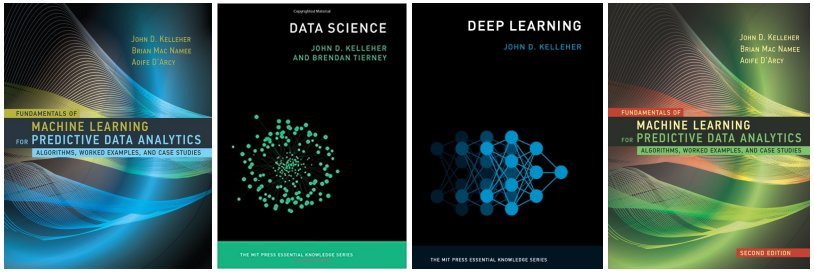
Green AI:
Reducing the Environmental Cost of AI

Prof. John D. Kelleher

john.d.kelleher@tudublin.ie

@johdkelleher

OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
TECHNOLOGICAL UNIVERSITY DUBLIN



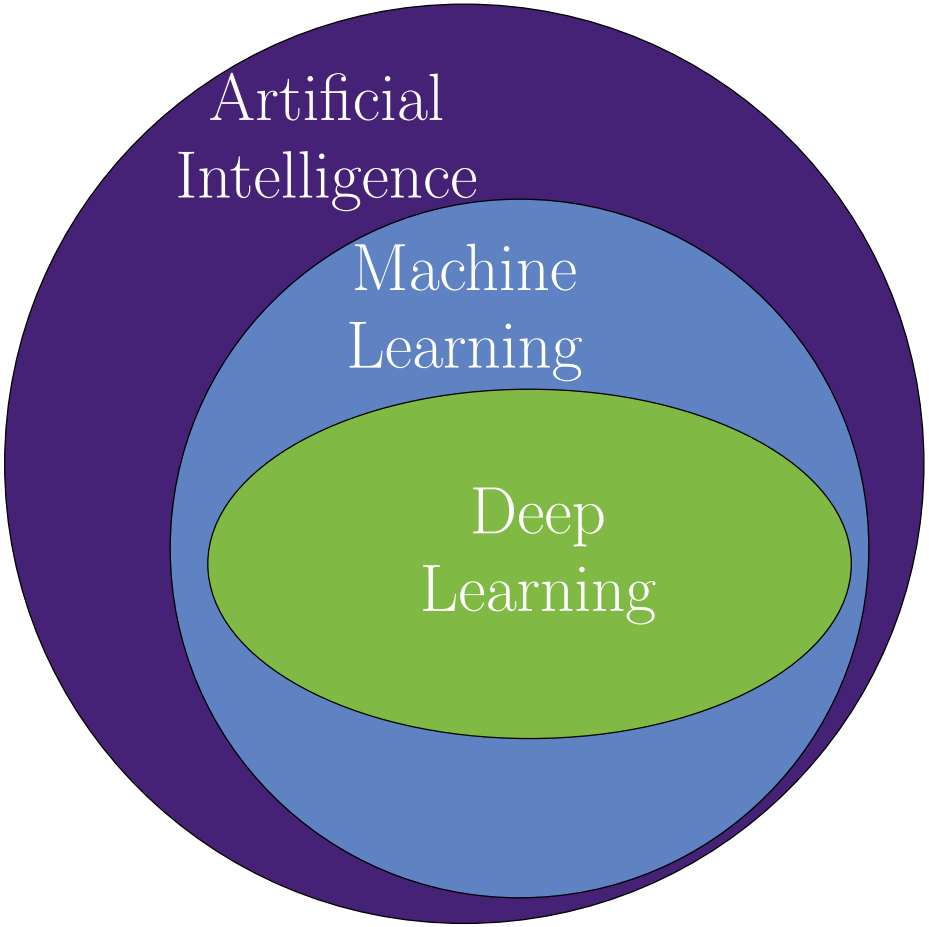
ADAPT Engaging Content Engaging People

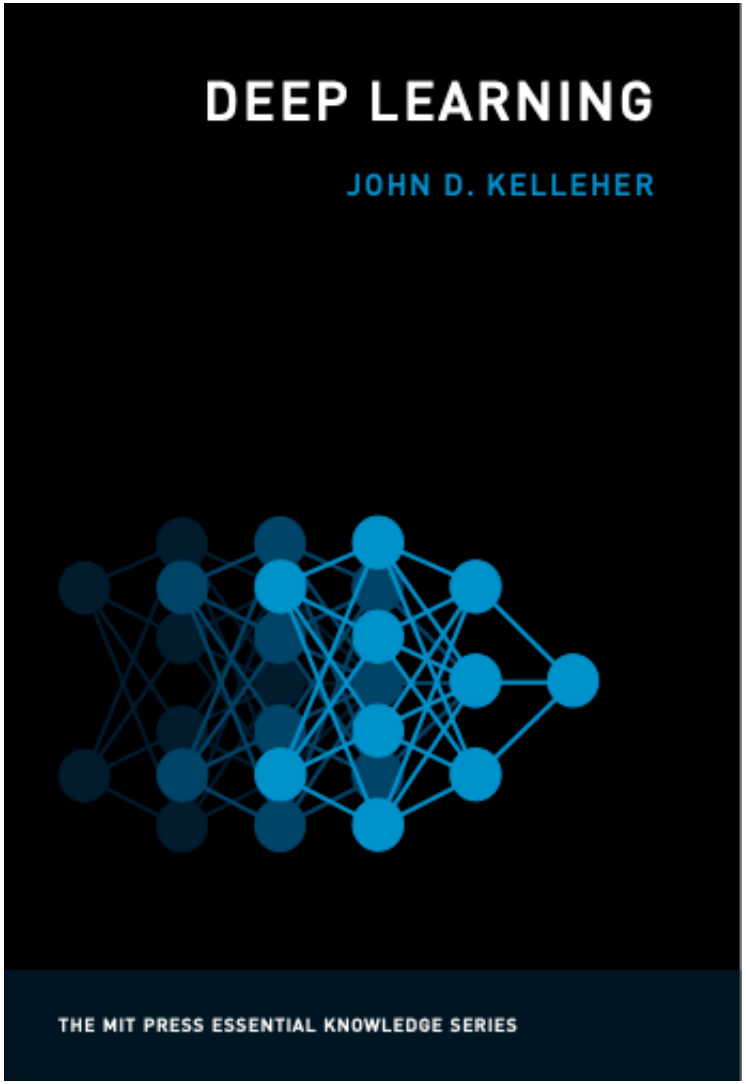
d-real SFI centre for research training in digitally-enhanced reality

PRECISE4Q PREDICTIVE MODELLING IN STROKE

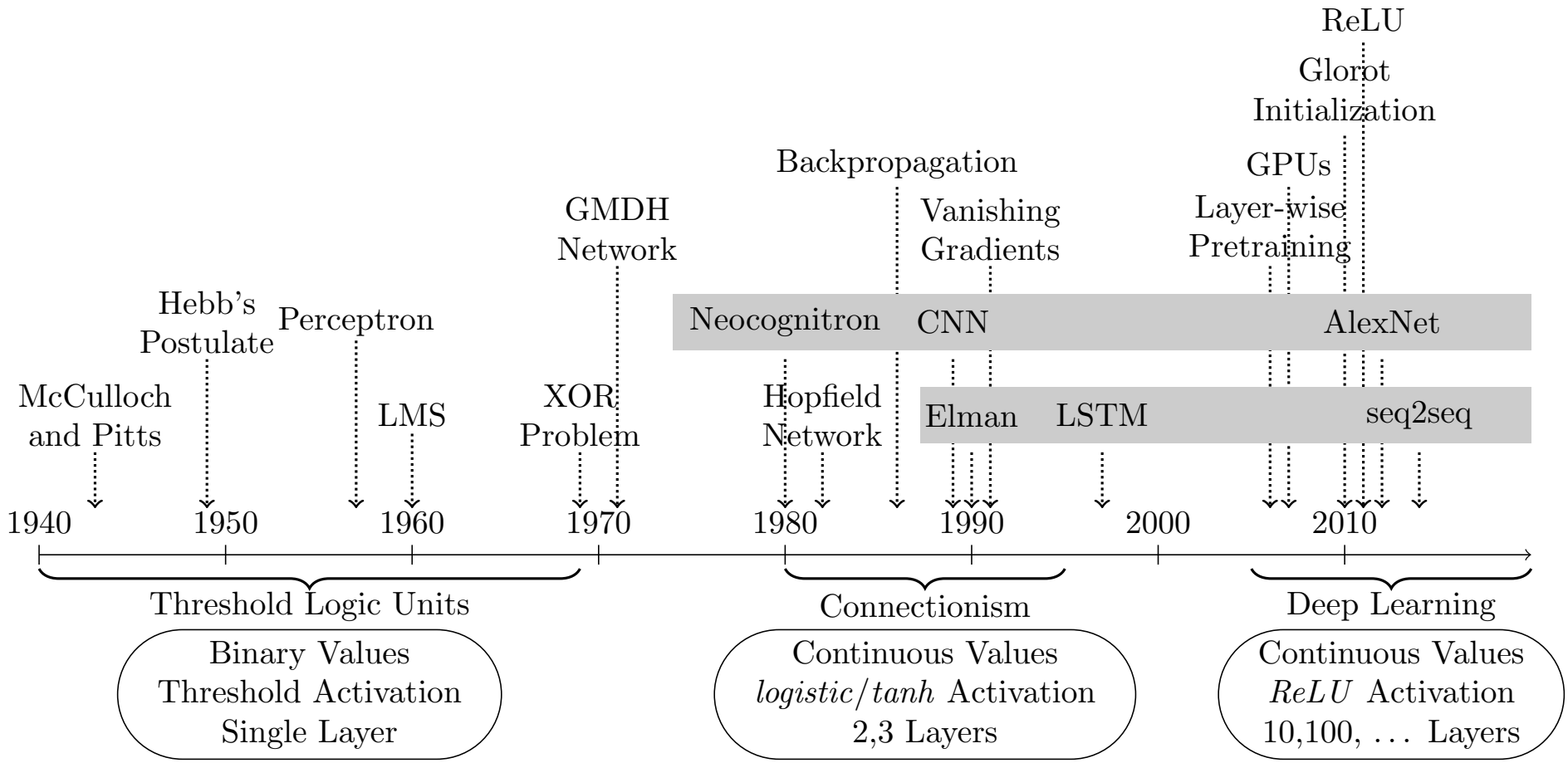
Real Smart Cities

Artificial Intelligence 2010-2020





Artificial Intelligence 2010-2020



Deep Learning. John D. Kelleher. MIT Press. 2019.

Artificial Intelligence 2010-2020

EXPERT OPINION

Contact Editor: **Brian Brannon,** bbrannon@computer.org

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

Eugene Wigner's article "The Unreasonable Effectiveness of Mathematics in the Natural Sciences"[1] examines why so much of physics can be neatly explained with simple mathematical formulas such as $f = ma$ or $e = mc^2$. Meanwhile, sciences that involve human beings rather than elementary particles have proven more resistant to elegant mathematics. Economists suffer from physics envy over their inability to neatly model human behavior. An informal, incomplete grammar of the English language runs over 1,700 pages.[2] Perhaps when it comes to natural language processing and related fields, we're doomed to complex theories that will never have the elegance of physics equations. But if that's so, we should stop acting as if our goal is to author extremely elegant theories, and instead embrace complexity and make use of the best ally we have: the unreasonable effectiveness of data.

One of us, as an undergraduate at Brown University, remembers the excitement of having access to the Brown Corpus, containing one million English words.[3] Since then, our field has seen several notable corpora that are about 100 times larger, and in 2006, Google released a trillion-word corpus with frequency counts for all sequences up to five words long.[4] In some ways this corpus is a step backwards from the Brown Corpus: it's taken from unfiltered Web pages and thus contains incomplete sentences, spelling errors, grammatical errors, and all sorts of other errors. It's not annotated with carefully hand-corrected part-of-speech tags. But the fact that it's a million times larger than the Brown Corpus outweighs these drawbacks. A trillion-word corpus—along with other Web-derived corpora of millions, billions, or trillions of links, videos, images, tables, and user interactions—captures even very rare aspects of human behavior. So, this corpus could serve as the basis of a complete model for certain tasks—if only we knew how to extract the model from the data.

Learning from Text at Web Scale

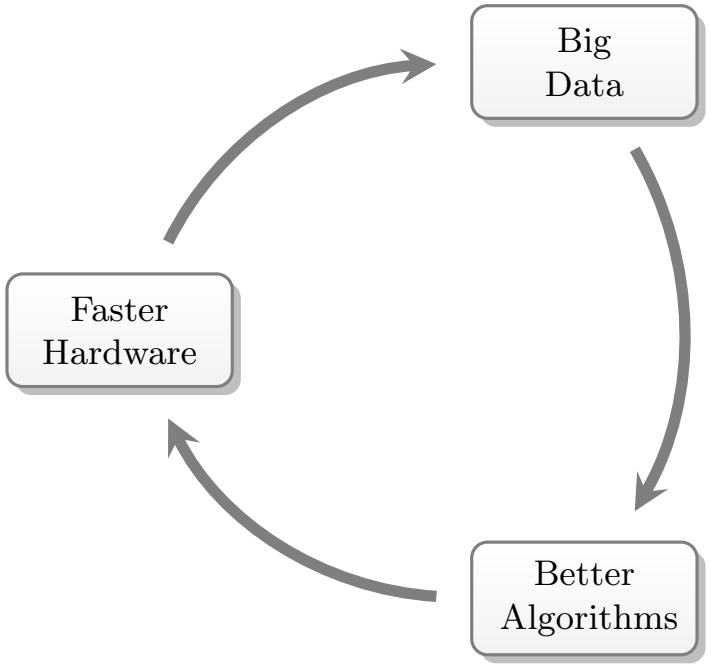
The biggest successes in natural-language-related machine learning have been statistical speech recognition and statistical machine translation. The reason for these successes is not that these tasks are easier than other tasks; they are in fact much harder than tasks such as document classification that extract just a few bits of information from each document. The reason is that translation is a natural task routinely done every day for a real human need (think of the operations of the European Union or of news agencies). The same is true of speech transcription (think of closed-caption broadcasts). In other words, a large training set of the input-output behavior that we seek to automate is available to us *in the wild*. In contrast, traditional natural language processing problems such as document classification, part-of-speech tagging, named-entity recognition, or parsing are not routine tasks, so they have no large corpus available in the wild. Instead, a corpus for these tasks requires skilled human annotation. Such annotation is not only slow and expensive to acquire but also difficult for experts to agree on, being bedeviled by many of the difficulties we discuss later in relation to the Semantic Web. The first lesson of Web-scale learning is to use available large-scale data rather than hoping for annotated data that isn't available. For instance, we find that useful semantic relationships can be automatically learned from the statistics of search queries and the corresponding results[5] or from the accumulated evidence of Web-based text patterns and formatted tables,[6] in both cases without needing any manually annotated data.

1541-1672/09/$25.00 © 2009 IEEE
Published by the IEEE Computer Society

invariably, ***simple models and a lot of data trump*** *more elaborate models based on less data*

- Excellence = accuracy
- Simple != small
- Simple != easy to understand

Artificial Intelligence 2010-2020

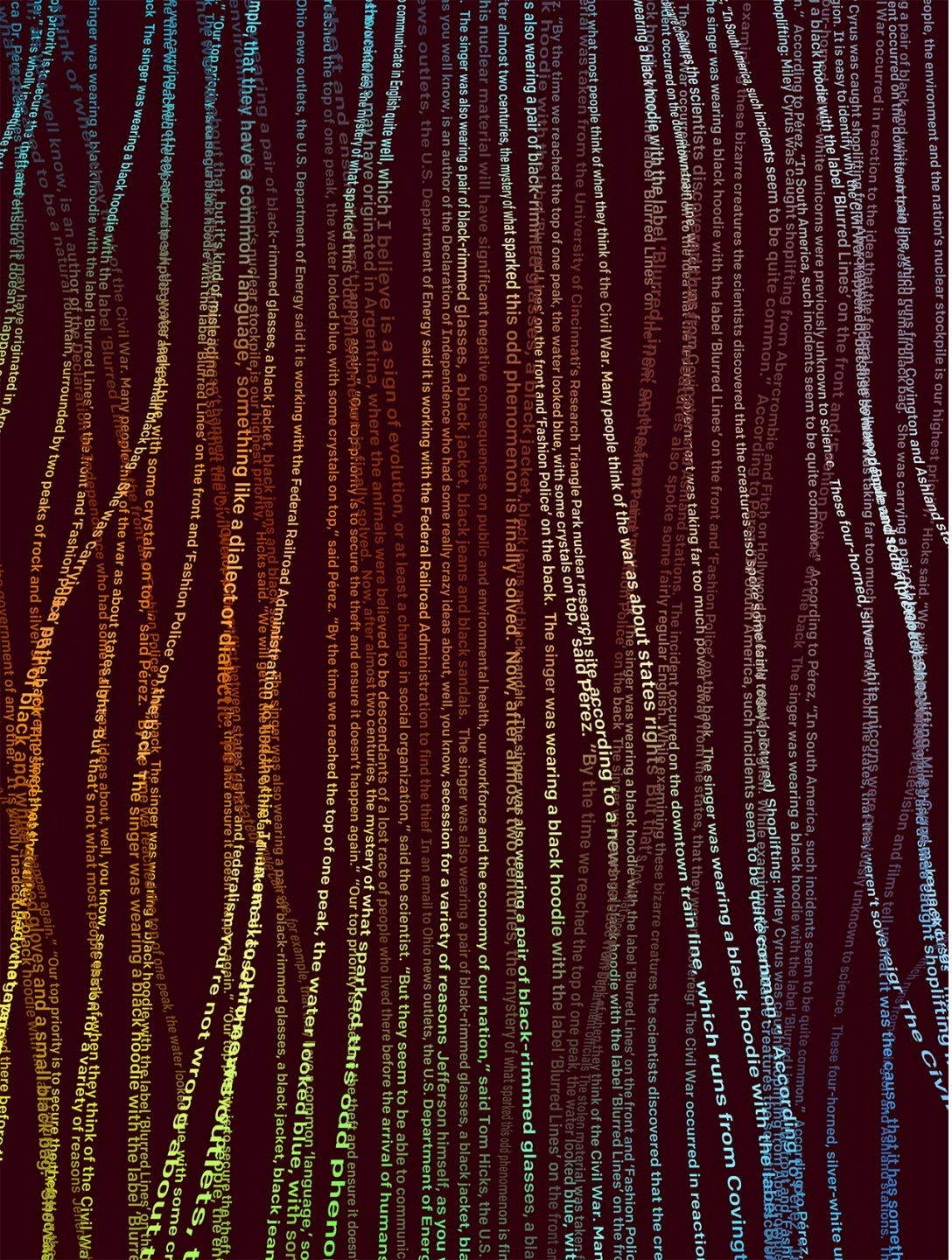


- Data centres consume approximately 2% of electricity worldwide, and this is set to reach 8% by 2030
- In Ireland EirGrid estimates that by 2028 data centres and other large tech users will consume nearly 30 per cent of Ireland's electricity.

Estimated CO_2 Emissions

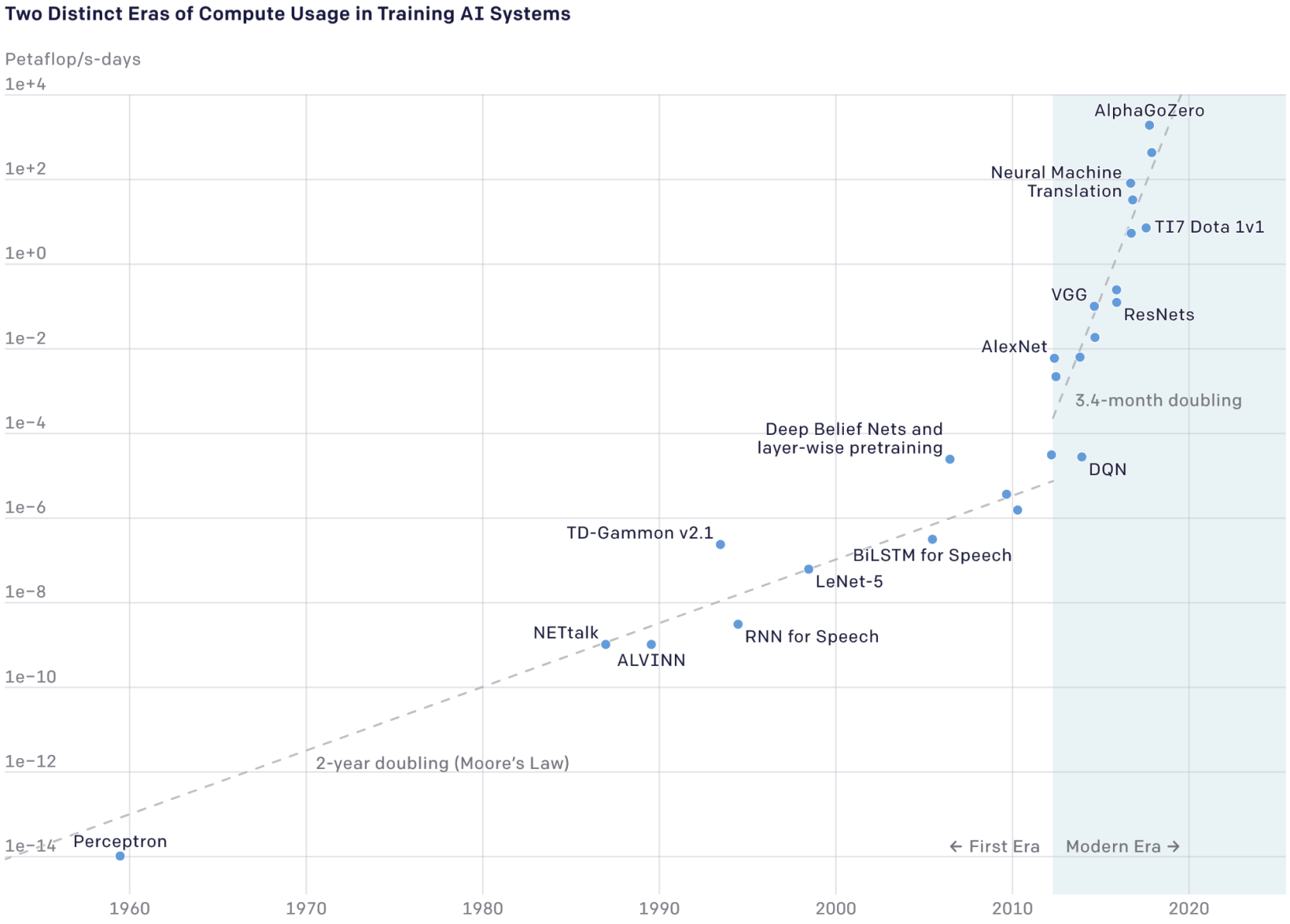
Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY <-> SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg. incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
with tuning and experimentation	78,468
Transformer (big)	192
with neural architecture search	626,155

Energy and Policy Considerations for Deep Learning in NLP. Emma Strubell, Ananya Ganesh, and Andrew McCallum. Proceedings of the 57th Annual Meeting of the ACL, 2019.



OpenAI

- GPT-3 5th June 2020
- 175 billion parameters
- 499 billion tokens
- 3.14^{23} FLOPS
- 355 years on Tesla V100
- $4.6m using cloud GPUs
- 700GB memory in FP32



Dario Amodei & Danny Hernandez, et al. "AI and Compute," OpenAI, May 16, 2018. https://openai.com/blog/ai-and-compute/ (accessed May 15, 2021)

Sustainable AI

Metrics

Precision

Sparsity

Metrics of Energy Efficiency

- Electricity Usage
- Elapsed Time
- Number of model parameters
- Floating Point Operations (FLOPs)

For more on metrics see:

- *Green AI.* Roy Schwartz et al., 2019. arXiv:1907.10597v3
- *Efficient Processing of Deep Neural Networks. Vivienne* Sze et al., 2020. Morgan Claypool.

Metrics of Energy Efficiency

Operation	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.10
16b FP Add	0.40
32b FP Add	0.90
8b Multiply	0.20
32b Multiply	3.10
16b FP Multiply	1.10
32b FP Multiple	3.70
32b SRAM Read (8kb)	10.00
32b DRAM Read	640.00

Energy consumption for a range of operations, adapted from: Computing's energy problem (and what we can do about it), M. Horowitz, 2014. doi: 10.1109/ISSCC.2014.6757323

Precision

- a measure of the detail in which the quantity is expressed, usually measured in bits or decimal digits.

- π (32 places) 3.1415926535897932384626433832795 0
- π (16 places) 3.1415926535897932

- Measurement precision is not a detail
- It links to the **energy cost of a calculation.**

Precision

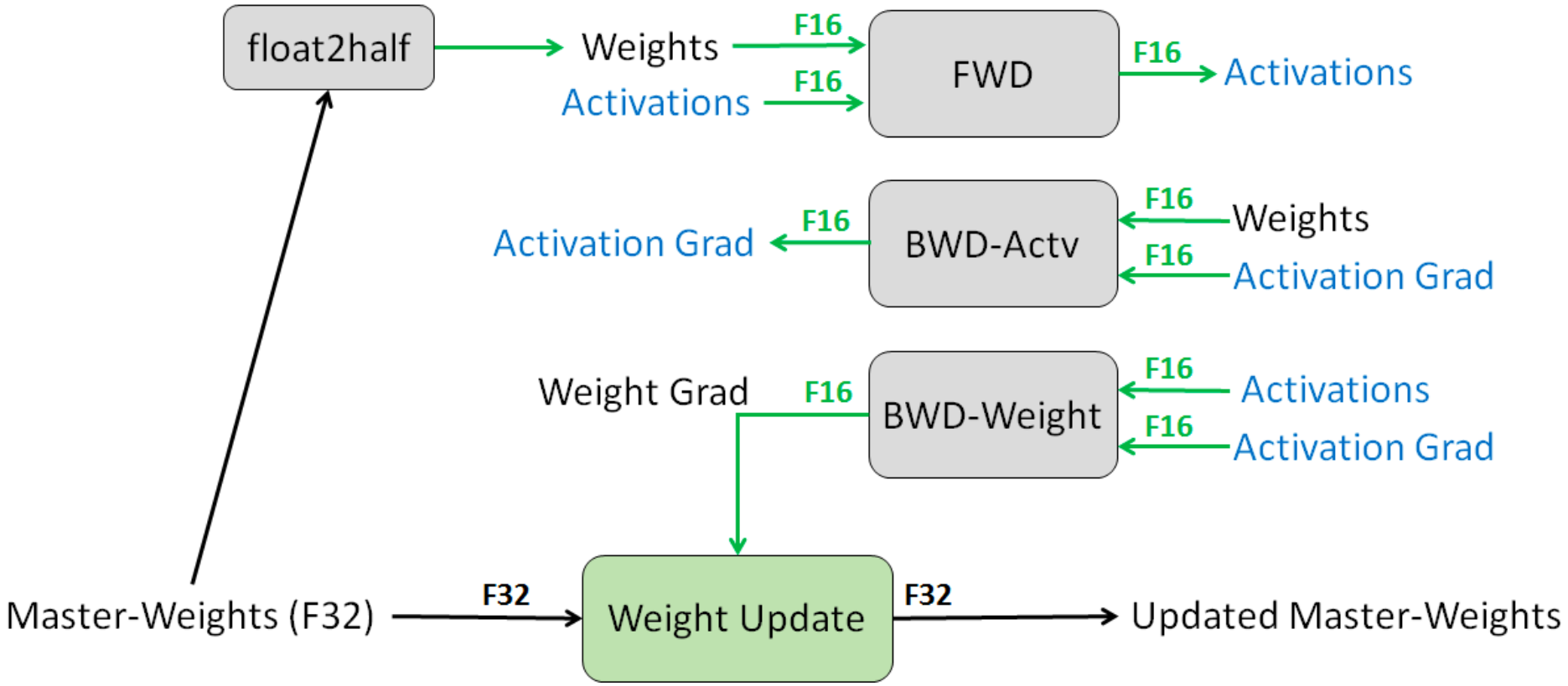
Operation	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.10
16b FP Add	0.40
32b FP Add	0.90
8b Multiply	0.20
32b Multiply	3.10
16b FP Multiply	1.10
32b FP Multiple	3.70
32b SRAM Read (8kb)	10.00
32b DRAM Read	640.00

- The energy consumed by an operation is dependent on the precision of the operation

Precision

- Reducing precision for weights/activations can reduces the energy expended on:
  - data movement
  - MAC operations
- The challenge here is to reduce precision while maintaining accuracy

Precision



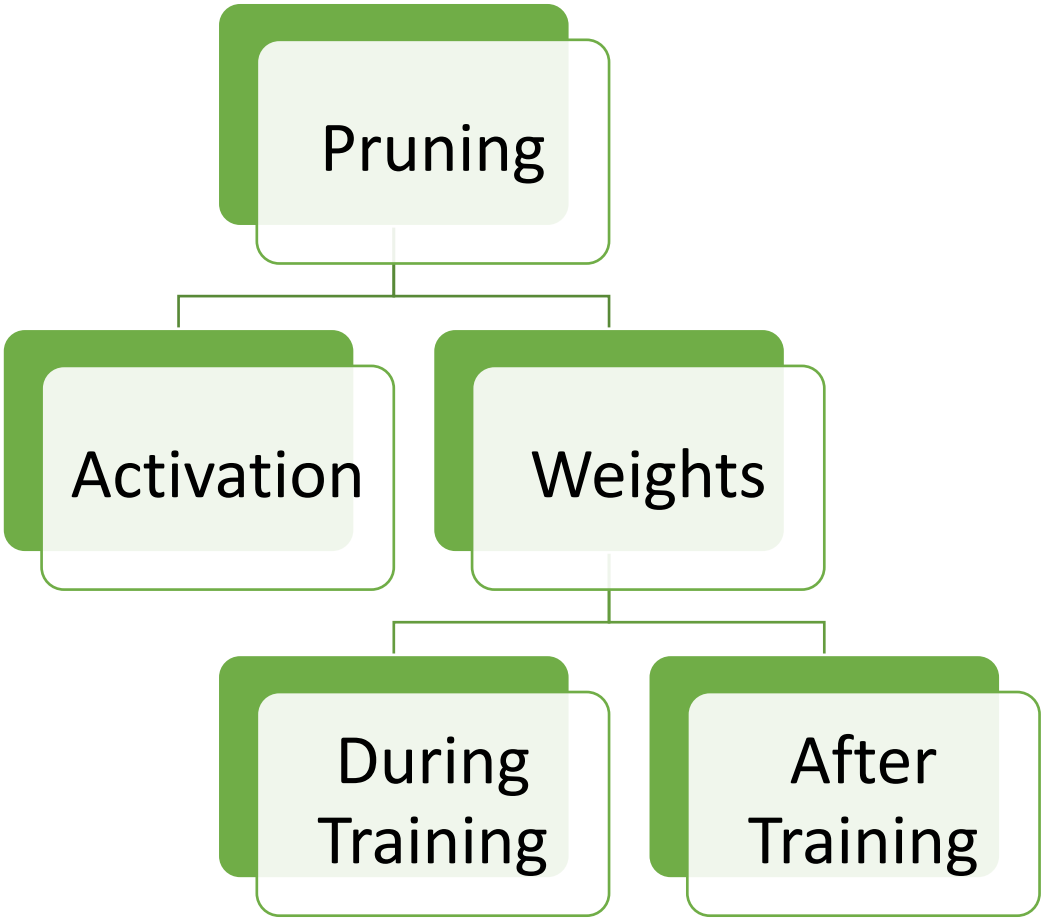
Mixed Precision Training. Sharan Narang, Paulius Micikevicius, et al. In Proceedings of the International Conference on Learning Representations (ICLR), 2018.

Sparsity

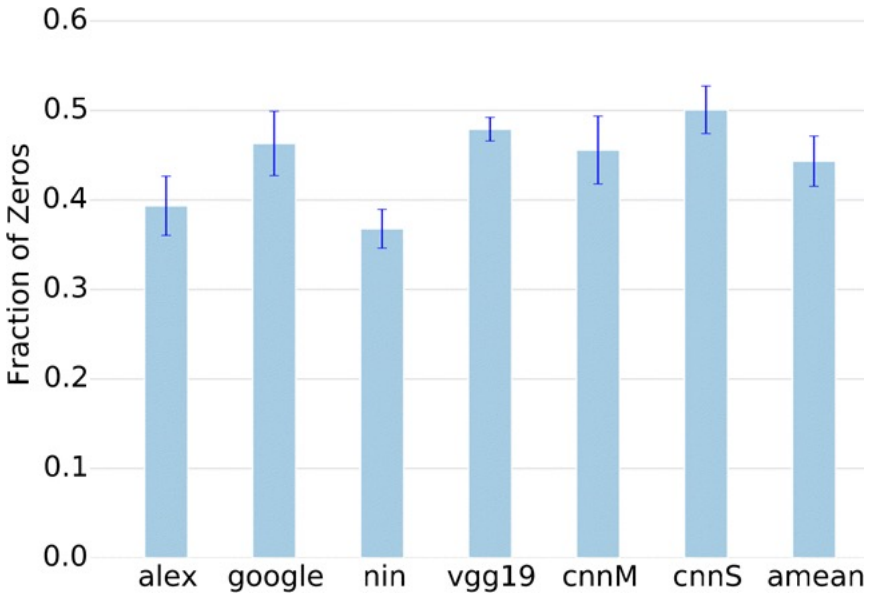
- Sparse data: many repeated values (often 0)

- Benefits of sparse data (weight/activations):
  - Reduce memory footprint
  - Reduce number of MACs

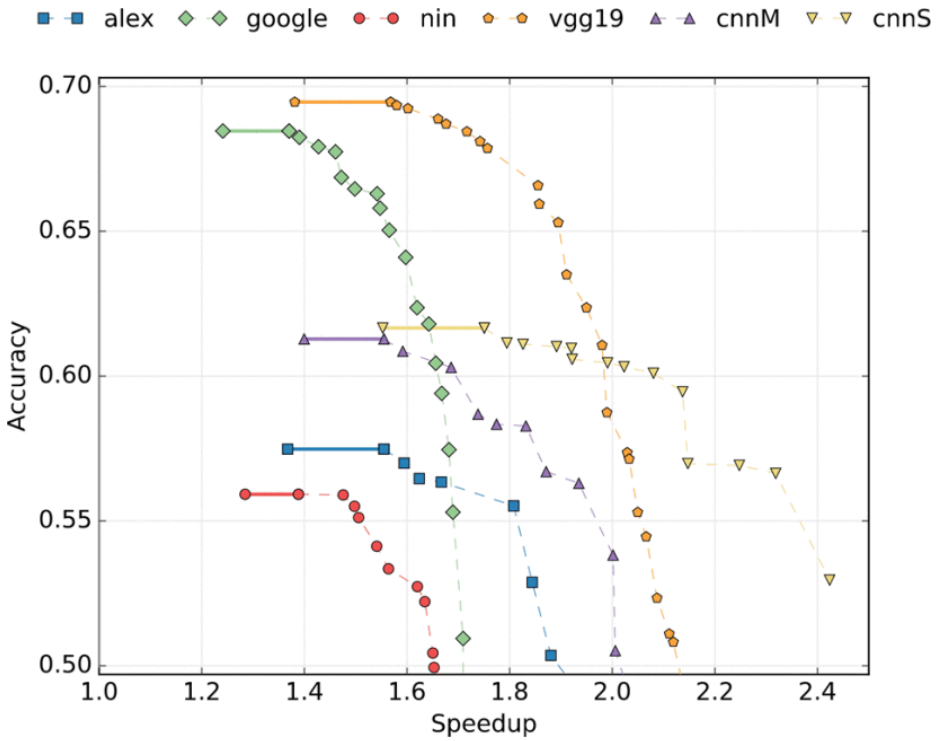
Sparsity


Pruning
Activation
Weights
During Training
After Training

Sparsity: Pruning Activations



Avg. fraction of convolution layer multiplication input values that are zero



Trade-off between accuracy and speedup as more neurons are set to zero

Cnvlutin: Ineffectual-Neural-Free Deep Neural Network Computing. Jorge Albericio et al. 2016. doi: 10.1109/isca.2016.11

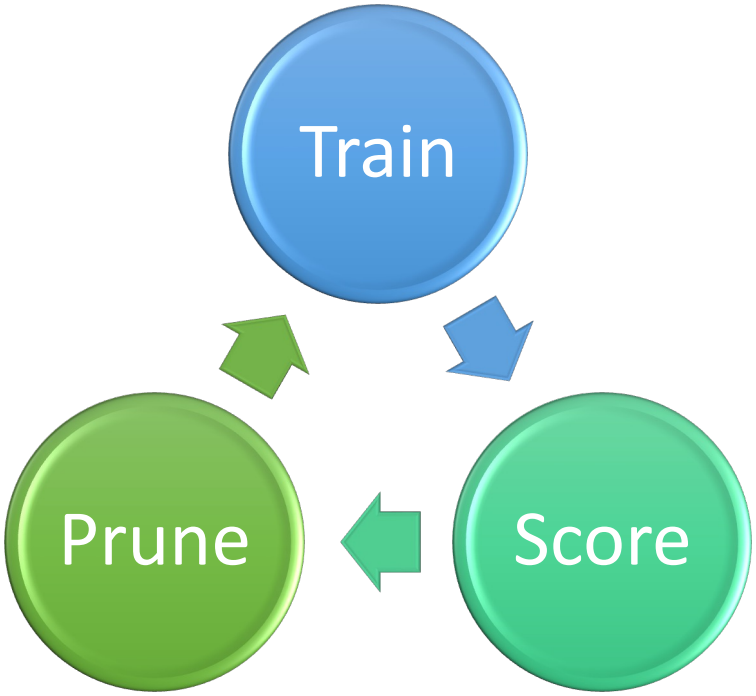
Sparsity: Pruning Weights

Weight Pruning

The majority of weight pruning approaches prune weights that have small magnitudes (this is known as *magnitude-based pruning*)

- **Unstructured** (fine-grained) pruning, individual weights
- **Structured** (coarse-grained) pruning, groups of weights (channels, filters, and so on)

Sparsity: Weight Pruning (during training)



- Prune a small fraction of convolutional filters at each epoch of training until the target pruning rate is achieved

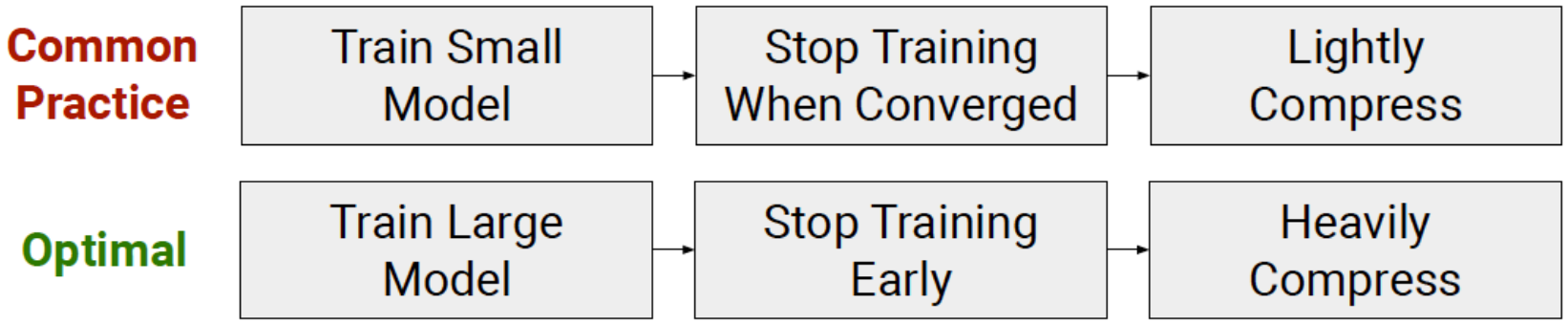
Pruning filters while Training for Efficiently Optimizing Deep Learning Networks. Sourjya Roy et al., 2020. arXiv: 2003.02800v1

Sparsity: Weight Pruning (during training)

- Three methods for choosing filter to prune:
  - L1 Normalization
  - Mean activation
  - Random
- Results indicate that:
  - L1 Normalization Pruning is best
  - Gradual pruning of filters during training enables successive epochs to compensate for any accuracy loss
  - Almost no-drop in accuracy on CIFAR10 after pruning ~80% of filters

Pruning filters while Training for Efficiently Optimizing Deep Learning Networks. Sourjya Roy et al., 2020. arXiv: 2003.02800v1

Sparsity: Weight Pruning (after training)



Train Large, Then Compress: Rethinking Model Size for Efficient Training and Inference of Transformers. Zhuohan Li et al., 2020. arXiv: 2002.11794

Sustainable AI: if data is the new oil, AI is the new CO_2

SOTA	When accuracy is the sole metric of progress there is an incentive to make ever larger models
CO_2	The environmental impact of this approach to research makes it unsustainable
Access	It excludes researchers who do not have access to super-computers
Small Data	Low resource scenarios are marginalised e.g. low resource languages

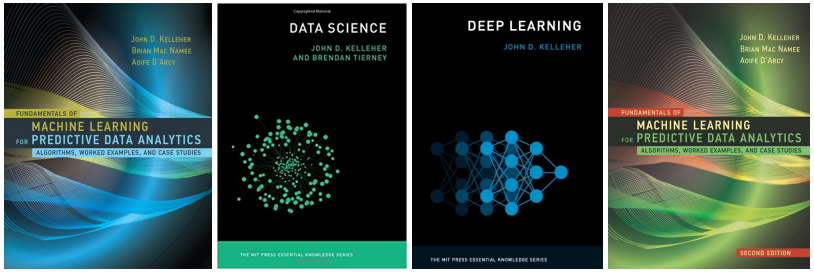
Beyond Accuracy: Reducing the Environmental Cost of AI

Prof. John D. Kelleher

john.d.kelleher@tudublin.ie

@johdkelleher

OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
TECHNOLOGICAL UNIVERSITY DUBLIN



ADAPT Engaging Content Engaging People

d-real SFI centre for research training in digitally-enhanced reality

PRECISE4Q PREDICTIVE MODELLING IN STROKE

Real Smart Cities