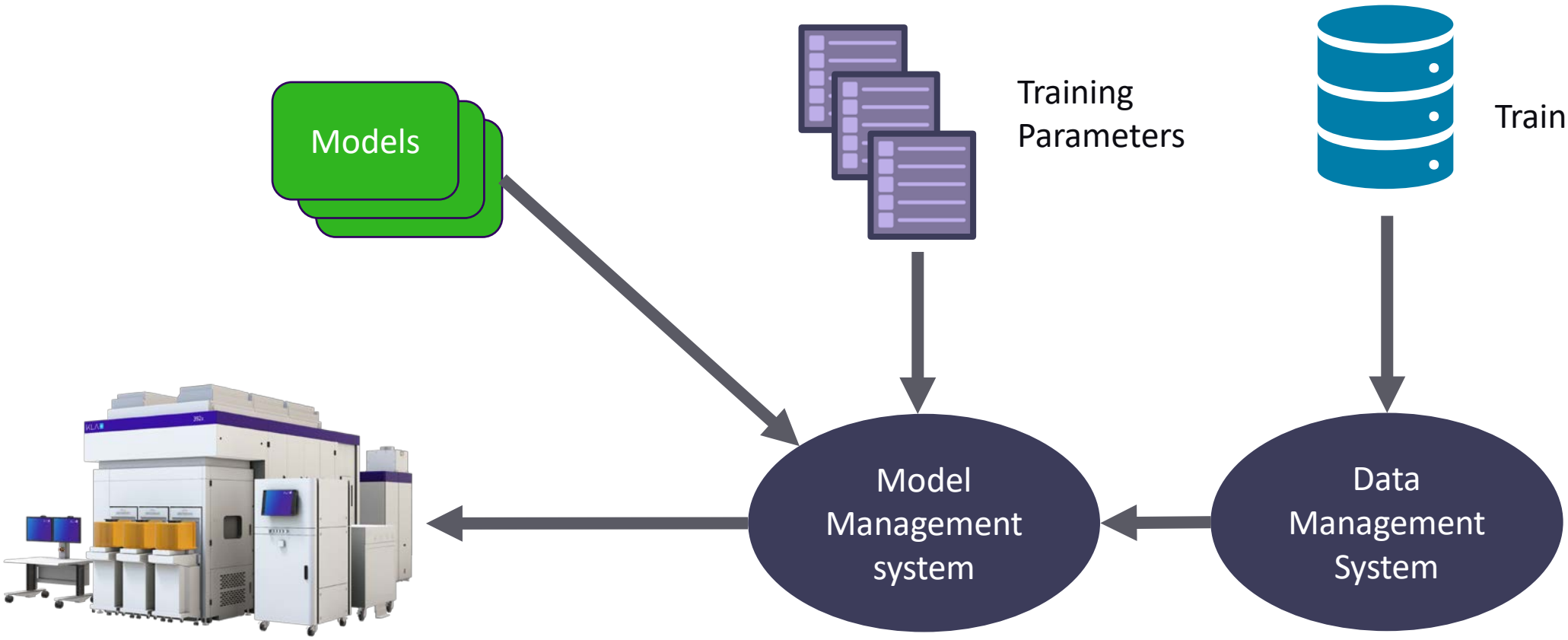
Defect Attributes (several hundreds)



- Interpretable
- Reproducible
- Easy to Maintain

- Reproducible: Need to store Training Set
- Bookkeeping required
- Need explainable ML

Model Maintenance



Need to have an ecosystem that would keep track of models

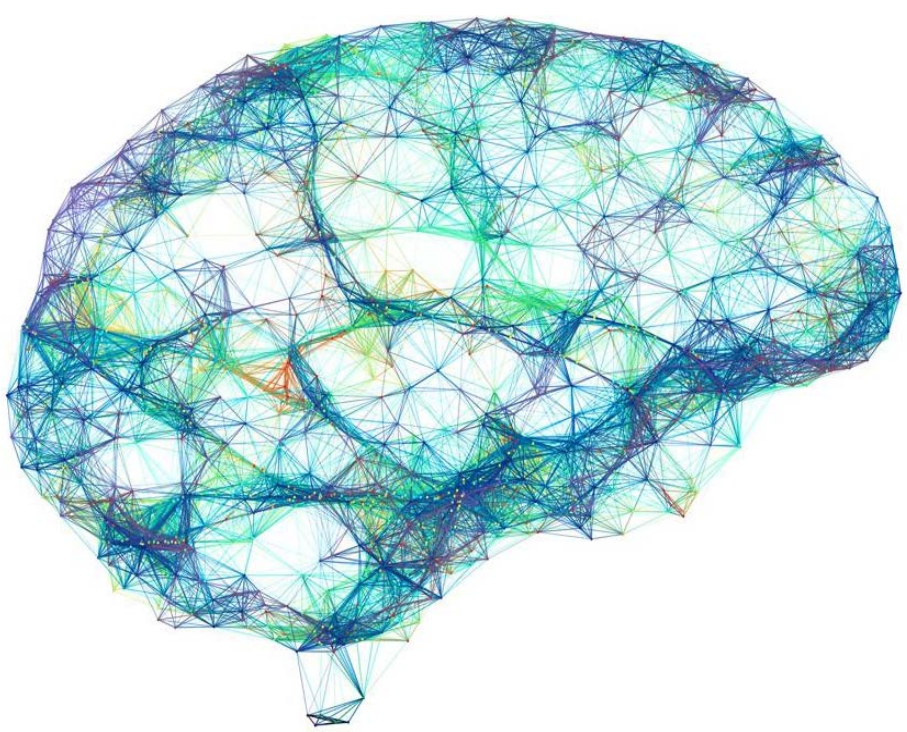
Throughput Challenges

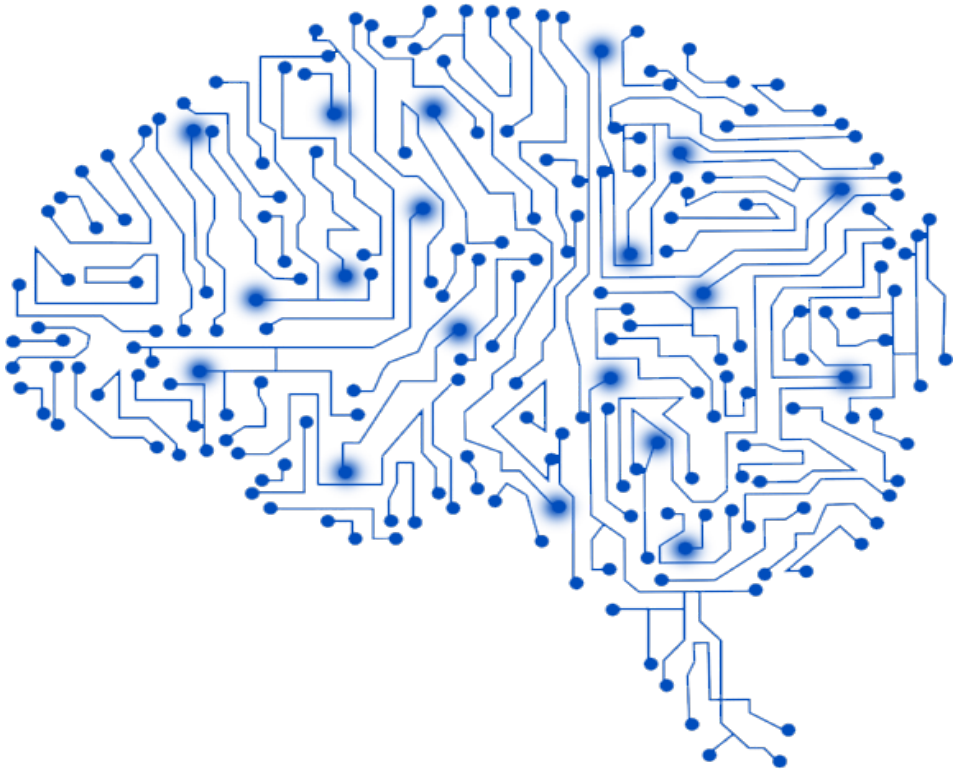
Achieving Throughput Expectation

Need to have cheaper compute

Optimize best on the hardware

Hardware Aware Algorithms





DL Inference

NVIDIA TensorRT
Programmable Inference Accelerator

cuDNN

Custom Implementations

NVIDIA CUDA

TENSOR CORES

Simple convolution (15x15)

Number of Images	Tensorflow Inference (seconds)	TensorRT (FP32) Inference(seconds)	CuDNN (Seconds)
1M	17.74	12.6	10.9

These frameworks are made to optimize large DL networks.

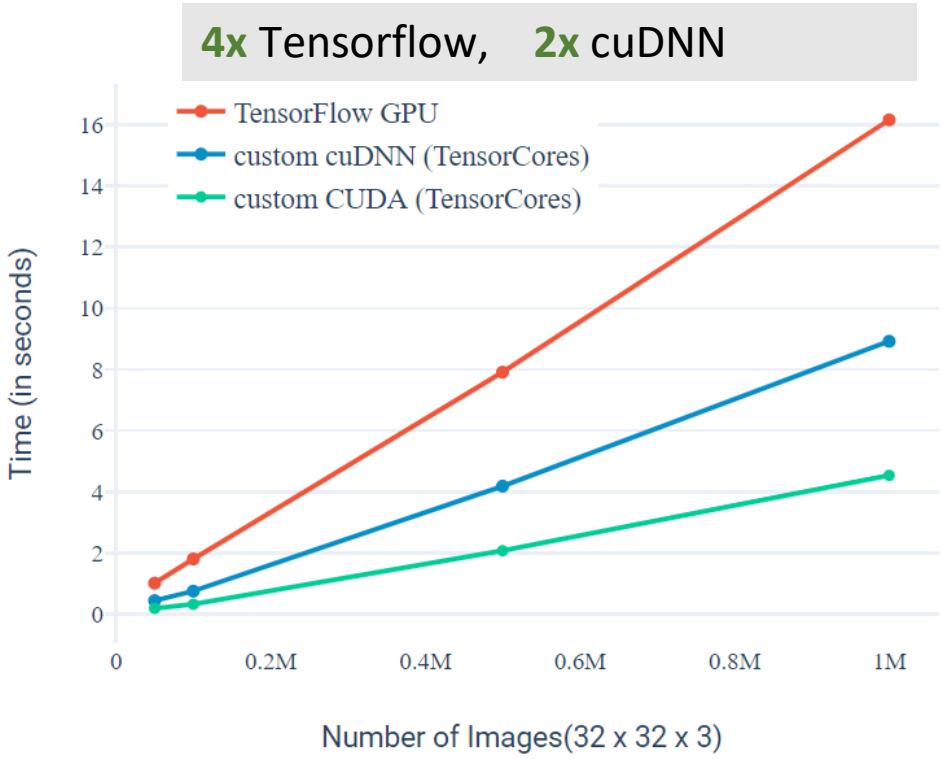
DL Inference

Can we create an optimized framework that could hyper boost DL inference tailored for inspection use case?

Layers	Parameters	Size	Cache	Max Capacity
Convolution 2D	32 x 5 x 5 x 3	4.6 KB	Constant Cache	8 KB per SM
Fully Connected	32 x 64	4 KB	Shared Memory	64 KB per Block
Fully Connected	64 x 2	0.2 KB	Shared Memory	64 KB per Block

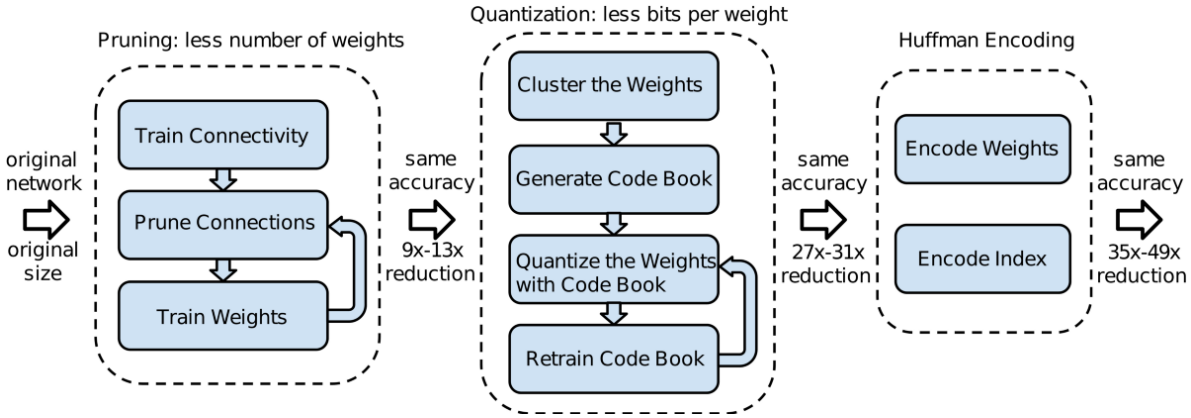
All Parameters are stored as float16 (2 Bytes)

No of Images	TensorFlow (in seconds)	Custom cuDNN (in seconds)	Custom CUDA using Tensor core (in seconds)
50k	1.02	0.45	0.20
100k	1.81	0.76	0.33
500k	7.91	4.19	2.08
1M	16.15	8.92	4.54



DL Inference

How can we train models that improves the inference speed?



	Baseline	Pruning (80%)	Post Training Quantization
	284kb	115kb	39kb (tflite)

No Accuracy degradation after quantization

Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding
Song Han, Huizi Mao, William J. Dally 2016

A model that fits into cache can be ninja optimized for inference

Need of the Hour…

- Expert in AI/ML
  - Explainable models
  - Few shot learning
  - Statistical ML
  - Model optimization for compute
- Embrace Modern C++
- Heterogenous Compute

Thank You