INCOSE
Healthcare
Working Group
5th Annual Systems Engineering in Healthcare Conference
May 1-2, 2019
Minneapolis, MN

Machine learning is a Systems Engineering problem

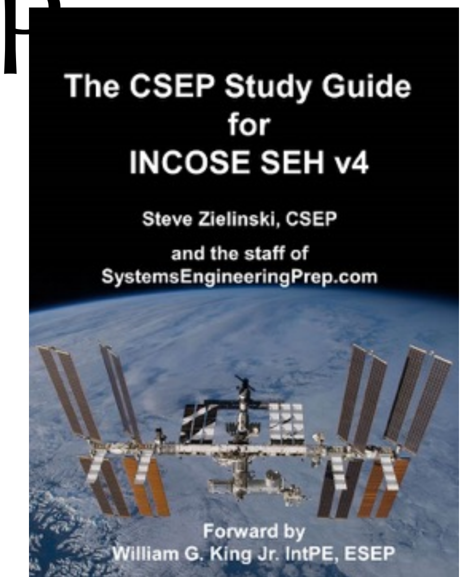
INCOSE HWGSEC19

Steve Zielinski, CSEP

Sr. Systems Engineering Manager
at Boston Scientific

BA University of St. Thomas
BEE University of Minnesota
MSDD University of St. Thomas

PMP, CSEP



In Case You Missed It

Original Investigation | Innovations in Health Care Delivery

FREE

December 13, 2016

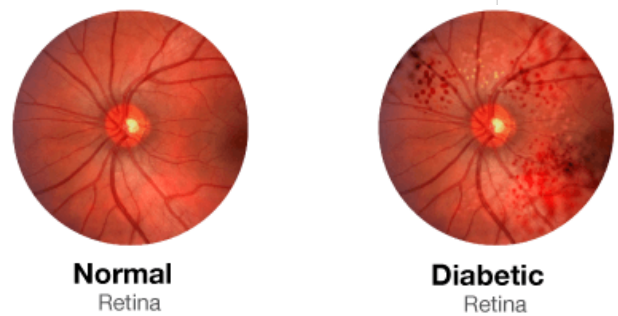
Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD[1]; Lily Peng, MD, PhD[1]; Marc Coram, PhD[1]; et al

Author Affiliations | Article Information

JAMA. 2016;316(22):2402-2410. doi:10.1001/jama.2016.17216

Machine Learning Website



INCOSE
Healthcare
Working Group
5th Annual Systems Engineering in Healthcare Conference
May 1-2, 2019
Minneapolis, MN

nature
International journal of science

Letter | Published: 25 January 2017

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun

Nature **542**, 115–118 (02 February 2017) | Download Citation

INCOSE Healthcare Working Group

5th Annual Systems Engineering in Healthcare Conference

May 1-2, 2019
Minneapolis, MN

Stanford ML Group

Cardiologist-Level Arrhythmia Detection With Convolutional Neural Networks

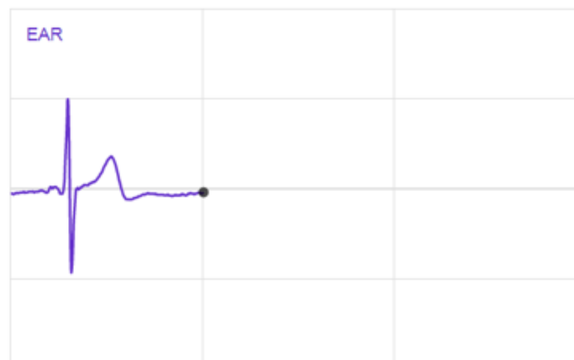
Pranav Rajpurkar*, Awni Hannun*, Masoumeh Haghpanahi, Codie Bourn, and Andrew Ng

A collaboration between Stanford University and iRhythm Technologies

We develop a model which can diagnose irregular heart rhythms, also known as arrhythmias, from single-lead ECG signals better than a cardiologist.

Key to exceeding expert performance is a deep convolutional network which can map a sequence of ECG samples to a sequence of arrhythmia annotations along with a novel dataset two orders of magnitude larger than previous datasets of its kind.

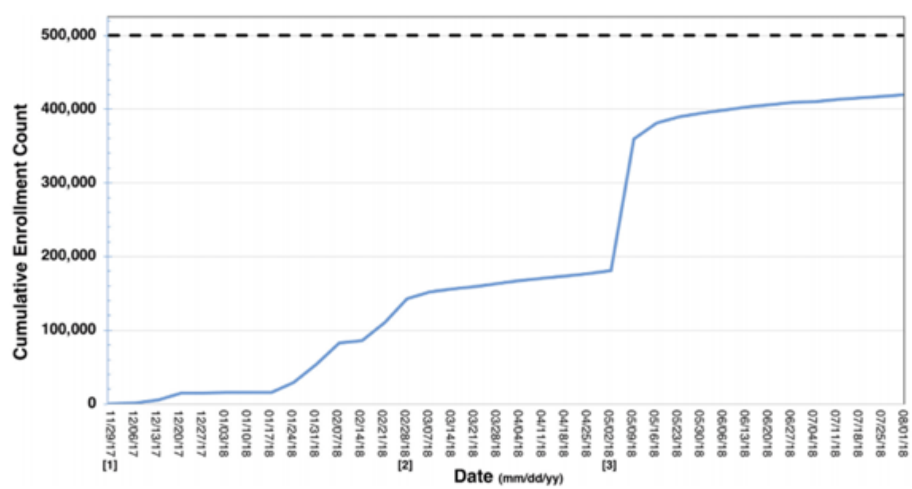
READ OUR PAPER



INCOSE
Healthcare
Working Group
5th Annual Systems Engineering in Healthcare Conference
May 1-2, 2019
Minneapolis, MN

Now we can enroll 400K+ participants within a year from different geographies

Figure 3



Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smartwatch: The Apple Heart Study

400K+ participants across the entire United States within a year. It's important to note that these are still iPhone users, who tend to be affluent.

CBINSIGHTS American Heart Journal

M | Data Driven Investor Follow

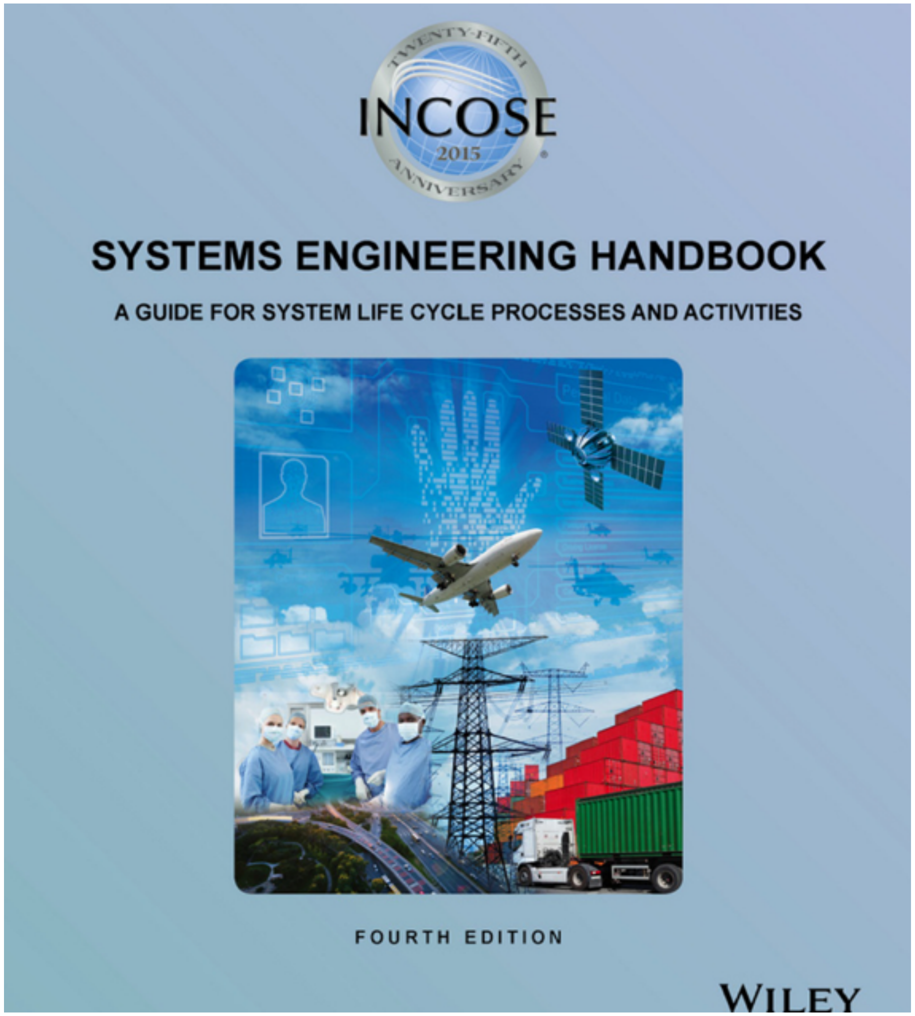
HOME FINTECH A.I. BLOCKCHAIN FINANCE ABOUT COLLABORATE SUBSCRIBE | DDI

Benefits of Robotic Process Automation in Healthcare

Craig Richardville Follow

Nov 20, 2018 · 3 min read

INCOSE
Healthcare
Working Group
5th Annual Systems Engineering in Healthcare Conference
May 1-2, 2019
Minneapolis, MN



"One of the systems engineer's first and most important responsibilities … is to establish nomenclature and terminology that support clear, unambiguous communication … "

INCOSE SEH 1.5

Types of Machine Learning

Machine Learning

Supervised	Unsupervised	Reinforcement
Task driven (Regression / Classification)	Data driven (Clustering)	Algorithm learns to react to an environment

https://medium.com/deep-math-machine-learning-ai/different-types-of-machine-learning-and-their-types-34760b9128a2

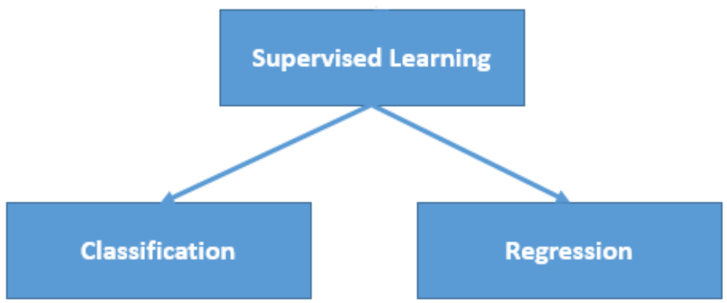
Supervised vs Unsupervised

- Supervised
  - You begin with a data set where the “correct” answer is known.

- Unsupervised
  - The examples in the data set don’t have a “correct” answer.

Supervised

- Regression: Predict continuous valued output (price)
- Classification
  - Classification Discrete valued output (0 or 1)



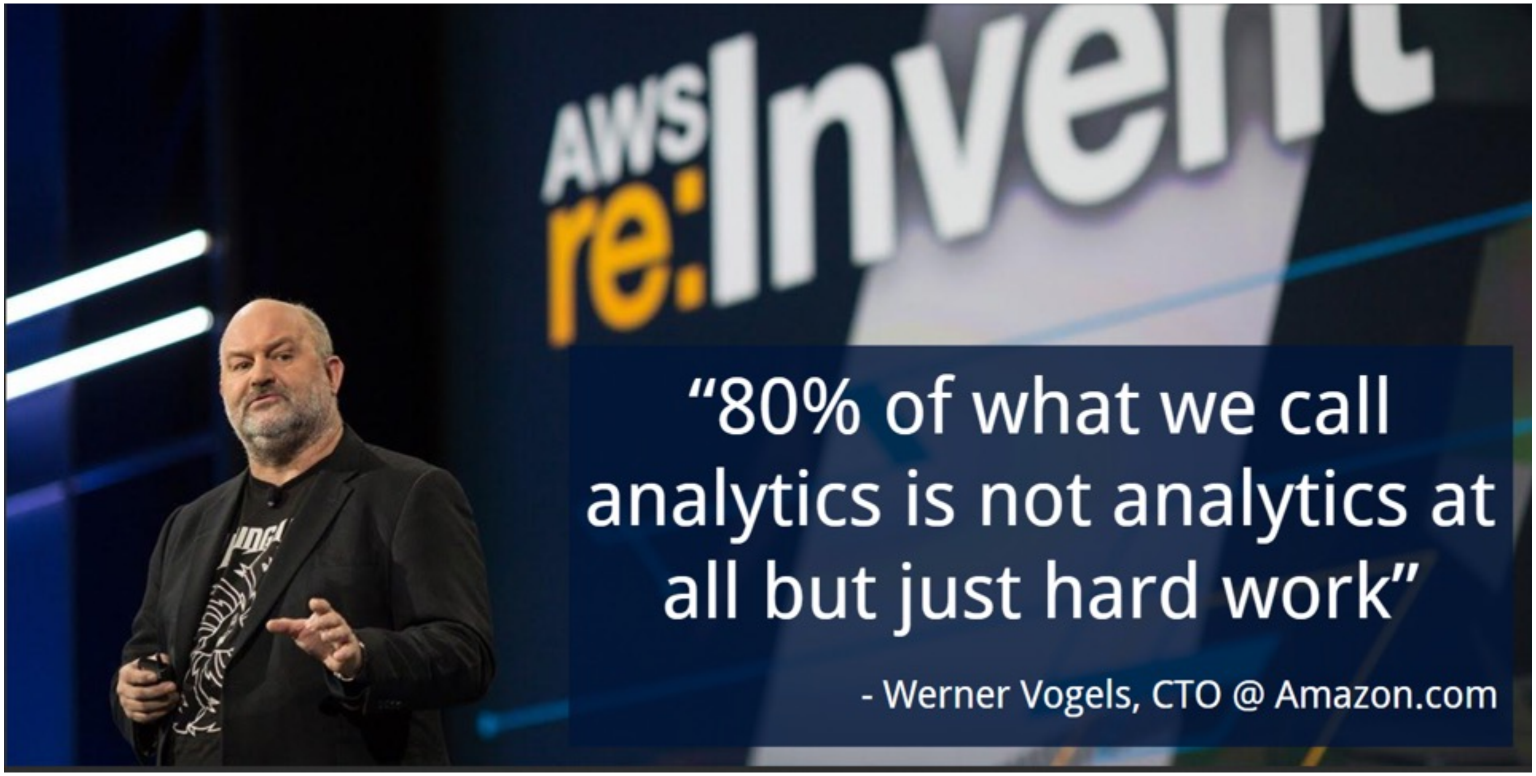
Unsupervised

- Given a database of customer data, automatically discover market segments and group customers into different market segments.

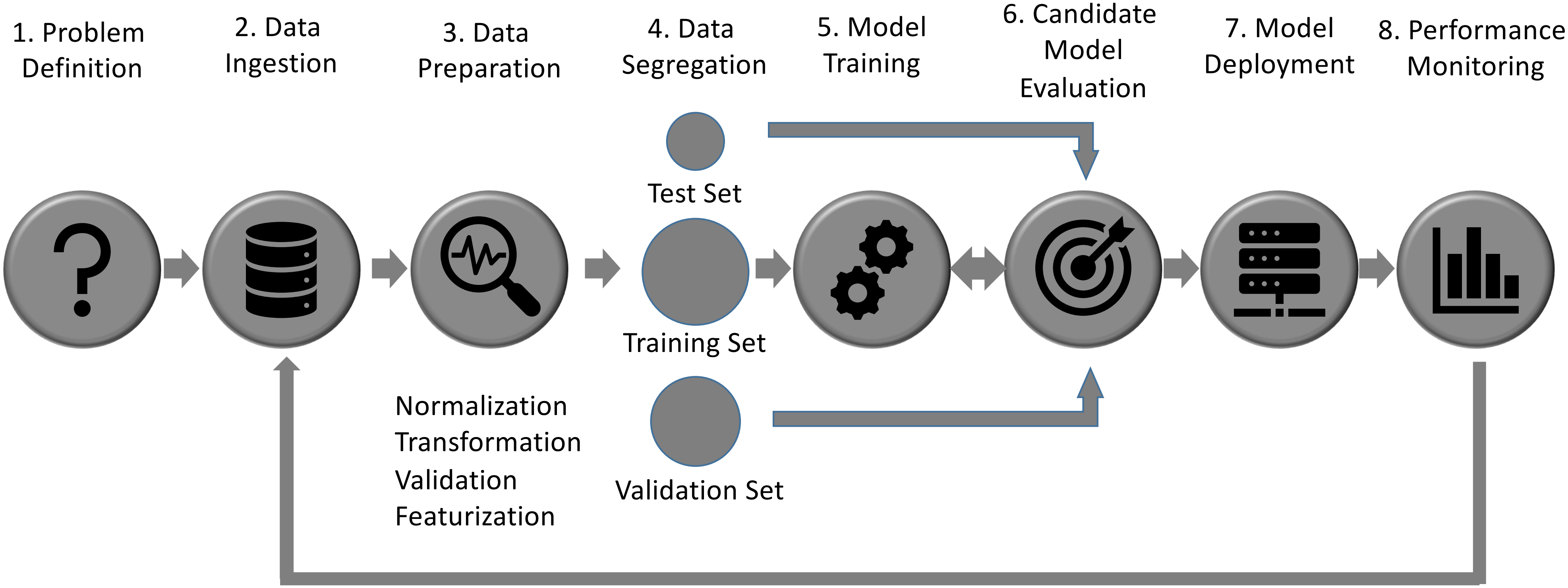
"Locked" vs Continuously Learning (Adaptive)

- "Locked" algorithms are those that provide the same result each time the same input is provided.
- Continuously learning algorithms take into account new data that (hopefully) improves their performance over time.

Machine learning is a Systems Engineering problem



https://www.youtube.com/watch?v=ZPbM2qGfH3s


1. Problem Definition
2. Data Ingestion
3. Data Preparation
4. Data Segregation
5. Model Training
6. Candidate Model Evaluation
7. Model Deployment
8. Performance Monitoring
Test Set
Training Set
Validation Set
Normalization
Transformation
Validation
Featurization

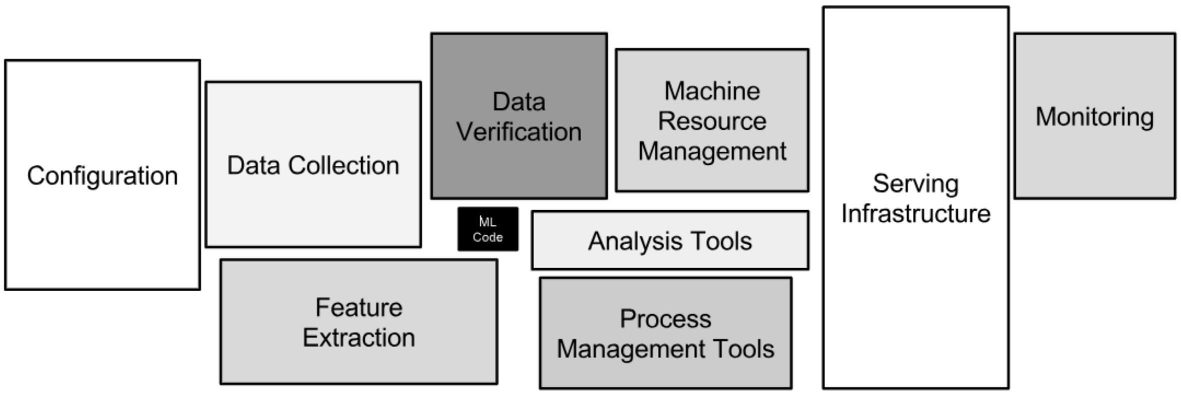


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

D. Sculley, et al. Hidden technical debt in machine learning systems. In Neural Information Processing Systems (NIPS). 2015

What's the FDA Got to Say?

"The traditional paradigm of medical device regulation was not designed for adaptive AI/ML technologies, which have the potential to adapt and optimize device performance in real-time to continuously improve healthcare for patients. The highly iterative, autonomous, and adaptive nature of these tools requires a new, total product lifecycle (TPLC) regulatory approach that facilitates a rapid cycle of product improvement and allows these devices to continually improve while providing effective safeguards."

from "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback"

"To date, FDA has cleared or approved several AI/ML-based SaMD. Typically, these have only included algorithms that are "locked" prior to marketing, where algorithm changes likely require FDA premarket review for changes beyond the original market authorization."

INCOSE
Healthcare Working Group
5th Annual Systems Engineering in Healthcare Conference
May 1-2, 2019
Minneapolis, MN

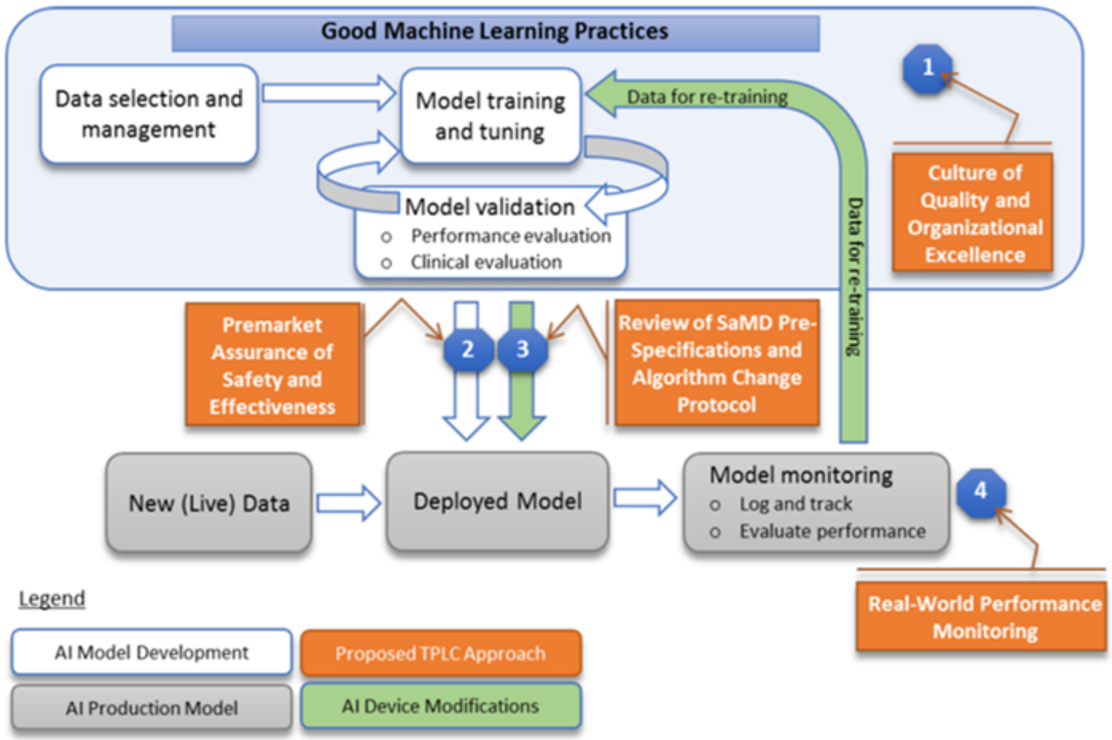


Figure 2: Overlay of FDA's TPLC approach on AI/ML workflow

How big is "big Data"?

INCOSE
Healthcare Working Group
5th Annual Systems Engineering in Healthcare Conference
May 1-2, 2019
Minneapolis, MN

Defining "Big": Algorithm vs Data

Stanford ML Group

Cardiologist-Level Arrhythmia Detection With Convolutional Neural Networks

Pranav Rajpurkar*, Awni Hannun*, Masoumeh Haghpanahi, Codie Bourn, and Andrew Ng

A collaboration between Stanford University and iRhythm Technologies

We develop a model which can diagnose irregular heart rhythms, also known as arrhythmias, from single-lead ECG signals better than a cardiologist.

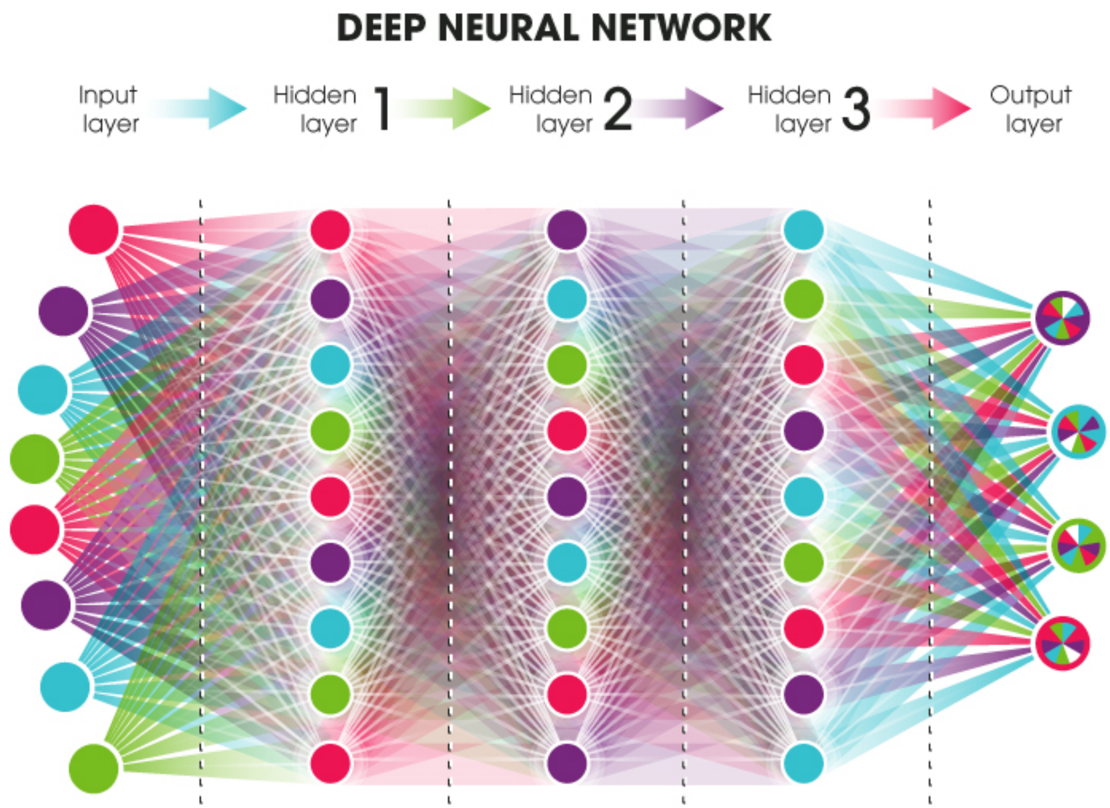
Key to exceeding expert performance is a deep convolutional network which can map a sequence of ECG samples to a sequence of arrhythmia annotations along with a novel dataset two orders of magnitude larger than previous datasets of its kind.

READ OUR PAPER

- 34-layer network
- ~ 64,121 ECG records

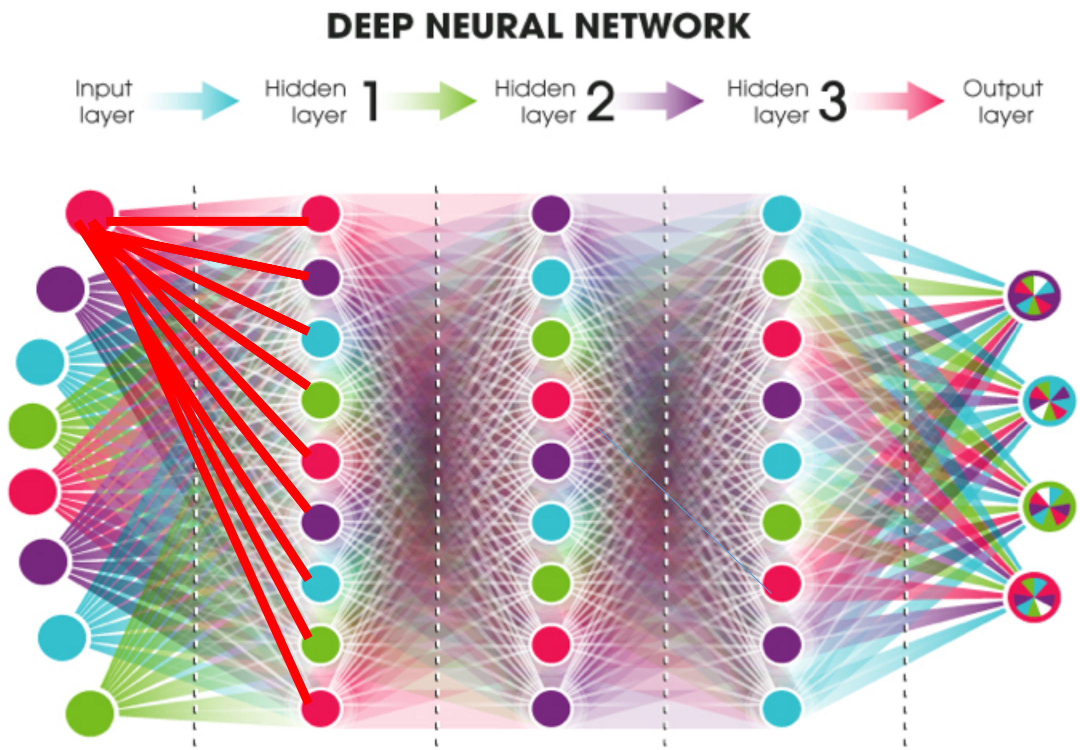
INCOSE
Healthcare
Working Group
5th Annual Systems Engineering in Healthcare Conference
May 1-2, 2019
Minneapolis, MN

Algorithm



neuralnetworksanddeeplearning.com - Michael Nielsen, Yoshua Bengio, Ian Goodfellow, and Aaron Courville, 2016.

Your task is to determine the weight (a real number) for everyone of those lines.



neuralnetworksanddeeplearning.com - Michael Nielsen, Yoshua Bengio, Ian Goodfellow, and Aaron Courville, 2016.

nature
International journal of science

Letter | Published: 25 January 2017

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun

Nature **542**, 115–118 (02 February 2017) | Download Citation

- 27 layer CNN

nature
International journal of science

Letter | Published: 25 January 2017

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun

Nature **542**, 115–118 (02 February 2017) | Download Citation

"We train a CNN using a dataset of 129,450 clinical images-two orders of magnitude larger than previous datasets …"

nature
International journal of science

Letter | Published: 25 January 2017

Dermatologist-level classification of skin cancer with deep neural networks

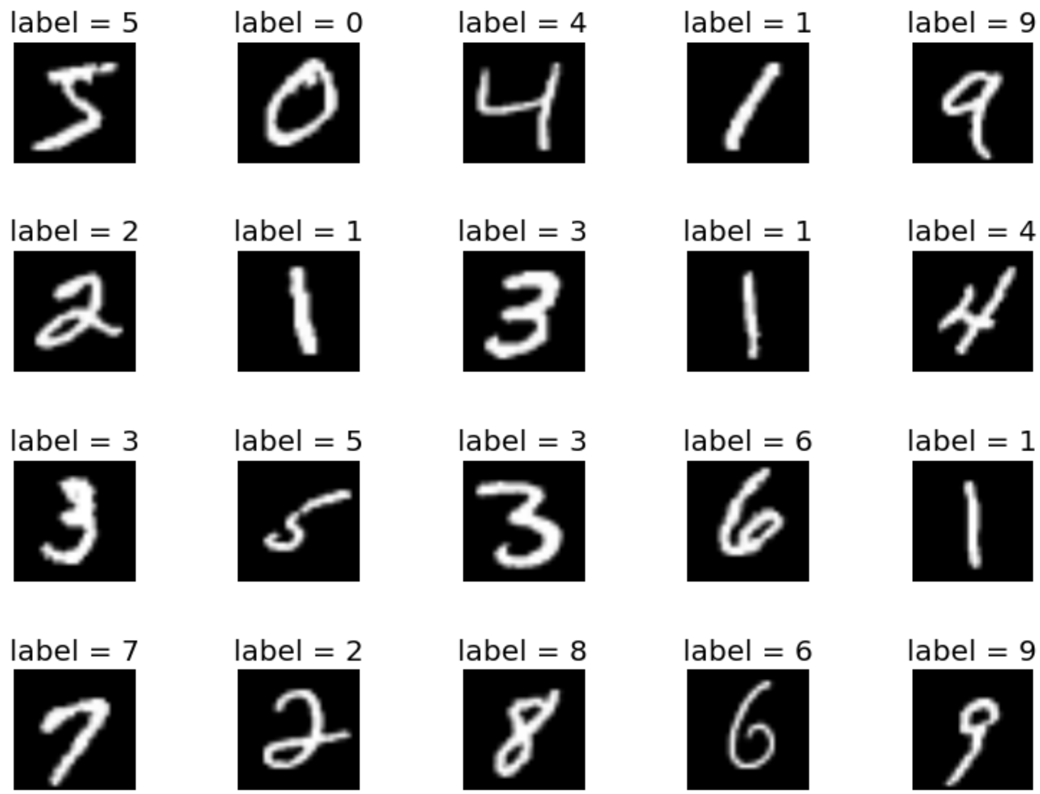
Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun

Nature **542**, 115–118 (02 February 2017) | Download Citation

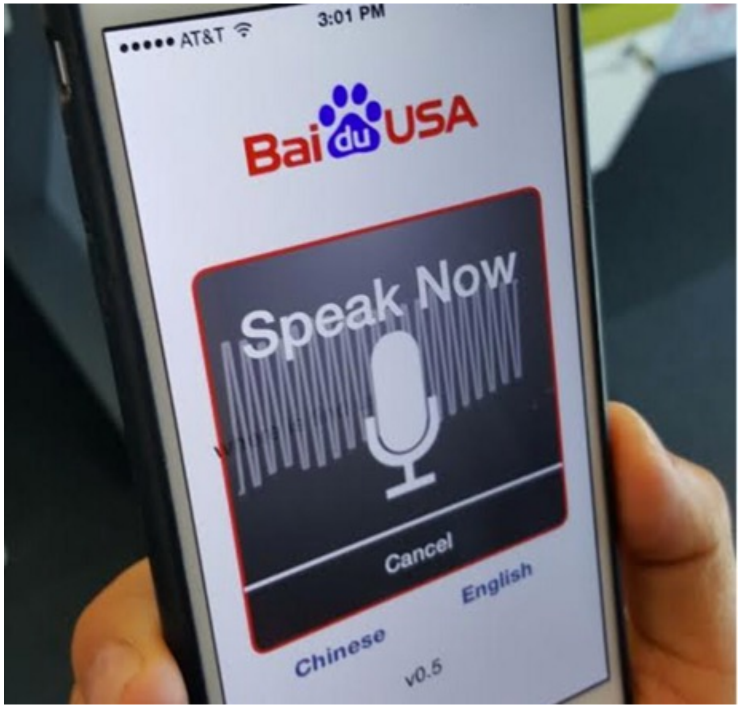
However, they didn't start from scratch!

"using a data-driven approach—1.41 million pre-training and training images"

MNIST database – 60,000 examples



Deep Speech 2



- 11,940 hours of English
- 9,400 hours of Mandarin
- tens of exaFLOPs that would require 3-6 weeks to execute on a single GPU.
- 16 GPUs –
  - 3 teraFLOP/second per GPU
  - cuts training times down to 3 to 5 days

1 exaFLOP = 10^{18} FLoating-point Operations
1 teraFLOP = 10^{12}

INCOSE
Healthcare
Working Group
5th Annual Systems Engineering in Healthcare Conference
May 1-2, 2019
Minneapolis, MN

The "cat" paper:
"Building High-level Features Using Large Scale Unsupervised Learning"

- 9-layers
- 1 billion connections
- 10 million 200x200 pixel images.
- 1,000 machine cluster
- 16,000 cores
- 3 days
- (from 2012)

- Training is the lengthy step.
- Presenting the model with an example following training can be very fast.
  - Autonomous cars go to the street with trained models.

Training vs Execution

What does this mean for hardware?

SPDRs

MarketWatch

The average smart phone has **2.7GB** of RAM memory and **43GB** of storage memory.

The average fully autonomous vehicle is expected to need **74GB** of RAM memory and **1TB (1,000GB)** in storage.

ce: Micron

INCOSE
Healthcare
Working Group
5th Annual Systems Engineering in Healthcare Conference
May 1-2, 2019
Minneapolis, MN

An AI-training server requires **6x more RAM memory** and **2x more storage memory** than a standard cloud server.

Source: Micron

ML and Hardware

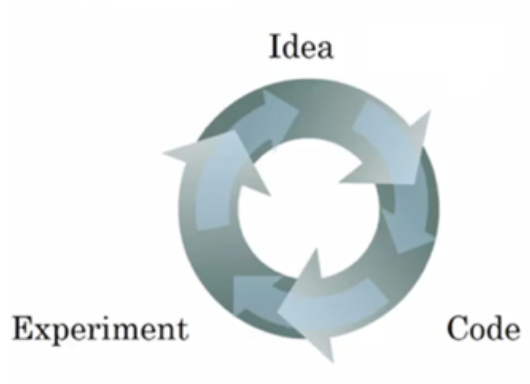
- " … the number of floating-point arithmetic operations needed to carry out machine learning neural networks is doubling every three and a half months."

- Whereas conventional neural nets may have hundreds of thousand of such weights that must be computed, or even millions, Google's scientists are saying "please give us a tera-weight machine," computers capable of computing a trillion weights. That's because "each time you double the size of the [neural] network, we get an improvement in accuracy." Bigger and bigger is the rule in AI.

-- Google software engineer Cliff Young opening keynote Linley Group Fall Processor Conference 2018

https://www.zdnet.com/article/google-says-exponential-growth-of-ai-is-changing-nature-of-compute/

Key SE Role



- Given the iterative nature of ML, SE must be concerned with optimizing the cycle time.

Is more data always better?

INCOSE
Healthcare Working Group
5th Annual Systems Engineering in Healthcare Conference
May 1-2, 2019
Minneapolis, MN

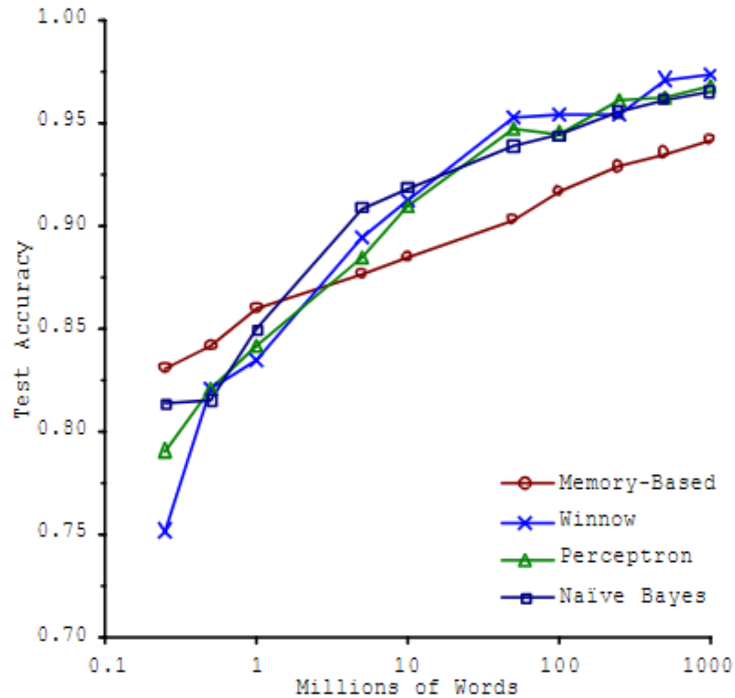
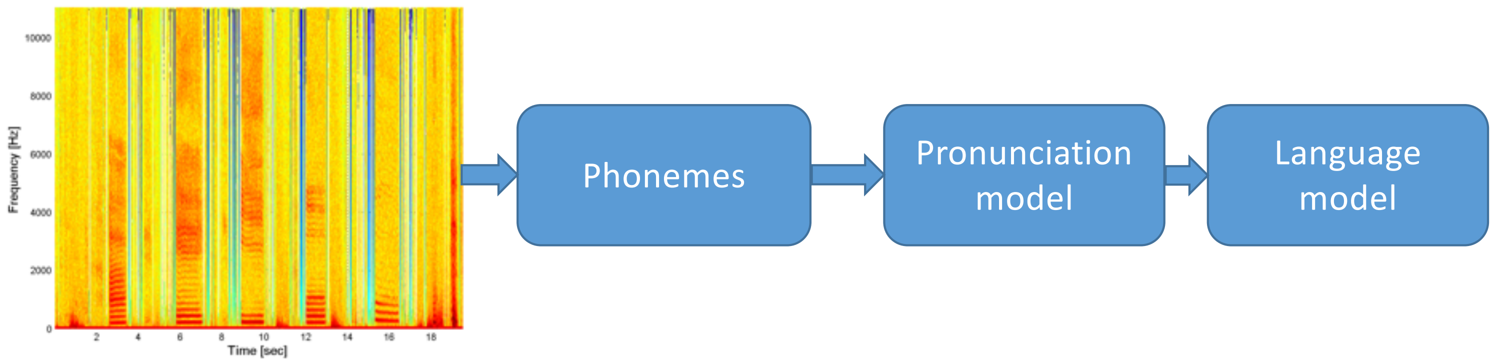


Figure 1. Learning Curves for Confusion Set Disambiguation

Banko and Brill, 2001

End-to-End Learning Adds to the "More Data" Argument

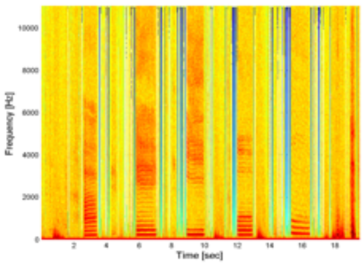
- Google Translate
  - The A.I. system [Translate] had demonstrated overnight improvements roughly equal to the total gains the old one had accrued over its entire lifetime. Started April 2006. AI switch occurred in November 2016.
  - Its speech-recognition team swapped out part of their old system for a neural network and encountered, in pretty much one fell swoop, the best quality improvements anyone had seen in 20 years.



“**A Bit of History**
Traditionally, speech recognition systems consisted of several components - an acoustic model that maps segments of audio (typically 10 millisecond frames) to phonemes, a pronunciation model that connects phonemes together to form words, and a language model that expresses the likelihood of given phrases. In early systems, these components remained independently-optimized.”

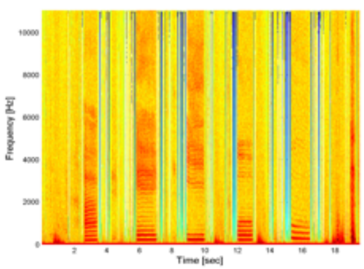
An All-Neural On-Device Speech Recognizer
Tuesday, March 12, 2019
Posted by Johan Schalkwyk, Google Fellow, Speech Team

INCOSE
Healthcare
Working Group
5th Annual Systems Engineering in Healthcare Conference
May 1-2, 2019
Minneapolis, MN



Phonemes → Pronunciation model → Language model → Transcript

VS



→ Learning algorithm → Transcript

This works when you have enough labeled (audio, transcript) data.

SE Question

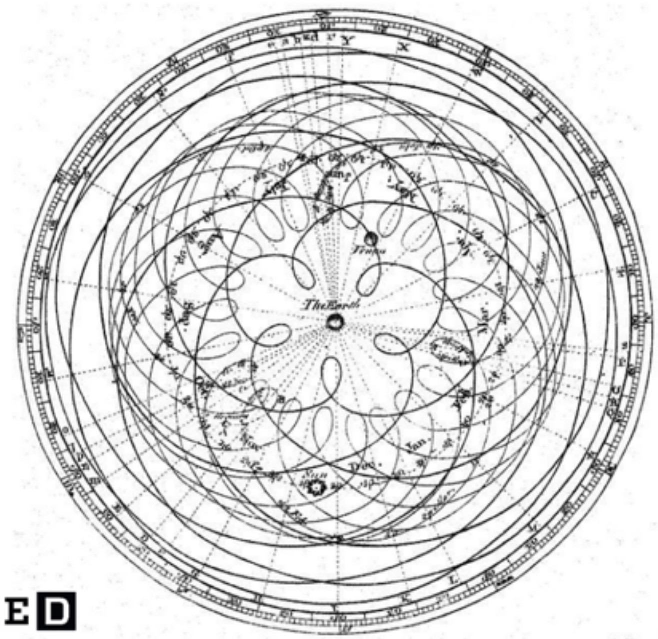
- What are you doing *today* to gather data?

Counterpoint

Domain Expertise is Critical

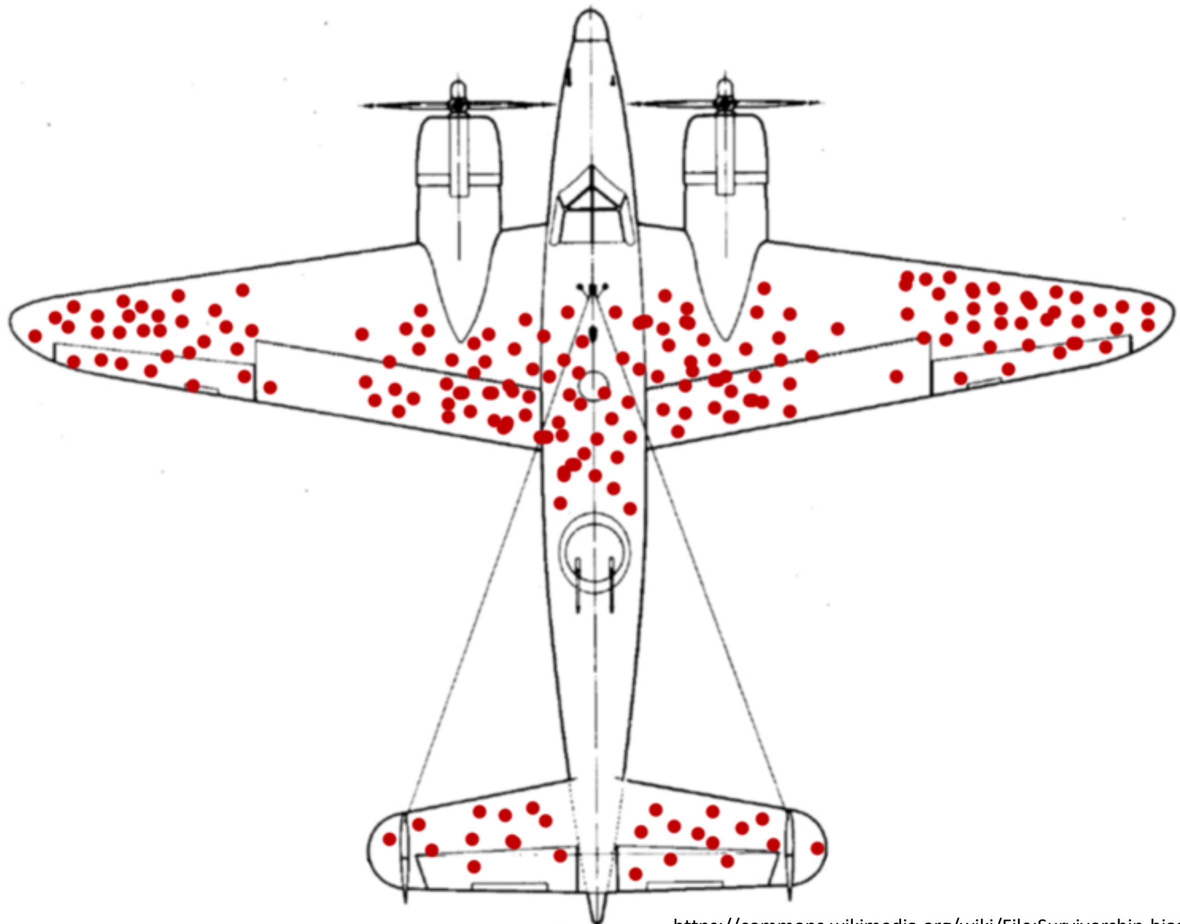
For big data to mature beyond marketing hype towards truly transformative solutions, it must "grow up" out of the computer science labs that gave birth to it and spend more time on understanding the domain-specific [problems] it is applied to than on the computing algorithms that operationalize them.

- Alev Leetaru WIRED



https://www.wired.com/2014/06/how-to-teach-heartless-computers-to-really-get-what-were-feeling/

How would a ML algorithm answer the question: Where should we add armor?

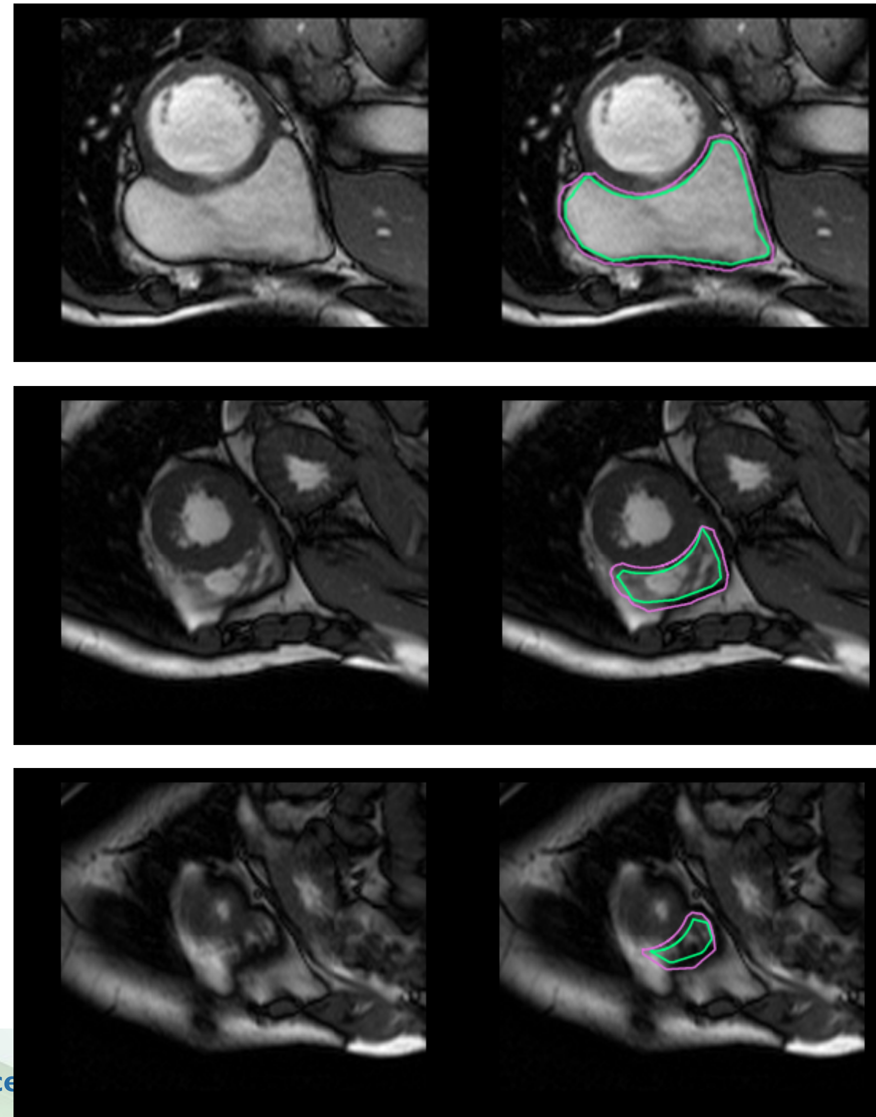


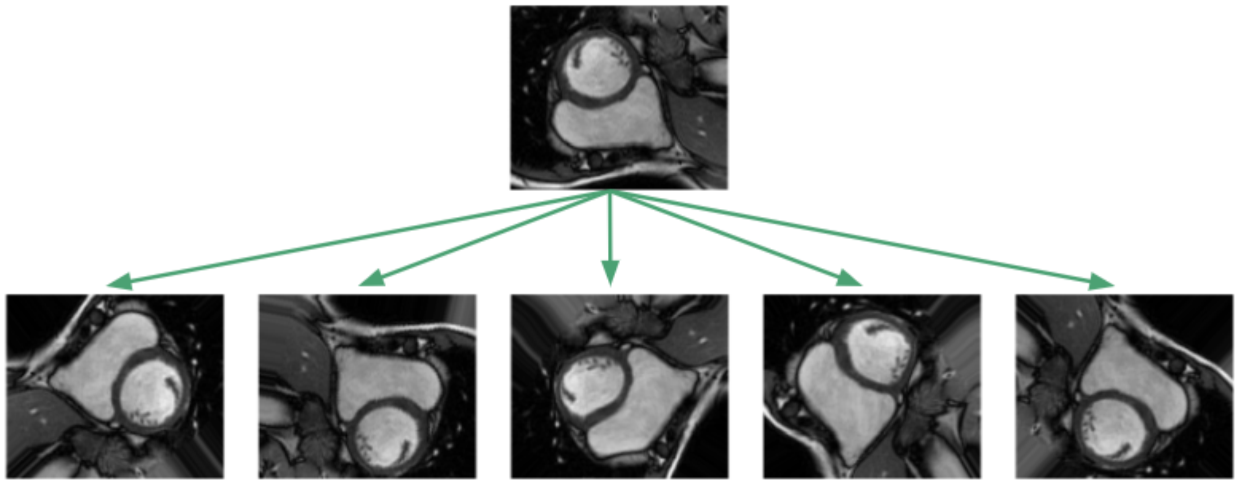
https://commons.wikimedia.org/wiki/File:Survivorship-bias.png

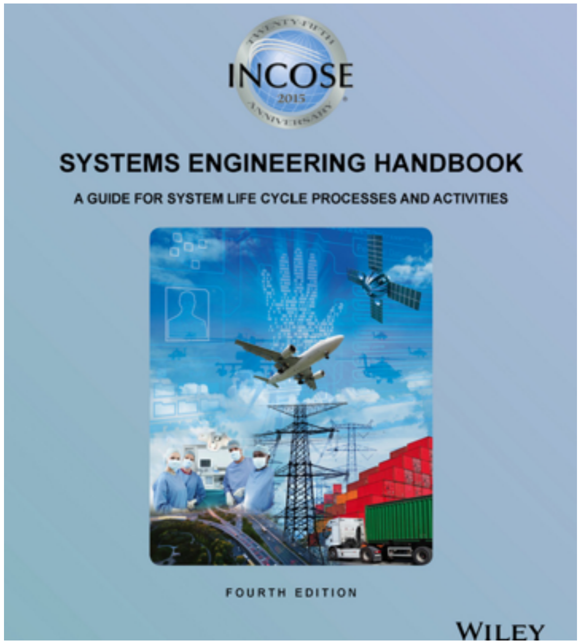
- Find the right-ventricle
  - The dataset contains only 243 physician-segmented images drawn from the MRIs of 16 patients.

https://chuckyee.github.io/cardiac-segmentation/

INCOSE
Healthcare
Working Group
5th Annual Systems Engineering in Healthcare Conference
May 1-2, 2019
Minneapolis, MN







What is success?

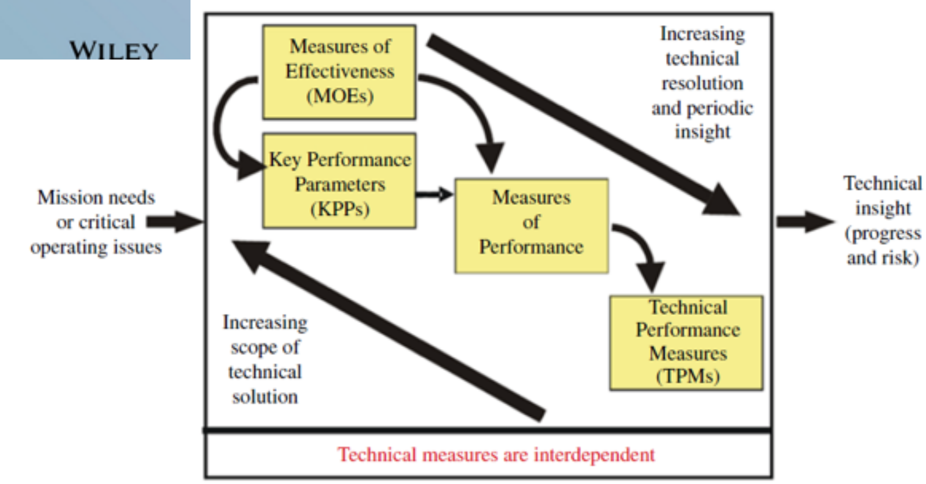


FIGURE 5.13 Relationship of technical measures. From INCOSE-TP-2003-020-01 Figure 1.1. Usage per the INCOSE Notices page. All other rights reserved.

The Problem with Accuracy

I identified 98% of the samples correctly!

Easy to do if the question is:

Does this biopsy contain cancer?

Just set: Result = Negative.

		Actual Class	
		1	0
Predicted Class	1	True Positive	False Positive
	0	False Negative	True Negative

$$\text{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$\text{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

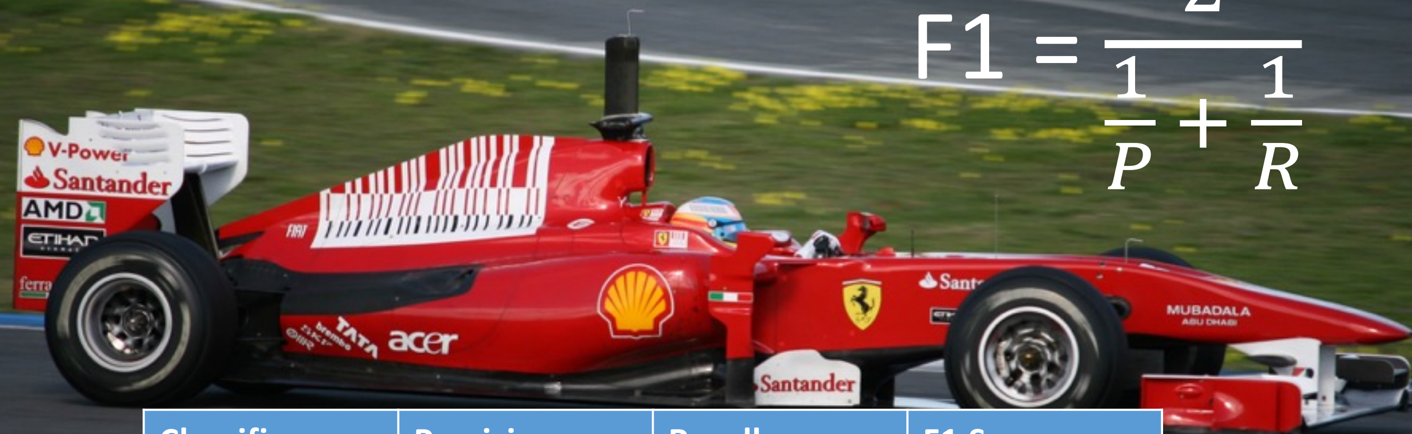
Which model is better?

Classifier	Precision	Recall
A	95%	90%
B	98%	85%

F1
Resolves the Precision vs Recall problem

- F1 “averages” precision and recall
- It’s actually the “harmonic mean”

$$F1 = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$



Classifier	Precision	Recall	F1 Score
A	95%	90%	92.4%
B	98%	85%	91.0%

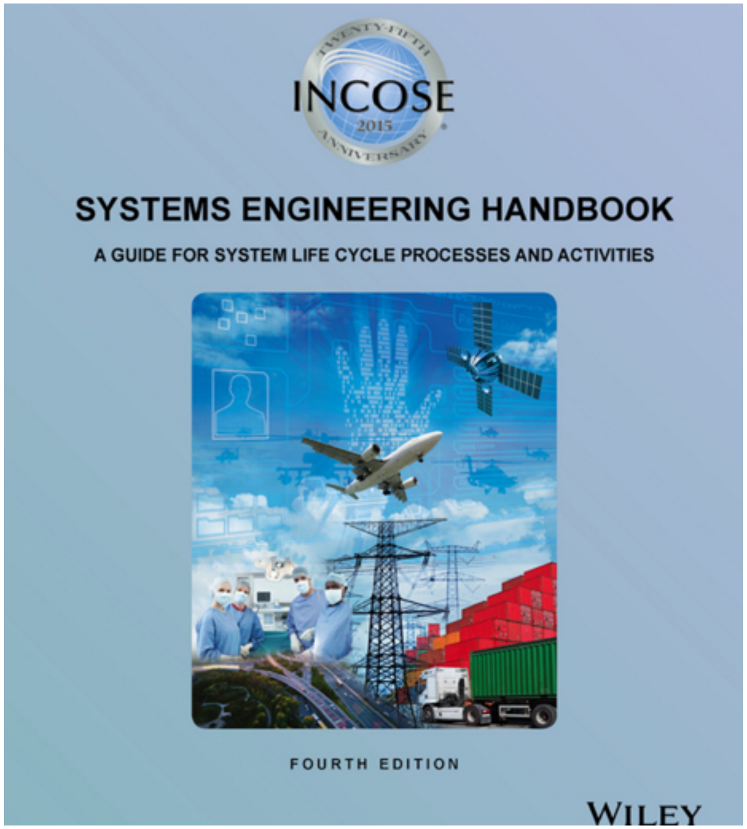
What to do when 1 number won't do?

- Satisfice!

- The term *satisficing*, a blend of *satisfy* and *suffice*,[2] was introduced by Herbert A. Simon in 1956

- If there are multiple things you care about
  1. Set a threshold value for one or more metrics (e.g. running time)
  2. Optimize around a different one

INCOSE
Healthcare
Working Group
5th Annual Systems Engineering in Healthcare Conference
May 1-2, 2019
Minneapolis, MN



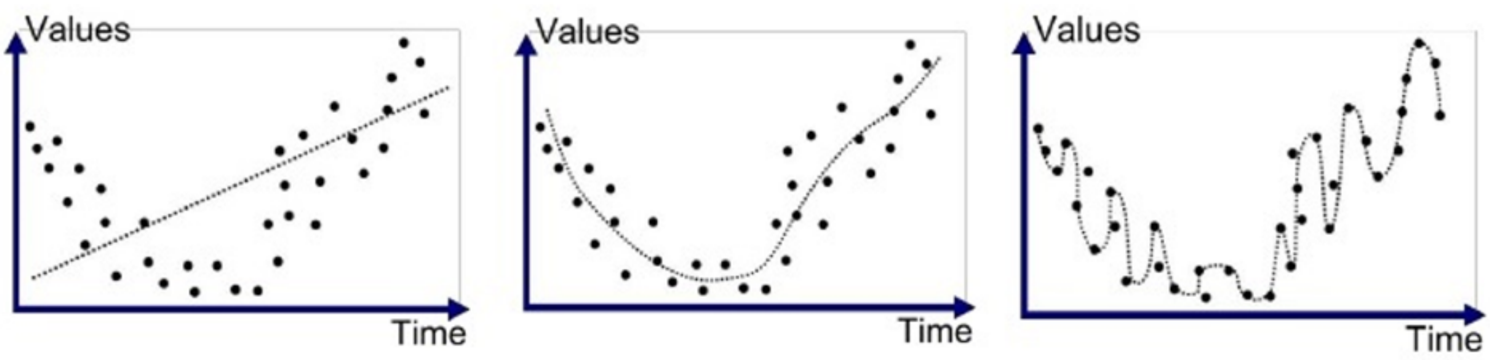
"… the human, an integral element of every system …"

--INCOSE SEH 10.13

Comparing to human-level performance

- Easier to build an ML system that automates a human task:
  - More data “easily” obtained from human labelers.
  - Error analysis can draw on human intuition
  - Better analysis of bias/variance.

Bias and Variance



Underfitted | Good Fit/Robust | Overfitted

High Bias

(Increase model size)
(More data does **not** help!)

High Variance

(Add training data)

Segment your data

3 segments is most common:

1. Training set
2. Dev set (development set) also called cross-validation set
3. Test set

Not long ago the recommended allocation was:

This makes sense for data sizes from 10 – 10,000

In many cases, the data set size has grown by 2 or more orders of magnitude. Data sizes of 1,000,000 are not uncommon.

INCOSE
Healthcare
Working Group
5th Annual Systems Engineering in Healthcare Conference
May 1-2, 2019
Minneapolis, MN

60% Train 20% Dev 20% Test

This makes sense for data sizes from 10 – 10,000

Current (with large datasets):

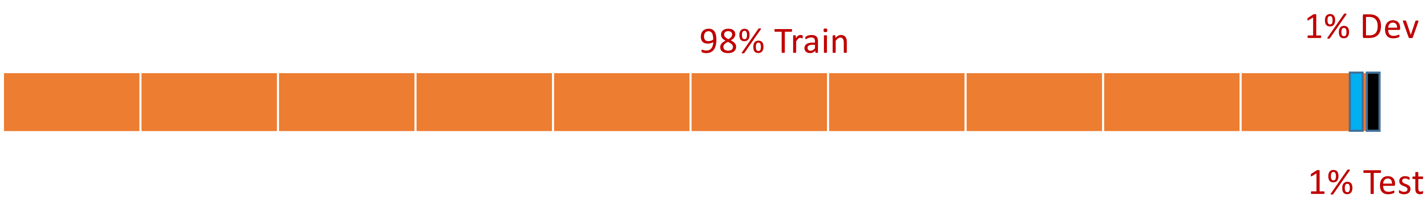
98% Train 1% Dev

1% Test

More data is always better!

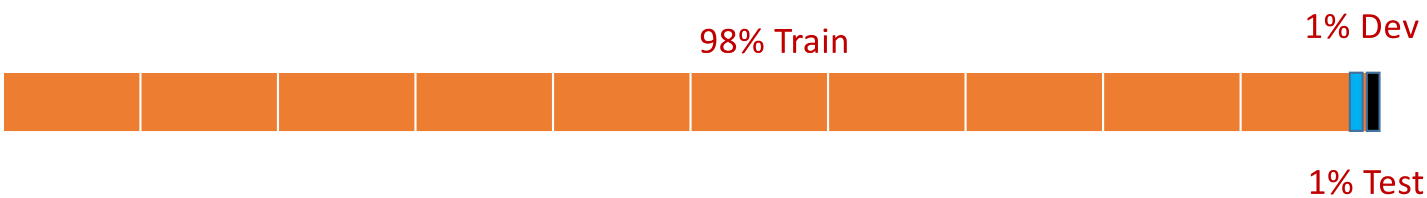
More Vocabulary

- Training set – used to determine the weights.
- Dev (development) set – tune parameters, select features, and make other decisions
- Test set – evaluate performance



Human-level error	1%
Training error	9%
Validation error	10%
Test error	11%

“Avoidable bias”



Human-level error	1%	
Training error	2%	"Overfitting/High Variance"
Validation error	10%	
Test error	11%	

Key SE role

- Considering the whole system:
  - Where are you getting additional data from?
    - Amazon Mechanical Turk
    - Sources of labeled data?
  - How frequently will you be updating your training/dev/test sets?
    - What is the expected life of your data? (Think software virus detection.)
      - A sophisticated enemy will be altering the patterns associated with their attacks/defenses.
  - Are the sets uniformly distributed? (Skewed sets will cause poor results.)

Engineering Your ML Strategy

- There are a lot of ways to potentially improve your results:
  - Collect more data
  - Collect more diverse data
  - Train longer
  - Change your optimization algorithm
  - Try bigger network
  - Try smaller network
  - Change your network architecture …
- What should you do?
  - The wrong decision can cost months!

INCOSE

Healthcare Working Group

5th Annual Systems Engineering in Healthcare Conference

May 1-2, 2019 Minneapolis, MN

Conclusion

Remember this picture!

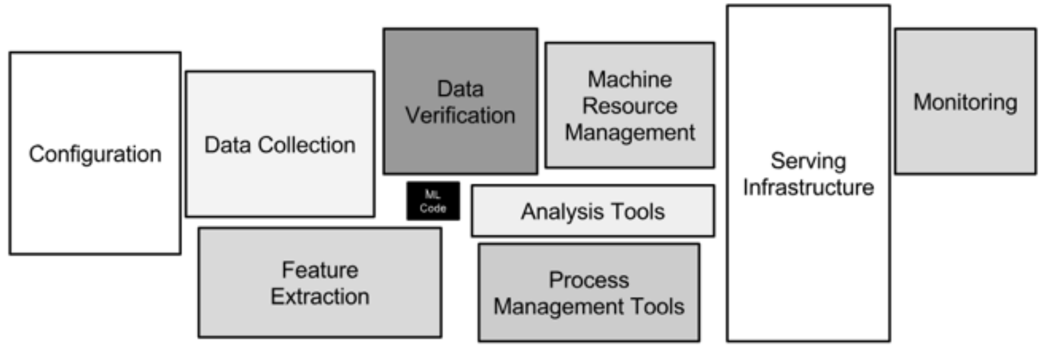
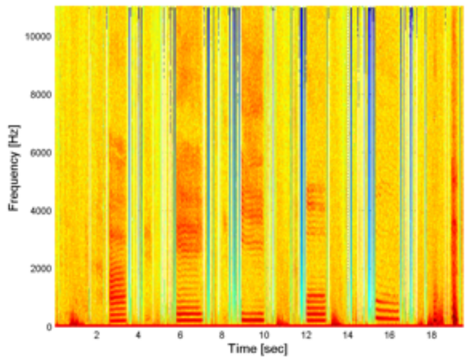


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

- And leave with an appreciation for:
  - Data sizes
  - Model sizes
  - Iterative nature
  - Evaluation
  - Vocabulary of ML


Frequency [Hz]
10000
8000
6000
4000
2000
0
2
4
6
8
10
12
14
16
18
Time [sec]

Thank you for attending!

Share your experiences at #HWGSEC