Device Manufacturing in an Era of Neural Networks

Dave Armstrong

Advantest Director of Business Development

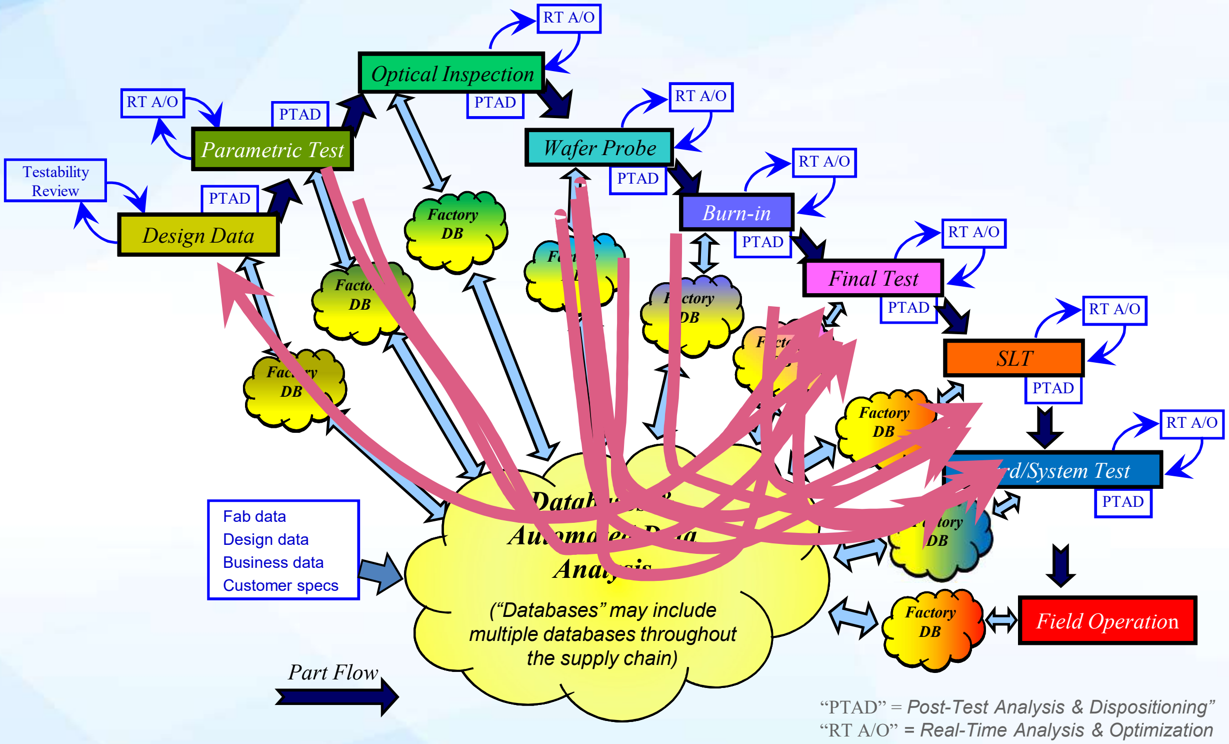
and

Heterogeneous Integration Roadmap, Chairman, Test Working Group

Agenda

1. What is Adaptive Test?
2. What are Convolutional Neural Networks?
3. How can these techniques be combined to improve the semiconductor manufacturing process?

Adaptive Test – The Vision



"PTAD" = *Post-Test Analysis & Dispositioning"*
"RT A/O" = *Real-Time Analysis & Optimization*

In theory….

Only a partial database exists today.

- Each test step can access DB to decide which tests would provide the most value for least cost.

In Practice….

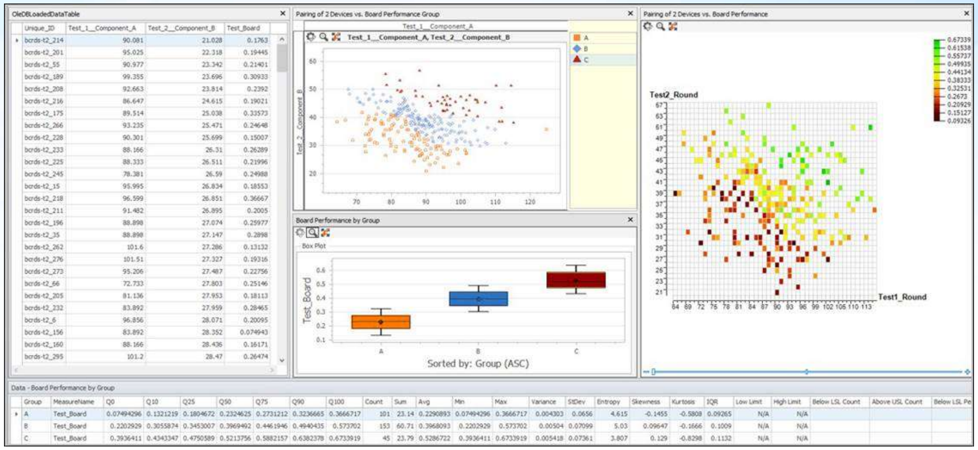
Wider data path broadly used today is Wafer Probe data being fed-forward to Final Test.

- All adaptive test today is feedforward in nature.
- Feedback data only helpful on future designs.

ADVANTEST

Adaptive Test – Today

- Data analysis today requires an expert data system and an expert data analysist.



	Unique_ID	Test_1__Component_A	Test_2__Component_B	Test_Board
▸	bcrds-t2_214	90.081	21.028	0.1763
	bcrds-t2_201	95.025	22.318	0.19445
	bcrds-t2_55	90.977	23.342	0.21401
	bcrds-t2_189	99.355	23.696	0.30933
	bcrds-t2_208	92.663	23.814	0.2392
	bcrds-t2_216	86.647	24.615	0.19021
	bcrds-t2_175	89.514	25.038	0.33573
	bcrds-t2_266	93.235	25.471	0.24648
	bcrds-t2_228	90.301	25.699	0.15007
	bcrds-t2_233	88.166	26.31	0.26289
	bcrds-t2_225	88.333	26.511	0.21996
	bcrds-t2_245	78.381	26.59	0.24988
	bcrds-t2_15	95.995	26.834	0.18553
	bcrds-t2_218	96.599	26.851	0.36667
	bcrds-t2_211	91.482	26.895	0.2005
	bcrds-t2_196	88.898	27.074	0.25977
	bcrds-t2_35	88.898	27.147	0.2898
	bcrds-t2_262	101.6	27.286	0.13132
	bcrds-t2_276	101.51	27.327	0.19316
	bcrds-t2_273	95.206	27.487	0.22756
	bcrds-t2_66	72.733	27.803	0.25146
	bcrds-t2_205	81.136	27.953	0.18113
	bcrds-t2_232	83.892	27.959	0.28465
	bcrds-t2_6	96.856	28.071	0.20095
	bcrds-t2_156	83.892	28.352	0.074943
	bcrds-t2_160	88.166	28.436	0.16171
	bcrds-t2_295	101.2	28.47	0.26474

Data - Board Performance by Group

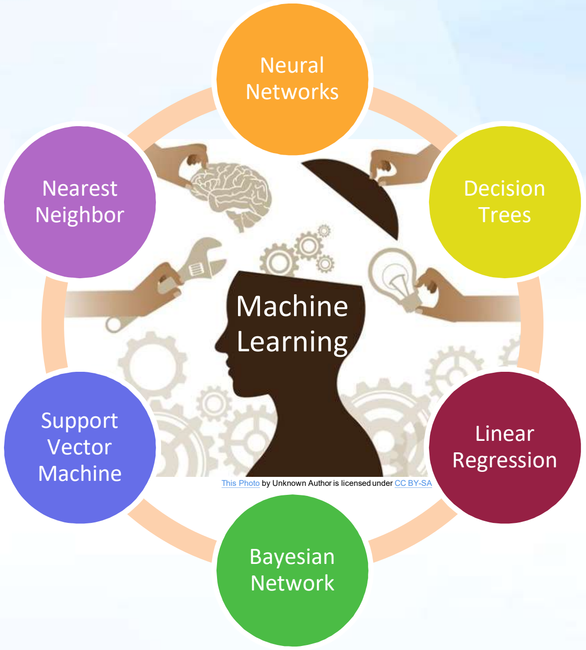
	Group	MeasureName	Q0	Q10	Q25	Q50	Q75	Q90	Q100	Count	Sum	Avg	Min	Max	Variance	StDev	Entropy	Skewness	Kurtosis	IQR	Low Limit	High Limit	Below LSL Count	Above USL Count	Below LSL Pe
▸	A	Test_Board	0.07494296	0.1321219	0.1804672	0.2324625	0.2731212	0.3236665	0.3666717	101	23.14	0.2290893	0.07494296	0.3666717	0.004303	0.0656	4.615	-0.1455	-0.5808	0.09265	N/A	N/A			
	B	Test_Board	0.2202929	0.3055874	0.3453007	0.3969492	0.4461946	0.4940435	0.573702	153	60.71	0.3968093	0.2202929	0.573702	0.00504	0.07099	5.03	0.09647	-0.1666	0.1009	N/A	N/A			
	C	Test_Board	0.3936411	0.4343347	0.4750589	0.5213756	0.5882157	0.6382378	0.6733919	45	23.79	0.5286722	0.3936411	0.6733919	0.005418	0.07361	3.807	0.129	-0.8298	0.1132	N/A	N/A			

TestVision 2020: “SQN Shared Quality Network” by Dan Sebban of OptimalPlus

In the future AI systems will do the analysis and the entire manufacturing flow will be tracked and optimized automatically.

The Machine Learning Toolbox

- Rich library of Machine Learning algorithms available
- Each algorithm is best for specific applications – no “one size fits all”
- Choosing the right algorithm is a mix of art and science – many factors must be considered
- Neural Network algorithms are particularly useful for recognizing patterns/relationships in complex data



Machine Learning

Neural Networks

Decision Trees

Linear Regression

Bayesian Network

Support Vector Machine

Nearest Neighbor

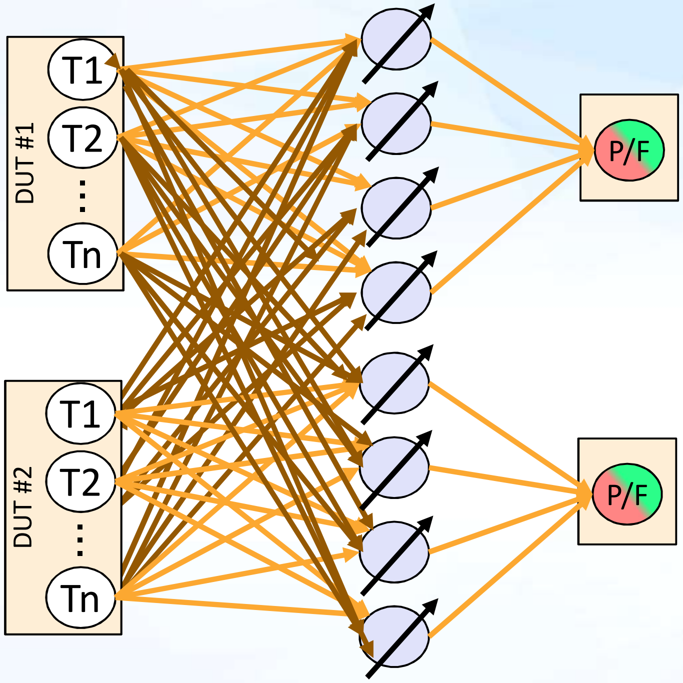
Types of Convolutional Neuromorphic Networks

... Determining the level of conductivity that is required for a neuromorphic implementation and then finding the appropriate hardware that can accommodate the level of connectivity is often a non-trivial exercise. **

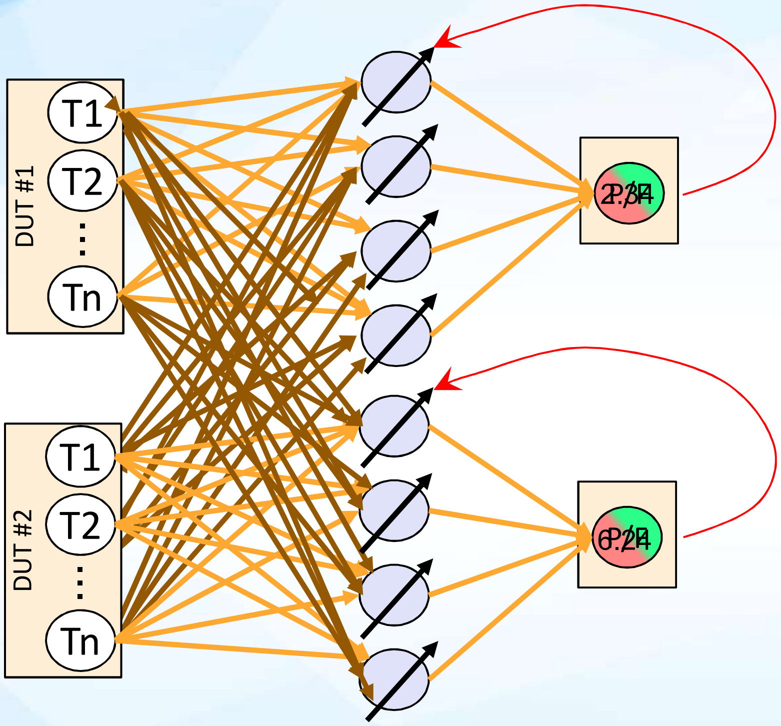
** A Survey of Neuromorphic Computing and Neural Networks in Hardware; Catherine Schuman et al; arXiv"1705/006963v1

What is a Convolutional Neural Network?

- A neural network takes test data, remaps it through one (or more) computational layers in order to determine pass/fail.
- Of course we have data for many devices per wafer/lot.
- Which gives us possibility for many more relationships.
- The convolutional step learns the feedforward coefficients over time. This is called "training".



Training



- Training is going to be the hardest part of setting up a Neural Network for semiconductor devices.
- What constitutes a “good” test result?
  - One with the most margin to spec?
  - One which help achieve the best yield?
  - One which meets spec with the minimum power consumption?
- Neural Networks today are typically trained with subjective (human) classifications.
- Once we have automatic loop training is we will truly have an artificially intelligent manufacturing environment.

Neural Networks are Good at Picture Recognition

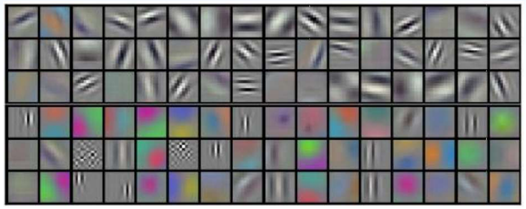
- Training requires a lot of data (something the semiconductor industry is not short of).

Input a bunch of Chihuahuas

Human Interpretation

Machine Learns to recognize Chihuahua patterns

MODEL = Set of Coefficients



“hmm, ok I learned what Chihuahuas look like”

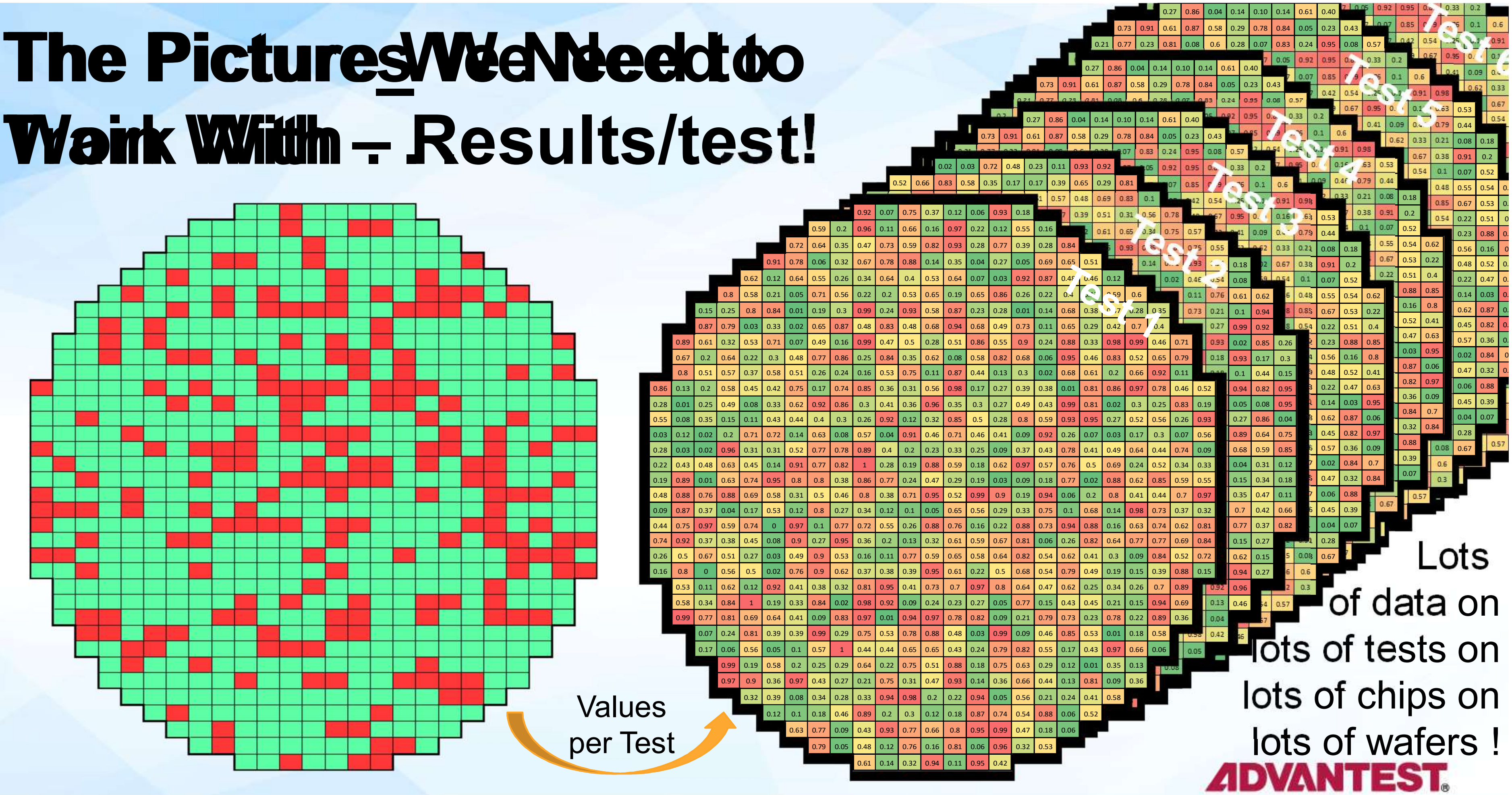
- Pointed ears
- Small typically dark nose
- Little beady eyes
- …

ADVANTEST

Training a Self-Driving Car

- Since the beginning, autonomous vehicle programs have analyzed videos of ~ 1 million traffic miles.
- But, statistics suggest ~ 770 accidents per billion miles driven. (0.8 per million miles!)
- Training these networks with real-world is both expensive, deadly, and slow.
- The industry is moving to simulated driving situations to get the vast quantity of data that’s needed to adequately train.

The Pictures We Need to Work With – Results/test!

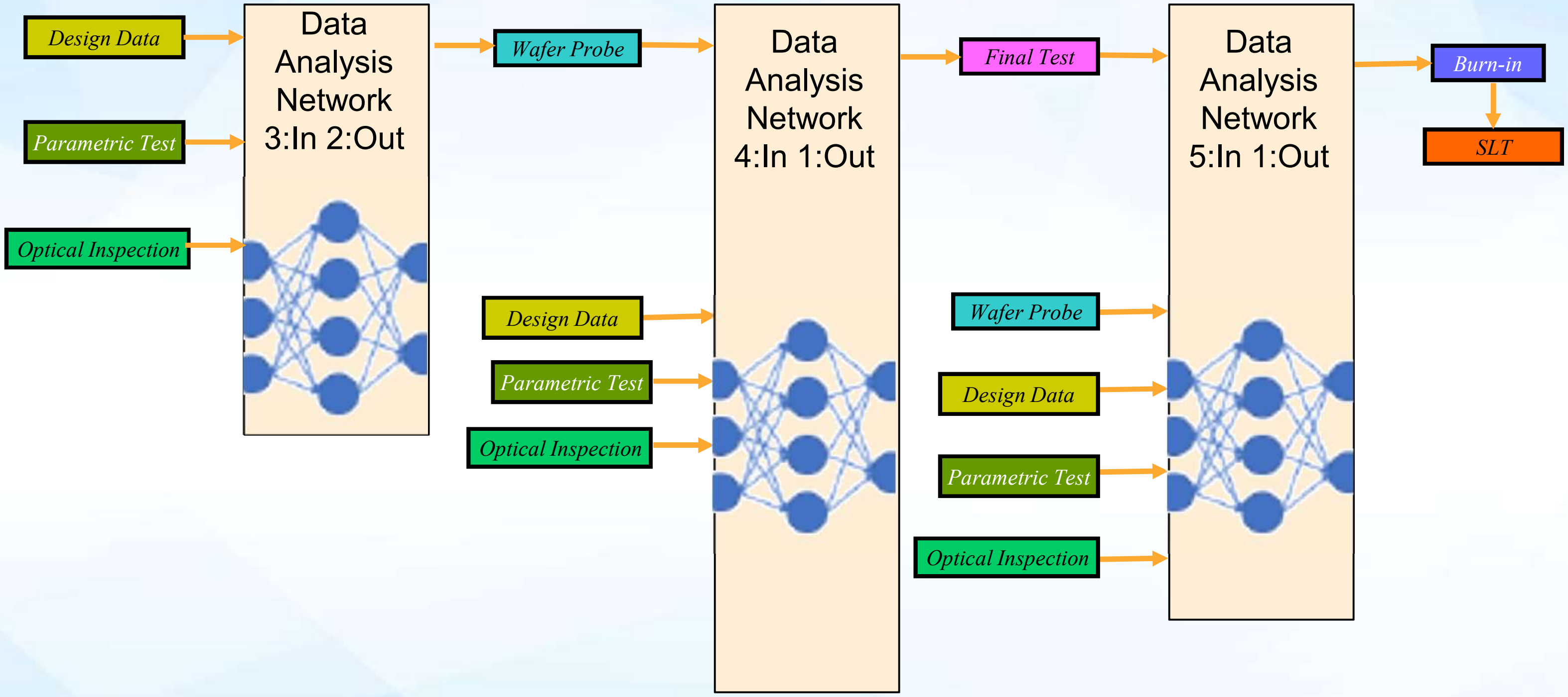


Lots
of data on
lots of tests on
lots of chips on
lots of wafers !

Key Advantages of Applying NN to Semiconductor Manufacturing

1. Lots of data available:
   - TSMC capacity is > 2M wafers/month
   - Each device likely has > 1K test data points (manufacturing & E-test data is in addition)
   - 500-10,000 devices per wafer
2. Semiconductor “images” don’t move around.
   - Unlike cars & dogs, correlations are always done in a fixed location.
3. Semiconductor industry models everything
4. Target to train against is fairly well understood.
   - Test results.
   - High yield.
   - Low cost.

Feed Forward Possibilities in Test



Design Data

Parametric Test

Optical Inspection

Data Analysis Network 3:In 2:Out

Wafer Probe

Design Data

Parametric Test

Optical Inspection

Data Analysis Network 4:In 1:Out

Final Test

Wafer Probe

Design Data

Parametric Test

Optical Inspection

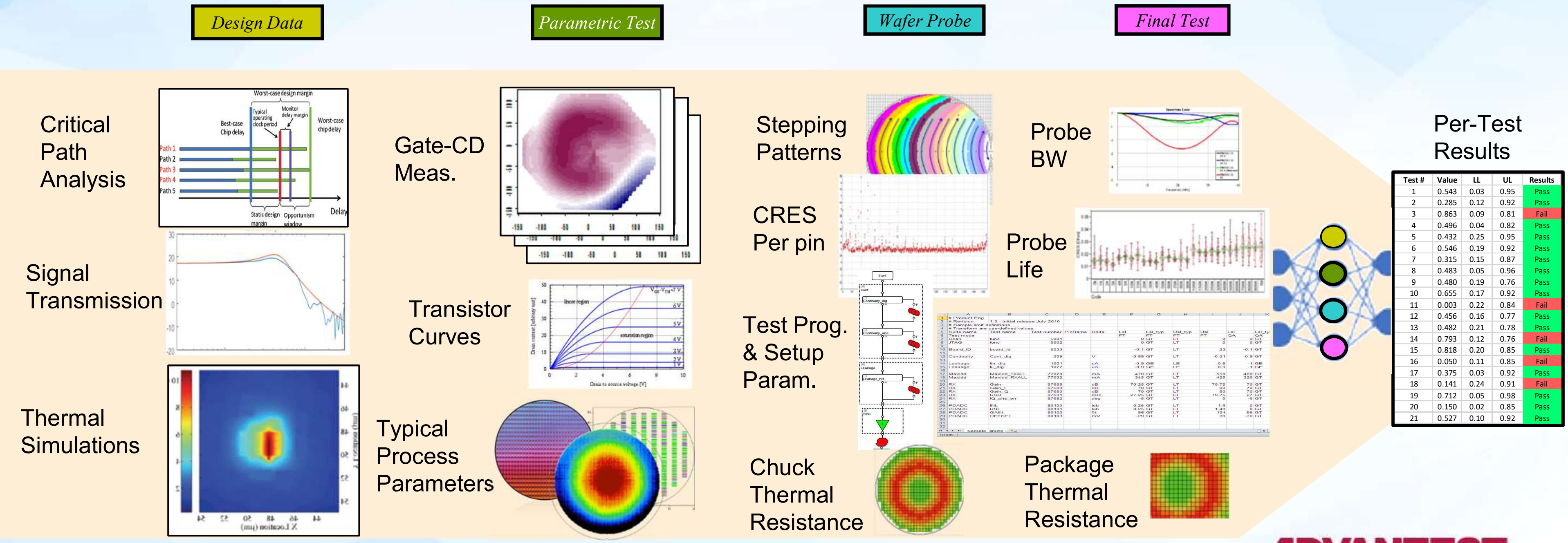
Data Analysis Network 5:In 1:Out

Burn-in

SLT

What are Some of the Neural Networks Inputs Which Might Make Sense?

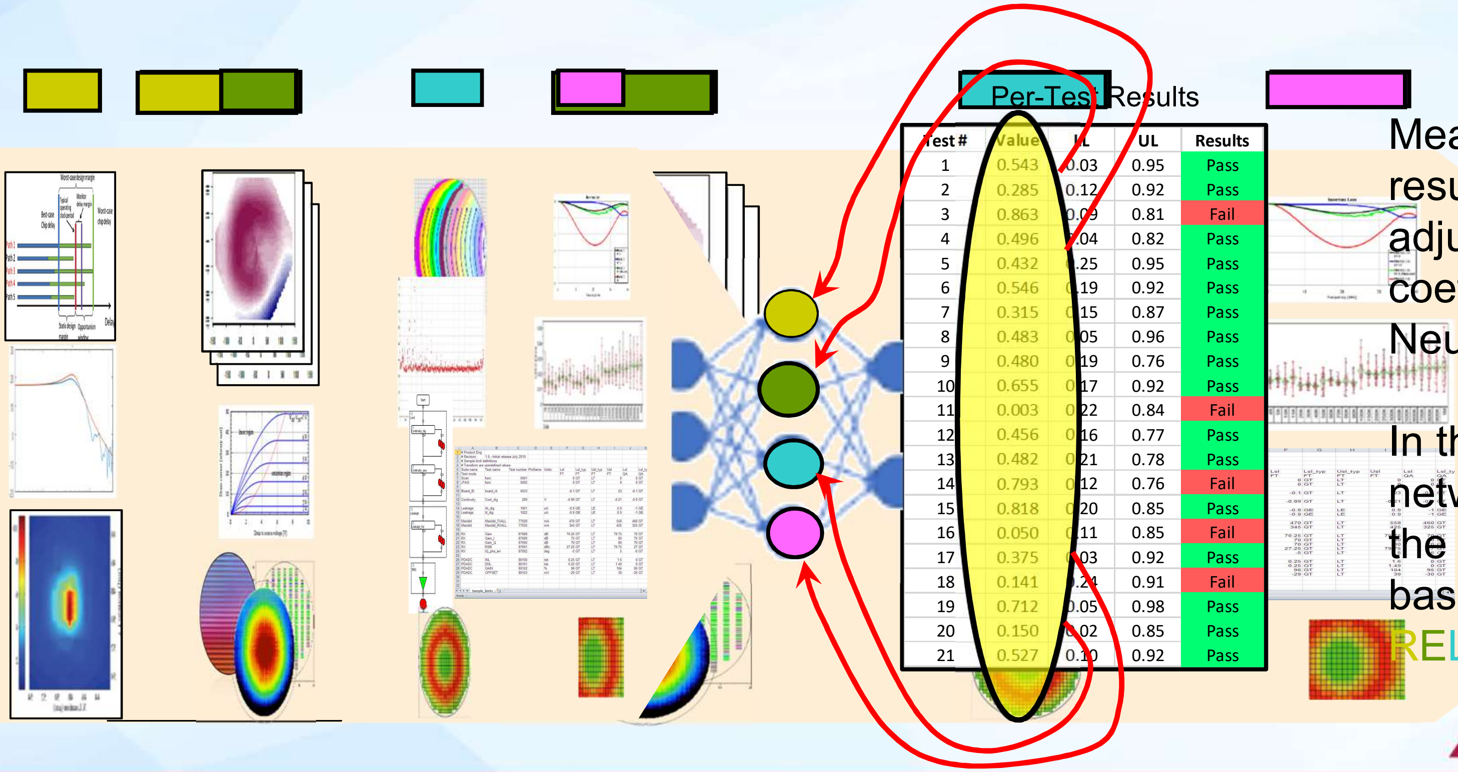
- Let’s look just at an Fmax Test . . .



Test #	Value	LL	UL	Results
1	0.543	0.03	0.95	Pass
2	0.285	0.12	0.92	Pass
3	0.863	0.09	0.81	Fail
4	0.496	0.04	0.82	Pass
5	0.432	0.25	0.95	Pass
6	0.546	0.19	0.92	Pass
7	0.315	0.15	0.87	Pass
8	0.483	0.05	0.96	Pass
9	0.480	0.19	0.76	Pass
10	0.655	0.17	0.92	Pass
11	0.003	0.22	0.84	Fail
12	0.456	0.16	0.77	Pass
13	0.482	0.21	0.78	Pass
14	0.793	0.12	0.76	Fail
15	0.818	0.20	0.85	Pass
16	0.050	0.11	0.85	Fail
17	0.375	0.03	0.92	Pass
18	0.141	0.24	0.91	Fail
19	0.712	0.05	0.98	Pass
20	0.150	0.02	0.85	Pass
21	0.527	0.10	0.92	Pass

Note: Data shown for example only.
Likely bears no resemblance to actual distributions.

How Will We Train this Neural Network?



Per-Test Results

Test #	Value	LL	UL	Results
1	0.543	0.03	0.95	Pass
2	0.285	0.12	0.92	Pass
3	0.863	0.09	0.81	Fail
4	0.496	0.04	0.82	Pass
5	0.432	0.25	0.95	Pass
6	0.546	0.19	0.92	Pass
7	0.315	0.15	0.87	Pass
8	0.483	0.05	0.96	Pass
9	0.480	0.19	0.76	Pass
10	0.655	0.17	0.92	Pass
11	0.003	0.22	0.84	Fail
12	0.456	0.16	0.77	Pass
13	0.482	0.21	0.78	Pass
14	0.793	0.12	0.76	Fail
15	0.818	0.20	0.85	Pass
16	0.050	0.11	0.85	Fail
17	0.375	0.03	0.92	Pass
18	0.141	0.24	0.91	Fail
19	0.712	0.05	0.98	Pass
20	0.150	0.02	0.85	Pass
21	0.527	0.10	0.92	Pass

Measurement results are used to adjust the tap coefficients of the Neural Network.

In this fashion, the network convolves to the optimal settings based on ALL RELEVENT inputs.

Per-Test Results

Test #	Value	LL	UL	Results
1	0.543	0.03	0.95	Pass
2	0.285	0.12	0.92	Pass
3	0.863	0.09	0.81	Fail
4	0.496	0.04	0.82	Pass
5	0.432	0.25	0.95	Pass
6	0.546	0.19	0.92	Pass
7	0.315	0.15	0.87	Pass
8	0.483	0.05	0.96	Pass
9	0.480	0.19	0.76	Pass
10	0.655	0.17	0.92	Pass
11	0.003	0.22	0.84	Fail
12	0.456	0.16	0.77	Pass
13	0.482	0.21	0.78	Pass
14	0.793	0.12	0.76	Fail
15	0.818	0.20	0.85	Pass
16	0.050	0.11	0.85	Fail
17	0.375	0.03	0.92	Pass
18	0.141	0.24	0.91	Fail
19	0.712	0.05	0.98	Pass
20	0.150	0.02	0.85	Pass
21	0.527	0.10	0.92	Pass

Note: Data shown for example only.
Likely bears no resemblance to actual distributions.

What is Required To Implement This?

1. Test measurement values are critical ➔ ATE needs an efficient way to gather and transmit large volumes of data.

2. Infrastructure needs to be upgraded to support big-data ➔ Data bandwidth improvements, elimination of data-silos, improved encryption with easy access for those who need the data.

3. Data analysts need to monitor and help direct the convolutional efforts ➔ Speed convergence, avoid instabilities.

4. Eventually, ATE will need to use Neural Network analysis to pinpoint good and bad devices ➔ "Low hanging" gains likely in both yield and device reliability.

Conclusions

- The potential gains associated with Machine Learning Algorithms such as Convolutional Neural Networks are considerable.

- The tools which support deploying this in semiconductor manufacturing are just arriving on the market.

- It will take some work to make this happen but the fruits of this labor will (in my opinion) revolutionize the semiconductor industry.

Backup

Adaptive Testing Today is Most Often Only Used Between Wafer & Final Test