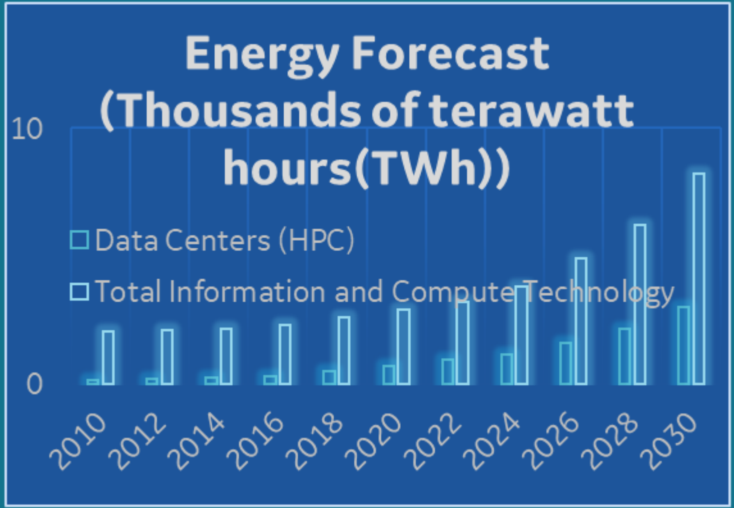
GE Research

The Entropy Economy: A New Paradigm for Carbon Reduction and Energy Efficiency for the Age of AI

Scott Evans, Tapan Shah and Achalesh Pandey
{evans, tapan.shah, achalesh.pandey}@ge.com

Introduction, Key Problem and Contribution

By 2030, over 20% of Energy Used Worldwide is expected to be consumed by Compute (10% from HPC's)



https://www.nature.com/articles/d41586-018-06610-y

1) Introduce the term **Entropy Economy,** which proposes to jointly optimizes compute, energy and waste heat
2) Capture practical examples illustrating **Entropy Economy**.
3) Introducing **Energy Aware Machine Learning (EAML)** as a driver of Entropy Economy
4) A detailed case study illustrating examples of EAML and some experimental results to highlight the salient points
5) Provide vision for future work

The Entropy Economy and Energy Aware Machine Learning are New Paradigm that Can Help Address This Challenge

Prior Work

Machine Learning, Energy and Entropy

- Strubell et al. *Energy and Policy Considerations for Deep Learning in NLP*, Arxiv, 2019
- Martin et al. *Estimation of energy consumption in machine learning, J. Parallel Distributed Computing, 2019*
- Still et al. *The thermodynamics of prediction,* Arxiv, 2012
- Bernstein et al. *Computing the Information Content of Trained Neural Networks*, Arxiv, 2021
- Shannon, Kolmogorov, Solomonov and others

Data centers and Energy

- Zheng et al. *Mitigating Curtailment and Carbon Emissions through Load Migration between Data Centers*, Joule, 2020
- Yang et al. Large-scale and Extreme-scale Computing with Stranded Green Power, IEEE T. Parallel and Distributed Systems, 2017

Computation and Entropy

- Kolchinsky et al. *Thermodynamic cost of Turing machines*, PRR, 2020
- Prokopenko et al. *Transfer Entropy and Transient Limits of Computation,* Scientific Reports, 2014
- Wolpert, *Stochastic Thermodynamics of Computation, Arxiv, 2019*
- Landauer, *Irreversibility and heat generation in the computing process,* IBM Jour. R&D, 1961
- Evans, et. a, 2006, 2007 – miRNA and Nucleotide analysis using Compression as learning with MDLcompress

Entropy Economy

Why Entropy Economy?

Entropy Is Intrinsic to Both Thermodynamics and Information Theory

Thermodynamics: Characterizes the Efficiency of cycles

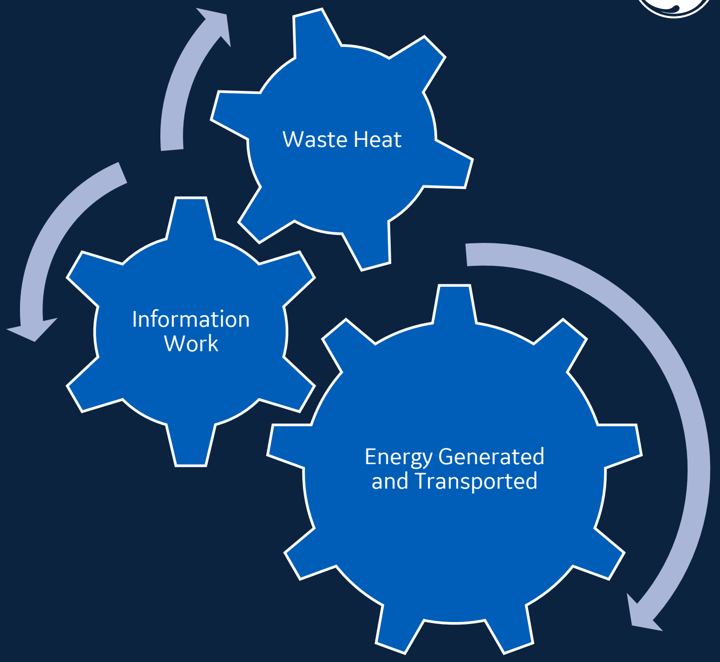
Information Theory: the Shannon Source Coding Theorem = bound-on compressibility of a data sequence.

For Machine Learning:

- Provides limits for learning
- Provides a heuristic to guide learning, e.g. Decision Trees
- Assessment of how much has been learned

Today these two worlds are optimized separately, e.g.

- Combined Cycle Power Plants – Optimal use of Waste Heat
- Renewable Powered HPC Centers to produce Carbon Credits
- CERN Compute Grid move Compute to Ide Processors



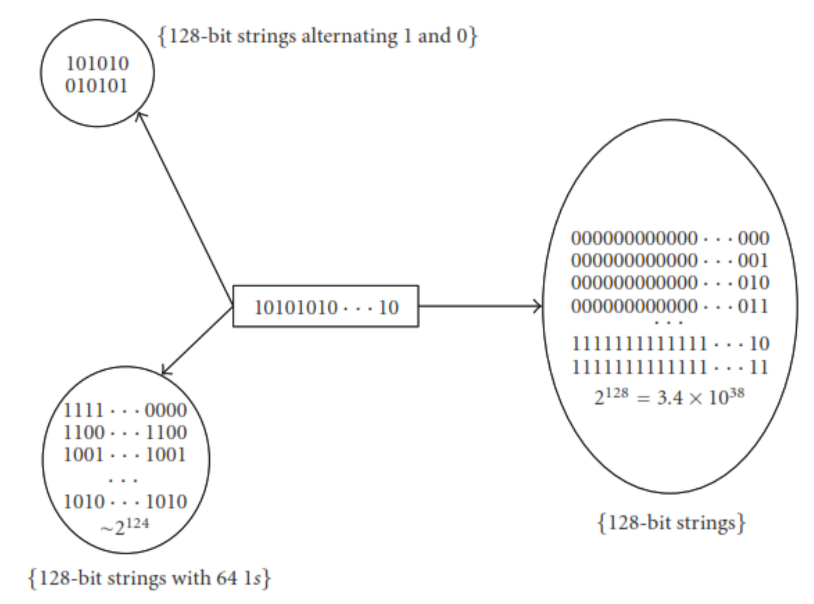
Joint Optimization: Minimize Waste Heat, Maximize Efficient Production of Energy, Maximize Learning

How Do we Assess The Energy Efficiency of Machine Learning? Compression is Learning!

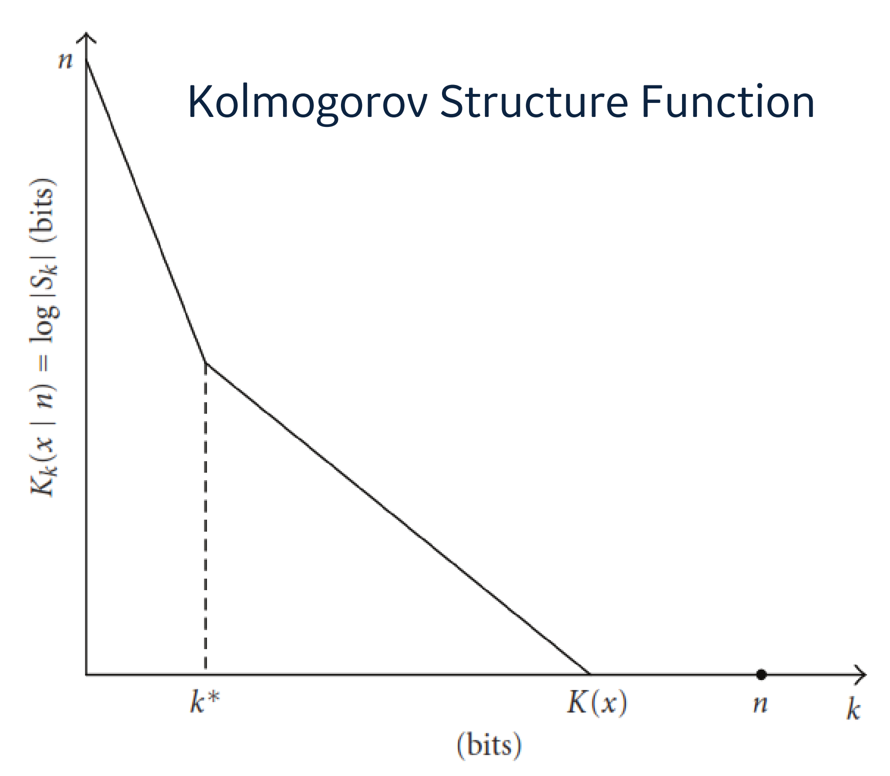
Kolmogorov Complexity is the compression bound for a data set

$$K_\varphi(x) = \left\{ \min_{\varphi(p)=x} l(p) \right\} = K_\varphi(x) \stackrel{+}{=} \{K(S) + \log_2 |S|\}$$

Consider a 128 bit string of alternating 1's and 0's



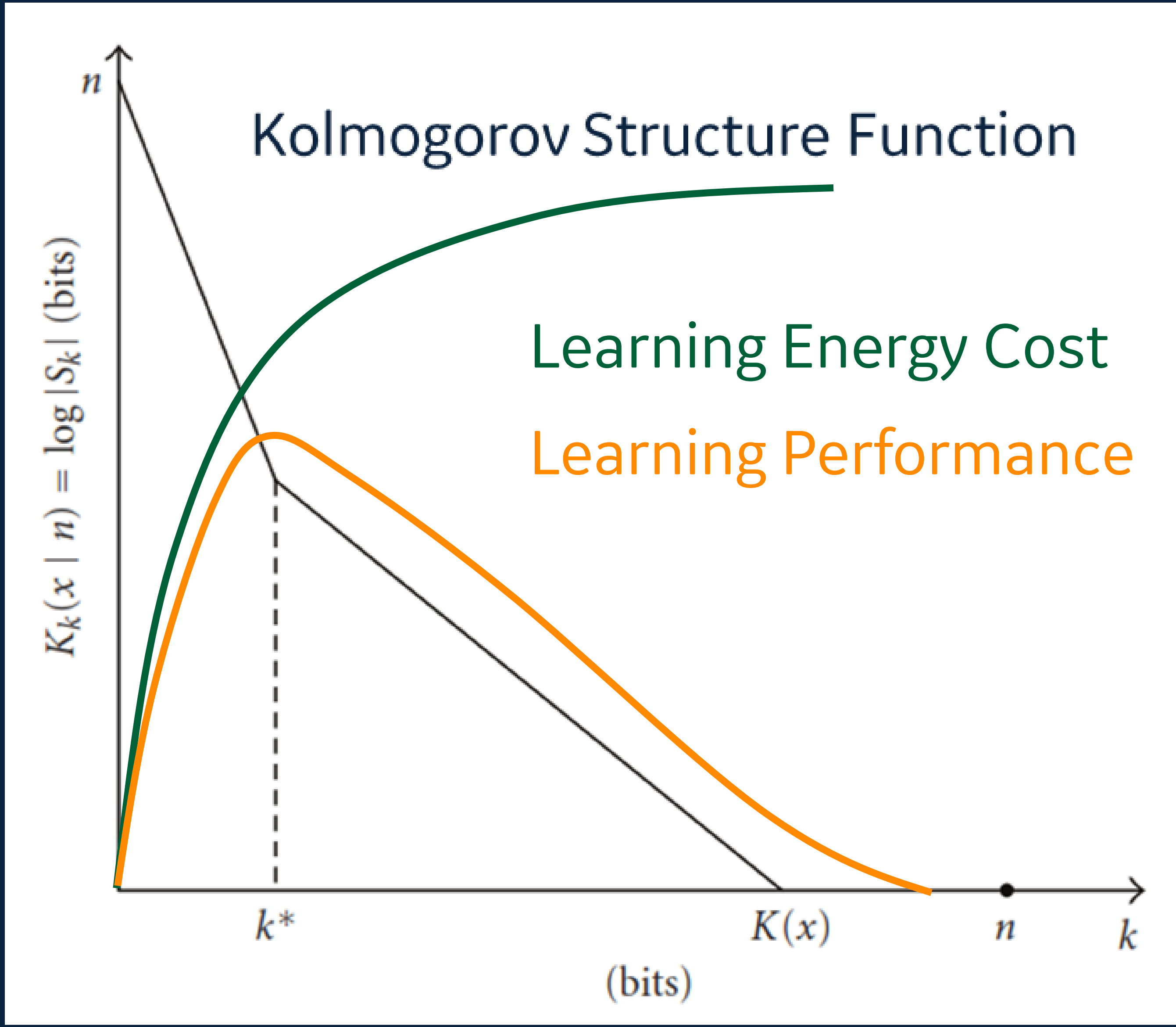
The smallest two-part representation of this string optimizes the tradeoff between Model Cost – Description of Typical Set, and Data Cost – the Cardinality of that set. This represents the Kolmogorov Minimum Sufficient Statistic



The Kolmogorov Structure Function provides a paradigm for finding optimal learning models without overfitting and by minimizing energy cost.

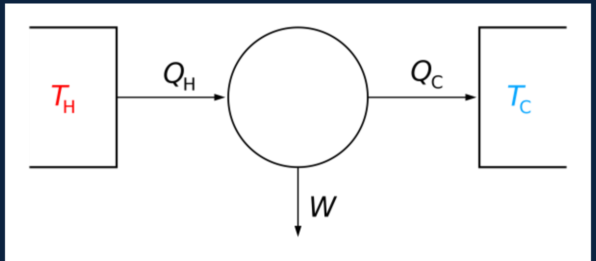
By Optimizing Entropy, we trade off Learning – Entropy Reduction, with Thermodynamic Efficiency – Entropy increase lost as waste heat

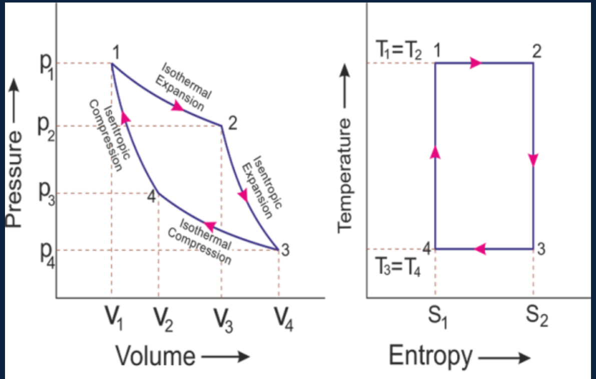
Learning structure function



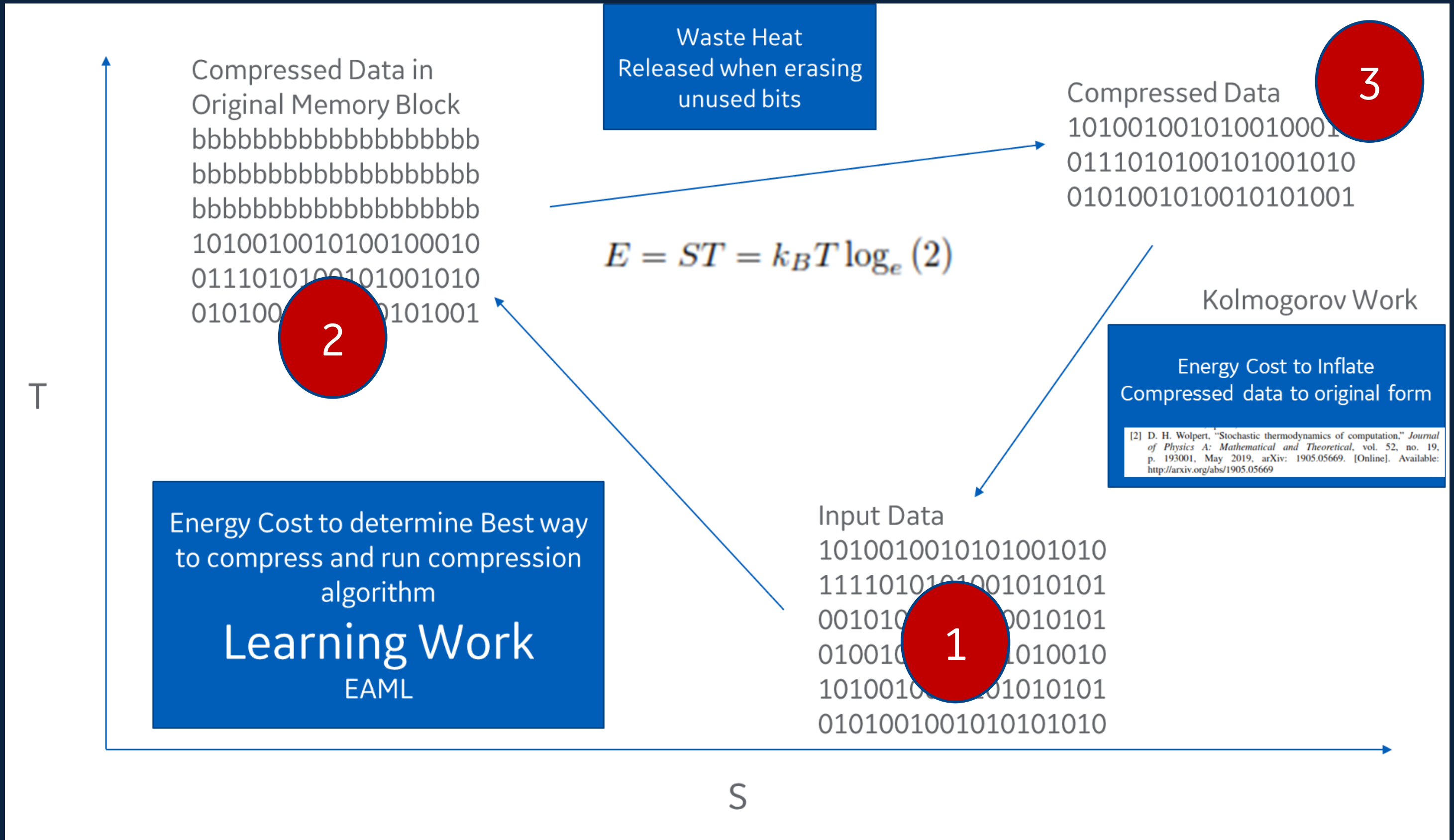
1) Equivalence to Kolmogorov Structure Function: There exists a k* that represents the optimal model complexity for optimal model generalizability
2) Beyond k*, the model is overfitting, and Learning Performance will Drop
3) For EAML, we want to go as close to k* **within Energy Constraints**

Energy Cost to Create an AI model





(a) p-v diagram (b) T-S diagram



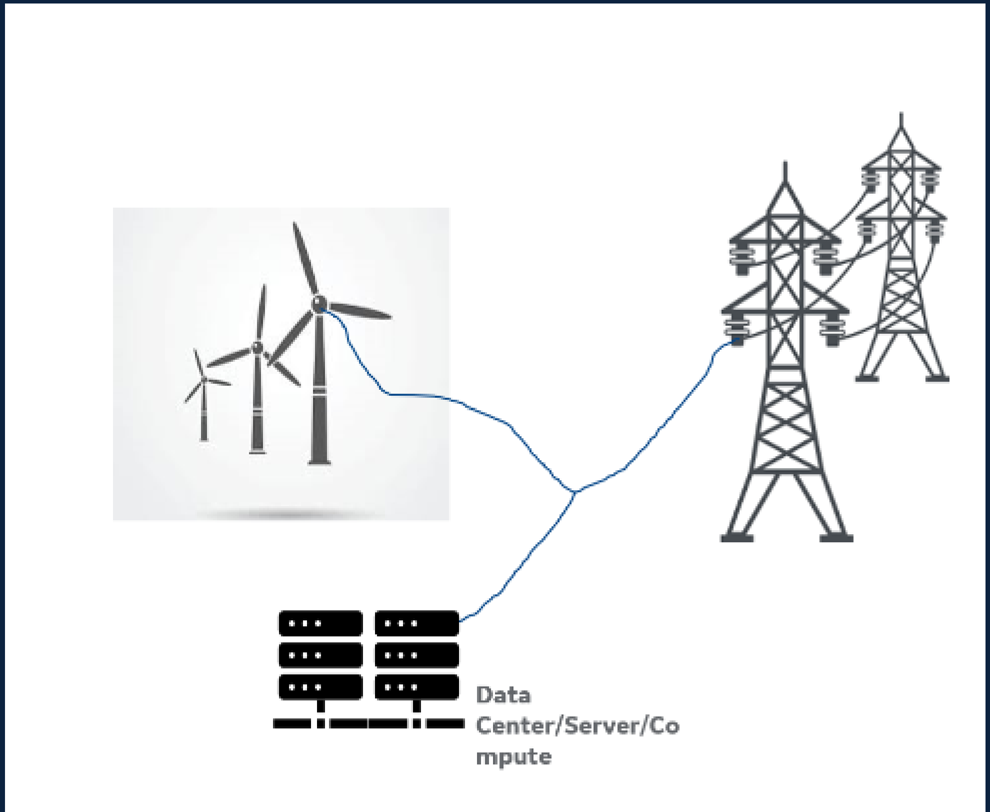
Carnot Engine: Efficiency of Heat Engine a function of Entropy Change at a given Temperature

Proposed "Kolmogorov Engine" Cycle: Wolpert introduces the concept of "Kolmogorov Work" = the energy required for a Turing Machine Execute a program of size K(X) to produce X.

We expand this "Kolmogorov Learning Cycle" to include the Energy Cost of learning model (the small program)

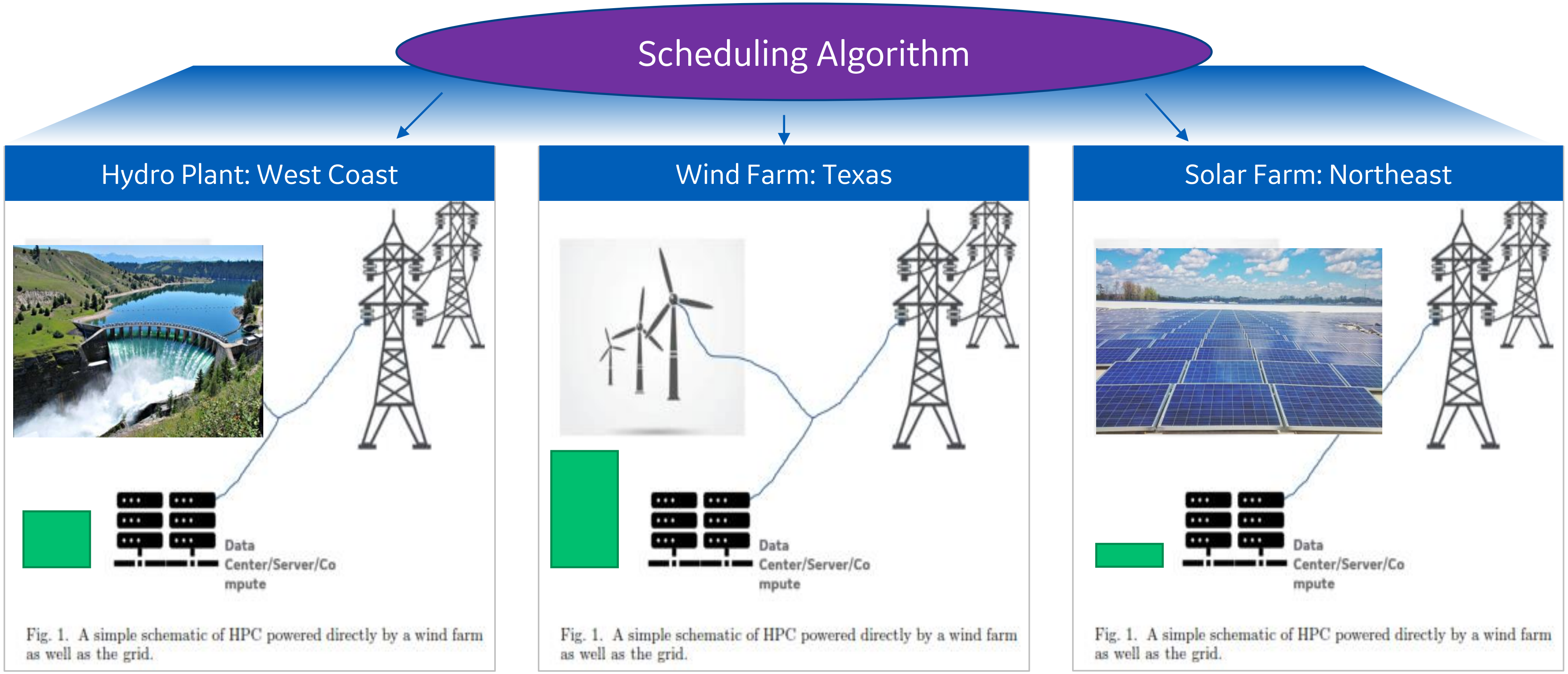
HPC in a wind farm: Example of entropy economy

1) Many wind farm curtail power to avoid overloading power grids, and/or limit size to match grid entitlement
2) This leads to loss of valuable "free" energy .
3) If HPC installed within wind farm, effectively use the “free” energy:
    1) Intensive “information work” when “free” energy available.
    2) Light “information work” when “free” energy not available.
4) Jointly optimize the common denominator: **Entropy.**



Matching HPC Load to Renewable Power Availability Optimizes Entropy Economy

Distributed HPC's Amplifies Opportunity to Optimize Entropy Economy



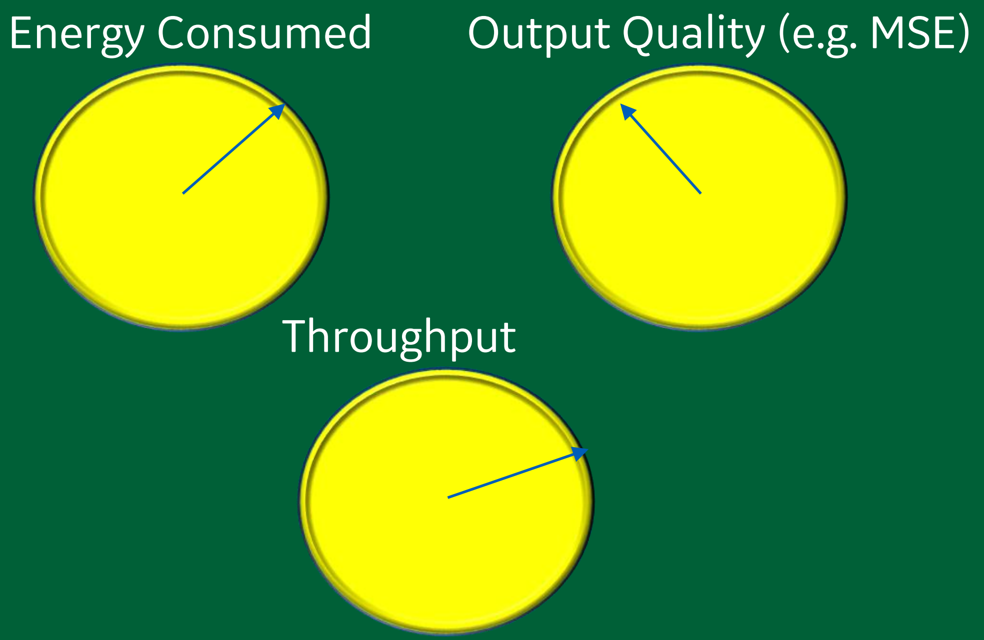
Entropy Economy: Managing the Precious Resource of Entropy Flow (Jointly Optimize Energy Production, Information Work and Disposition of Waste Heat)

REDUCE CARBON BY MATCHING INFORMATION LOAD / RENEWABLE POWER

Energy Aware Machine Learning

What is Energy Aware Machine Learning?

High Performance Compute Engine
(EAML Gives 3 new knobs that do not exist today)
How much of your 1MW Compute Engine Do you want to use and how do you want to use it?



Adjust Any Two Knobs to Determine Third

This requires algorithms that can:

1) Trade Energy Consumption for Quality of Output through:
   - Bit Quantization
   - Number of Tree's/Threads/Bootstraps
   - Dimensionality Reduction
   - Learning Rate
2) Adapt to a given Energy Profile
3) Provide Predicted Energy Entitlement for desired quality and throughput

Ultimate Smart Load to Stabilize Grid, Maximize Renewables, Reduce Carbon. Key Challenge: Algorithms!

Entropy Economy: Practical Scenarios

Case 1:

Wind Farm Produces 500MWH
(100MWH are lost due to curtailment)

Gas Generator Produces 500MWH
(When Wind Not Blowing)

50MWH are lost due to I2R Losses

HPC Data Center Uses Fixed Energy Profile and Consumes 500MWH to produce 5 Machine Learning Models with Output Quality in Spec

50 Tons of Carbon are produces

Case 2:

Wind Farm Produces 600MWH
(HPC Load Moved to Consume Excess Power Previously Curtailed)

Gas Generator Produces 300MWH
(When Wind Not Blowing)

HPC Data Center Uses Variable Energy Profile and Consumes 400MWH to produce 7 Machine Learning Models with Output Quality in Spec

20 tons of carbon are produced

Case 2 illustrates a practical example of Entropy Economy driven by EAML

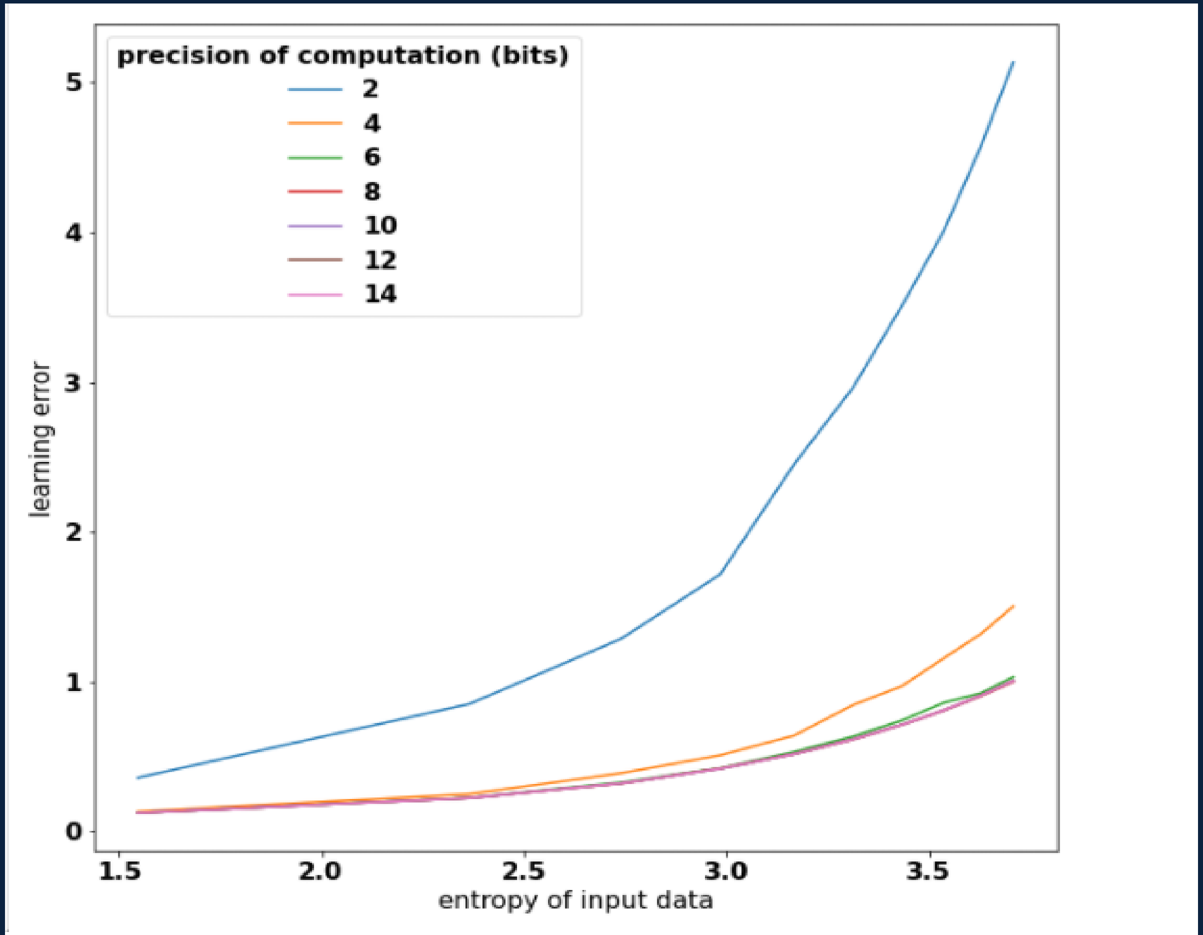
Experiments 1

Problem:
Computational resources required for a simple learning problem (regression) ?

Objective:
Can we effectively use low precision (energy) computation for low entropy data?

Conclusion:

1) For very low precision (2 or 4 bit) computers, learning error increases sharply with increase in input entropy.
2) For higher precision (6-10 bits) computers, learning error increases moderately with increase in input entropy



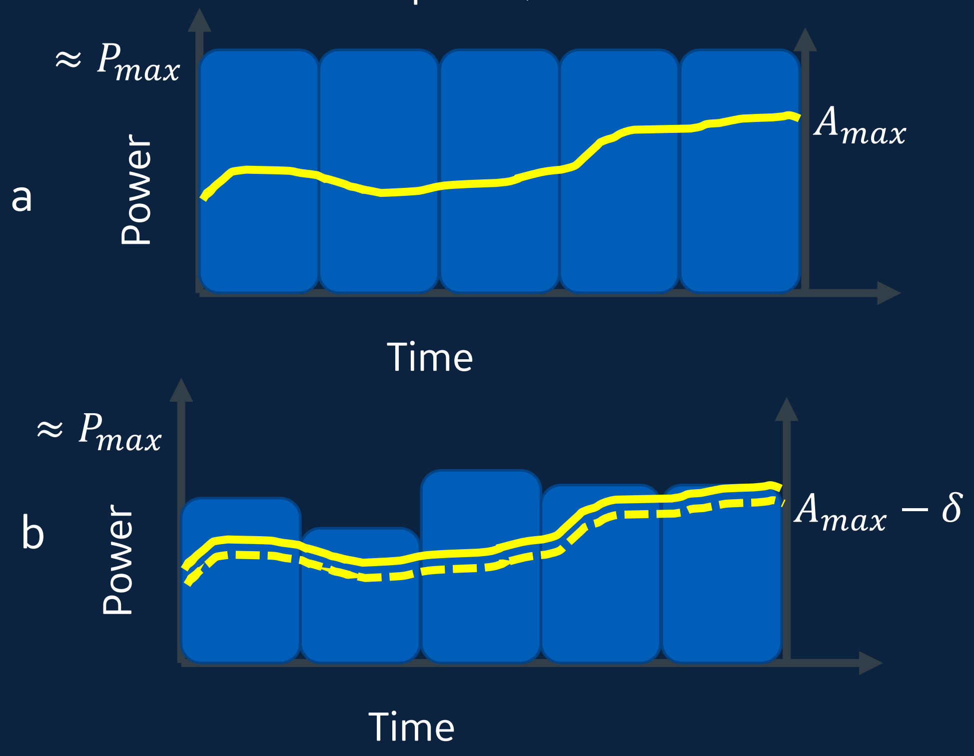
Allocate lower entropy data random data on lower resource computers?

Conclusion

Future work

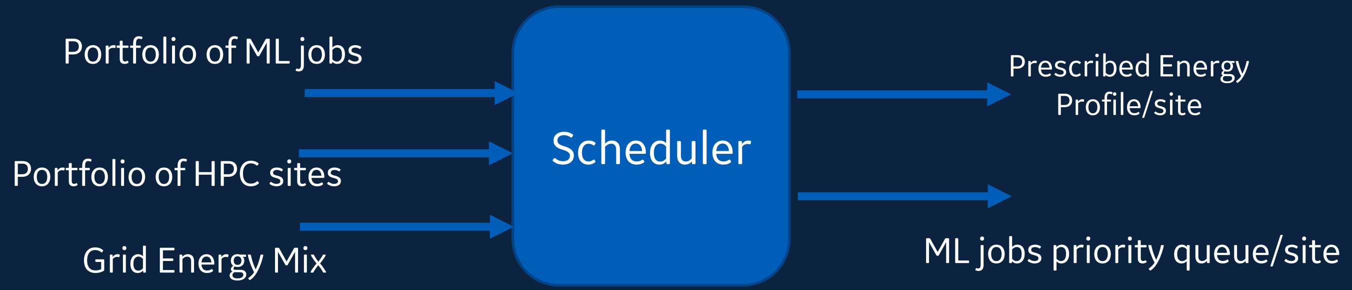
Problem 1

Can we achieve similar ML quality performance with reduced and variable input energy profile (e.g. wind power) ?



Problem 2

How do we schedule a portfolio of ML jobs to get "optimal" performance and provide load balancing to the grid ?



Hypothesis

1) **Smart Quantization/Dimension reduction**
2) **Estimation of Performance prediction/KWh using historical data**

The New Paradigm: From Energy Economy to Entropy Economy

FROM	TO
Spinning Reserve of (Dumb) Generation Capacity	Spinning Reserve of Information Work Jobs, Movable Across Gid
Machine Learning Algorithms the Do Not Consider Energy Costs	EAML algorithms that act to Optimize and Stabilize Grid while maximizing Algorithm Quality of Service for Energy Cost
Stand Alone HPC's	Integrated Network of HPC's leveraging Stranded Power
Efficiency of Computation not considered	tCO2e / Cost Benefit of Computation
High I2R losses and Congested Grid	Grid Optimized and Stabilized through HPC network
Renewable Projects Limited by Grid Capacity	Renewable Projects optimized for Wind/Solar/Hydro Resource Available
Carbon Credits Giving the Right to Pollute	Joint Optimization of Energy and AI Reducing total Carbon to Generate Compute Work

Thank You!

Building a world that works

evans@ge.com