Advanced Natural Language Processing (ANLP)

Lecture 10: Sustainability

Gabriel Stanovsky & **Roy Schwartz**

So far

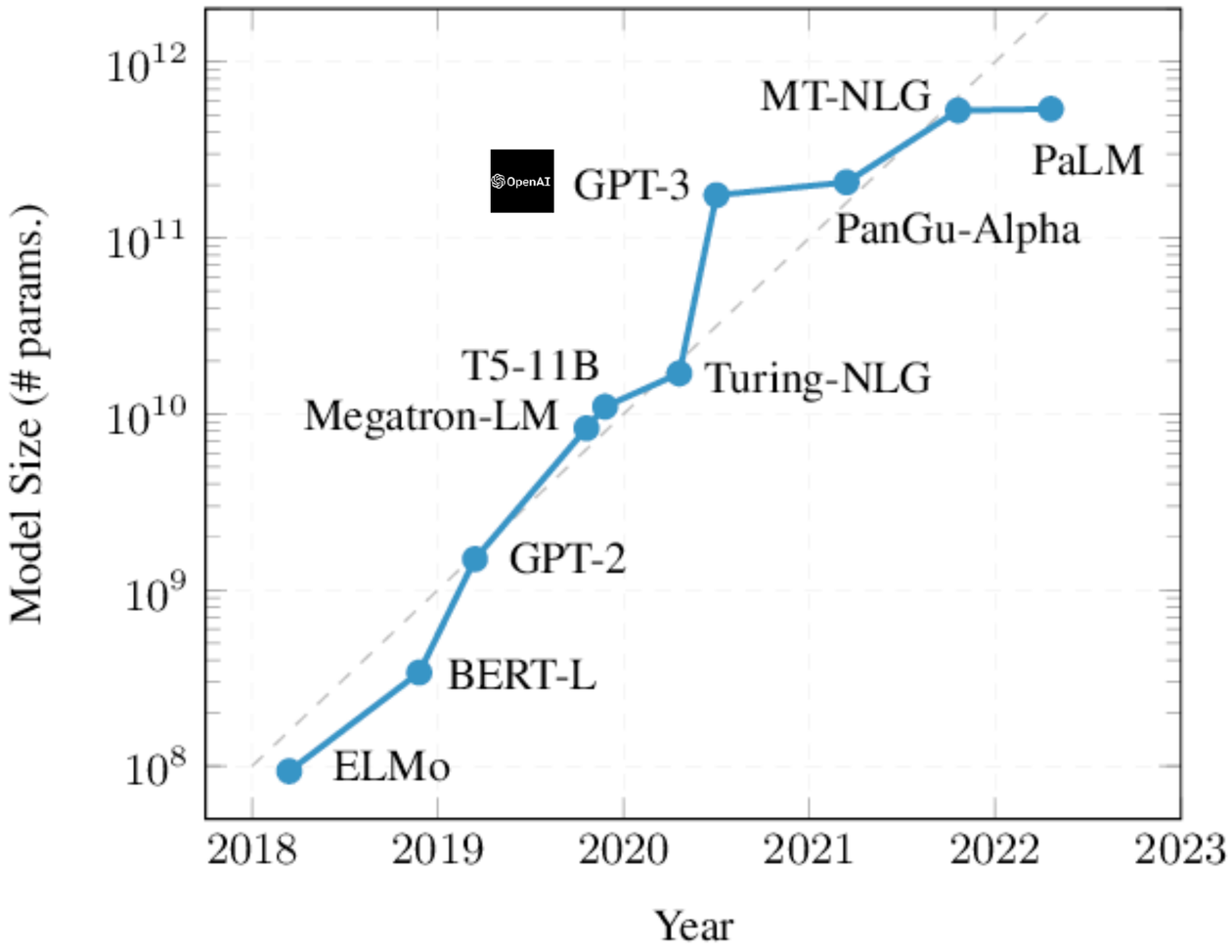
- Part 1: What is Modern NLP on and How Does it Work?
- Part 2: What do we Really Know about our Tools?
- **Part 3: What can Go Wrong?**
  - Week 8: Social biases and toxicity
  - Week 9: How and what do large LMs **know**?
  - Week 10: Other languages
  - Week 11: Language + other modalities
  - Week 12: Sustainability and efficiency
  - Week 13: Summary

In Reality

- Part 1: What is Modern NLP on and How Does it Work?
- Part 2: What do we Really Know about our Tools?
- **Part 3: What can Go Wrong?**
  - ~~Week 8: Social biases and toxicity~~
  - ~~Week 9: How and what do large LMs **know**?~~
  - Week 8: Other languages
  - Week 8b: Language + other modalities
  - Week 9: Sustainability and efficiency
  - Week 10: Summary

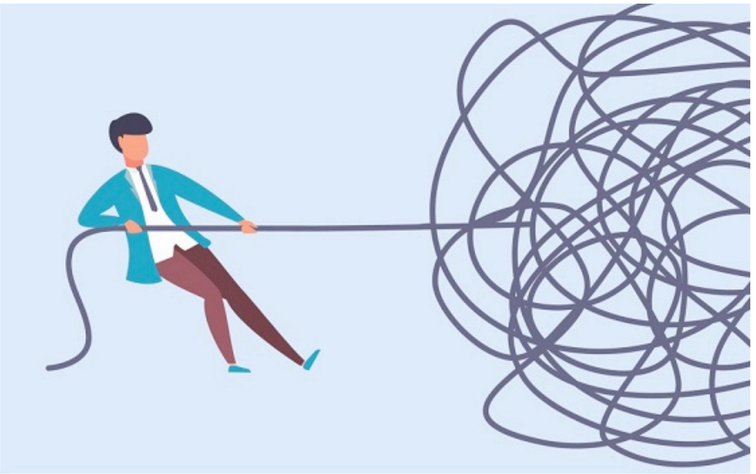
Large Models

5,000X in 4 Years



Taken from Lakim et al. (2022)

Outline



Challenges **Opportunities** **Mitigation**

Challenges with Scaling

Inclusiveness

https://syncedreview.com/2019/06/27/the-staggering-cost-of-training-sota-ai-models/

Training Costs

- BERT (Devlin et al, 2019) was trained on **16** Cloud TPUs for **4** days
- RoBERTa (Liu et al., 2019) was trained on **1024** V100 GPUs for approximately **1** day
- PaLM (Chowdhery et al., 2022) was trained on **6144** TPU v4 chips for **50** days and **3072** TPU v4 chips for **15** days

Number of Authors

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*

Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry

Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan

Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

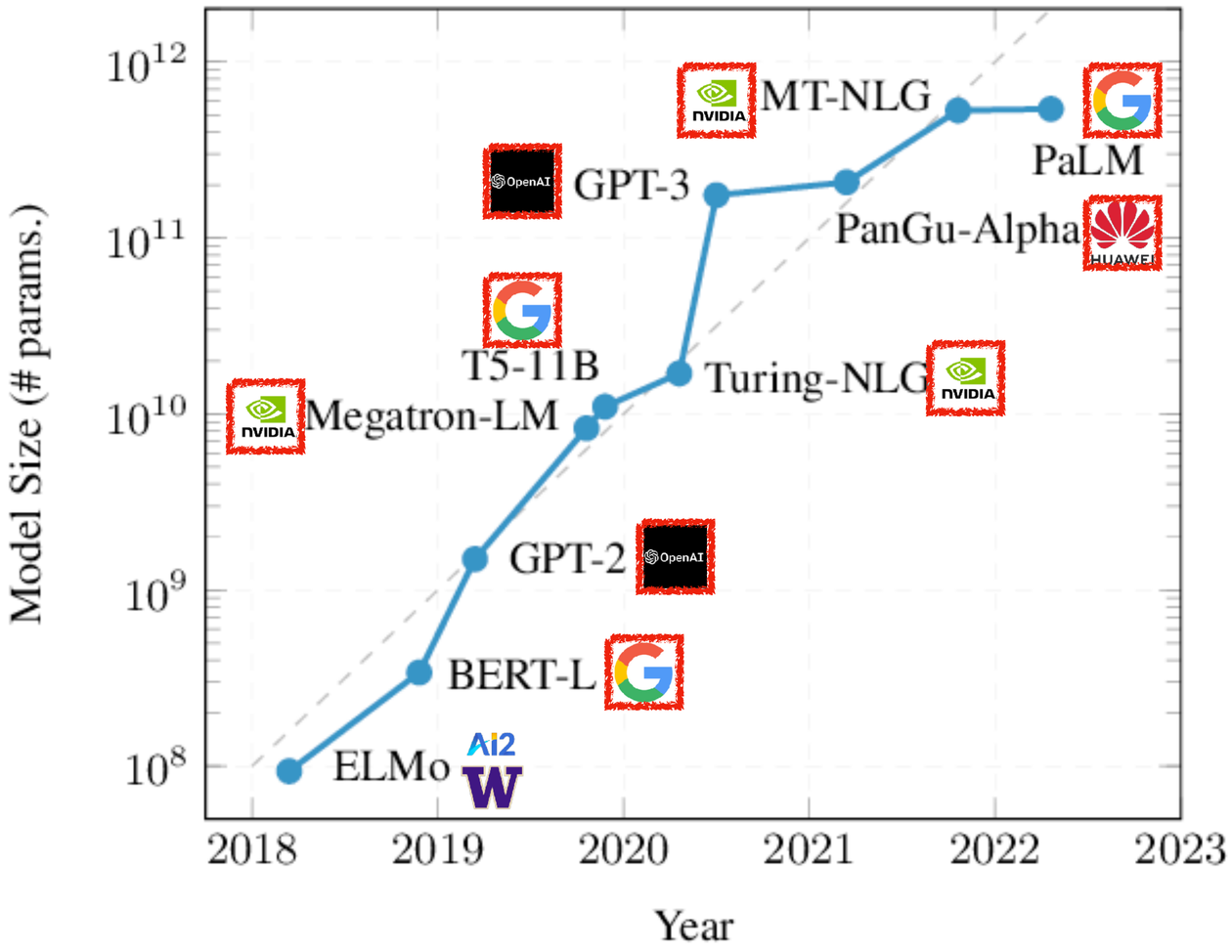
OpenAI

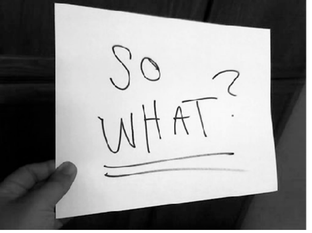
PaLM: Scaling Language Modeling with Pathways

Aakanksha Chowdhery* Sharan Narang* Jacob Devlin*
Maarten Bosma Gaurav Mishra Adam Roberts Paul Barham
Hyung Won Chung Charles Sutton Sebastian Gehrmann Parker Schuh Kensen Shi
Sasha Tsvyashchenko Joshua Maynez Abhishek Rao† Parker Barnes Yi Tay
Noam Shazeer‡ Vinodkumar Prabhakaran Emily Reif Nan Du Ben Hutchinson
Reiner Pope James Bradbury Jacob Austin Michael Isard Guy Gur-Ari
Pengcheng Yin Toju Duke Anselm Levskaya Sanjay Ghemawat Sunipa Dev
Henryk Michalewski Xavier Garcia Vedant Misra Kevin Robinson Liam Fedus
Denny Zhou Daphne Ippolito David Luan‡ Hyeontaek Lim Barret Zoph
Alexander Spiridonov Ryan Sepassi David Dohan Shivani Agrawal Mark Omernick
Andrew M. Dai Thanumalayan Sankaranarayana Pillai Marie Pellat Aitor Lewkowycz
Erica Moreira Rewon Child Oleksandr Polozov† Katherine Lee Zongwei Zhou
Xuezhi Wang Brennan Saeta Mark Diaz Orhan Firat Michele Catasta† Jason Wei
Kathy Meier-Hellstern Douglas Eck Jeff Dean Slav Petrov Noah Fiedel

Google Research

It's a Rich Man's World





Lack of Inclusiveness

- Lack of diversity
  - Of opinions
  - Of cultures
  - Of backgrounds
- Who gets to decide
  - How does state-of-the-art look like?
  - What to work on?
  - Which data to use?
- Closed models

Closed AI Models Make Bad Baselines

24 minute read

> *This post was authored by Anna Rogers, with much invaluable help and feedback from Niranjan Balasubramanian, Leon Derczynski, Jesse Dodge, Alexander Koller, Sasha Luccioni, Maarten Sap, Roy Schwartz, Noah A. Smith, Emma Strubell (listed alphabetically)*
> *Header image credit: Sasha Luccioni*

Energy and Policy Considerations for Deep Learning in NLP

ACL 2019

Emma Strubell **Ananya Ganesh** **Andrew McCallum**
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu

Consumption	**CO_2e (lbs)**
Air travel, 1 person, 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
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Is AI really creating an environmental problem?

Google’s Answer: No!

BLOG ›

Good News About the Carbon Footprint of Machine Learning Training

TUESDAY, FEBRUARY 15, 2022

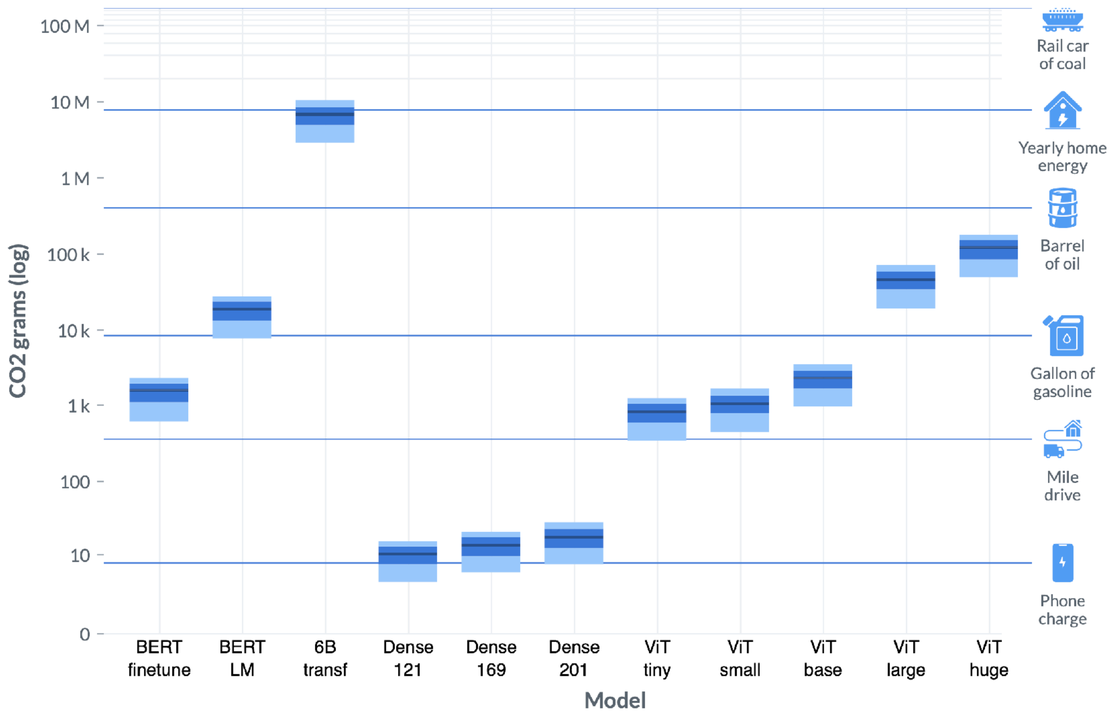
Posted by David Patterson, Distinguished Engineer, Google Research, Brain Team

*Strubell et al.’s energy estimate for NAS ended up **18.7X** too high for the average organization (…) and **88X** off in emissions for energy-efficient organizations like Google*

Our Answer: Maybe?

Measuring the Carbon Intensity of AI in Cloud Instances

JESSE DODGE, Allen Institute for AI, USA
TAYLOR PREWITT, University of Washington, USA
REMI TACHET DES COMBES, Microsoft Research Montreal, USA
ERIKA ODMARK, Microsoft, USA
ROY SCHWARTZ, Hebrew University of Jerusalem, Israel
EMMA STRUBELL, Carnegie Mellon University, USA
ALEXANDRA SASHA LUCCIONI, Hugging Face, USA
NOAH A. SMITH, Allen Institute for AI and University of Washington, USA
NICOLE DECARIO, Allen Institute for AI, USA
WILL BUCHANAN, Microsoft, USA



CO2 Relative Size Comparison

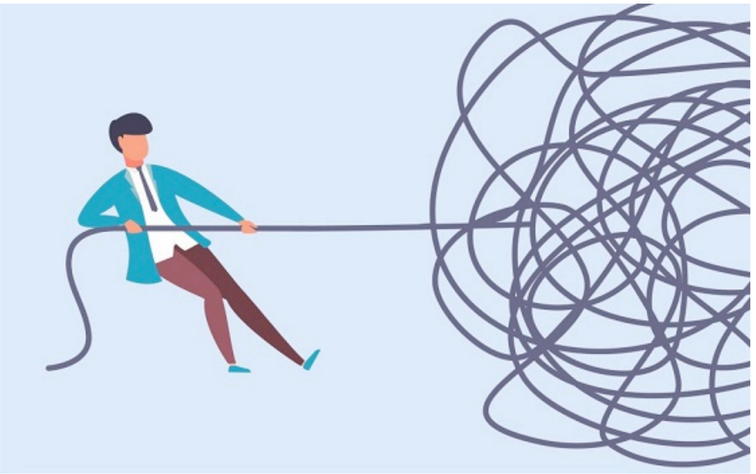
Green Software Foundation

Microsoft Azure

AI and the Environment

- Evidence around the **most expensive experiments**
  - More recent models consume 2-3 orders of magnitude more CO_2 (Luccioni et al., 2022)
  - These are typically run very few times
- What about “normal” experiments?
  - Much cheaper, but run hundreds / thousands of times a day?
- What about inference operations?
  - Very cheap (though increasingly more expensive)
  - Run billions of times a day?
  - 80-90% of AI computation is spent on inference
- Much harder to **measure**!

Outline



Challenges **Opportunities** **Mitigation**

Stop training large models?

Large Models are Important

- Push the limits of SOTA
- Released large pre-trained models save compute
- Large models are potentially faster to train
  - Li et al. (2020)
- But, large models have **concerning side affects**
  - Inclusiveness, environment
- Our goal is to **mitigate these side affects**

Mitigating the Challenges of Scaling

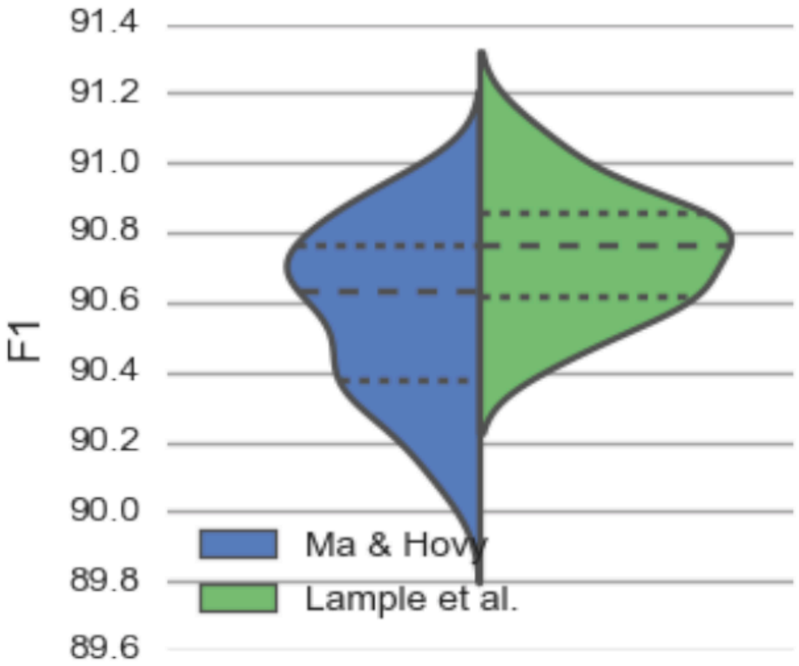
Enhanced Reporting

Efficient Methods

Is Model A > Model B?

Reimers & Gurevych (2017)

Model	F1
Model A	**91.21**
Model B	90.94



Unfair Comparison

Is **Model A > Model B**?

Better(?) Comparison

Is **Model A > Model B?** | ***Budget***

Expected Validation

Dodge, Gururangan, Card, Schwartz & Smith, 2019

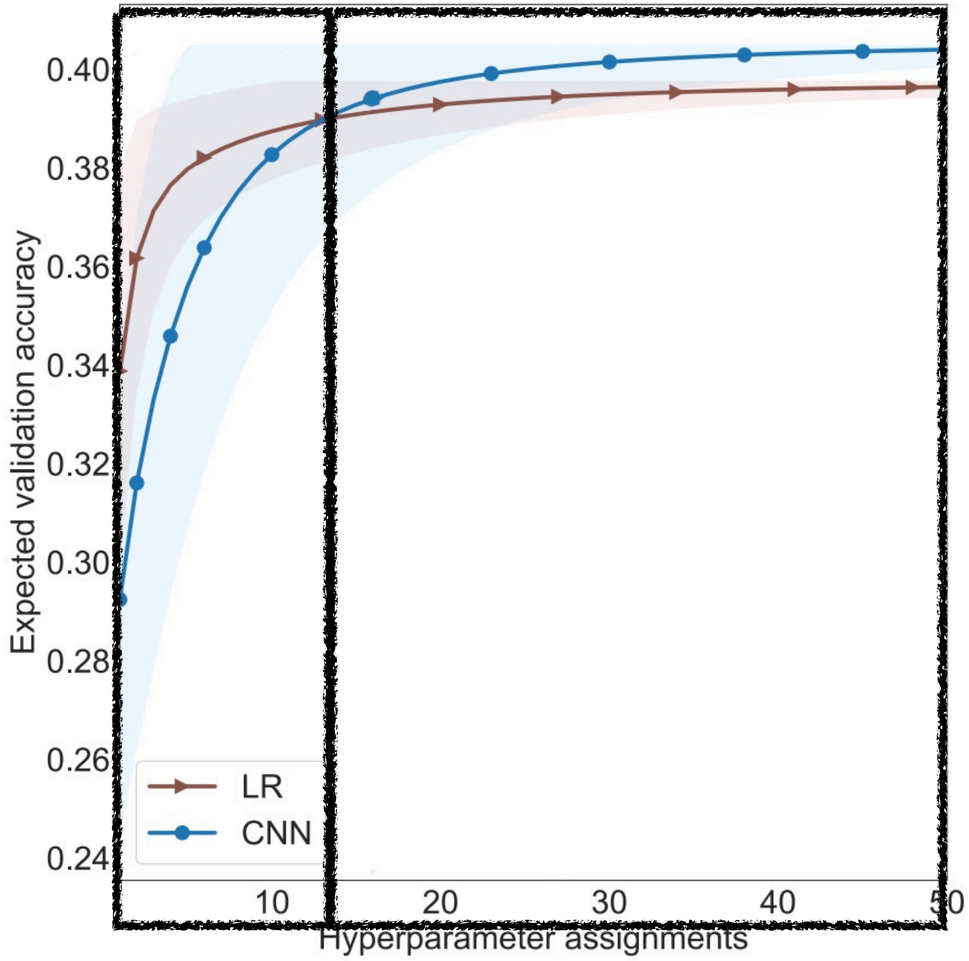
- Input: a set of experimental results $\{V_1, \ldots, V_n\}$
- Define $V_k^* = max_{i \in \{1,\ldots,k\}} V_i$
- **Expected validation performance**: $\mathbb{E}[V_k^* \mid k]$
  - k=1: $mean(\{V_1, \ldots, V_n\})$
  - k=2: $mean(\{max(V_i, V_j) \forall 1 \leq i < j \leq n\})$
  - k=n: $V_n^* = max_{i \in \{1,\ldots,n\}} V_i$

Expected Validation

Example

High budget:

CNN



Reporting

Open Questions

- How much will we gain by pouring **more compute**?
  - Kaplan et al. (2020); Clark et al. (2022); Hoffmann et al. (2022); Muennighoff et al. (2023)
- What should we report?
  - Number of experiments
  - Time
  - FLOPs
  - Energy (KW)
  - Carbon?

Mitigating the Challenges of Scaling

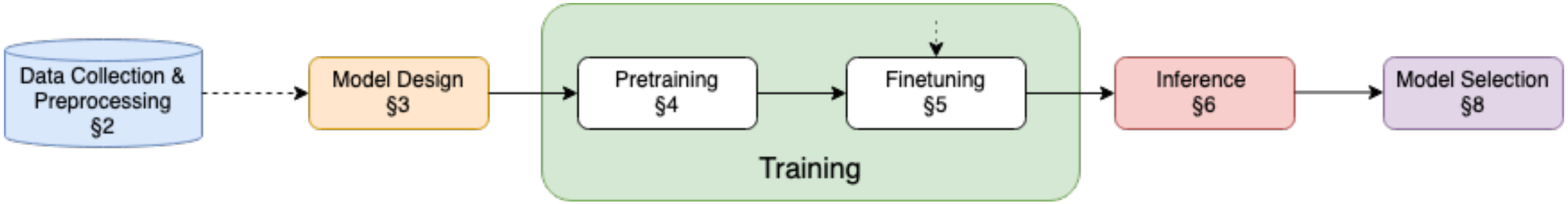
Enhanced Reporting

Efficient Methods

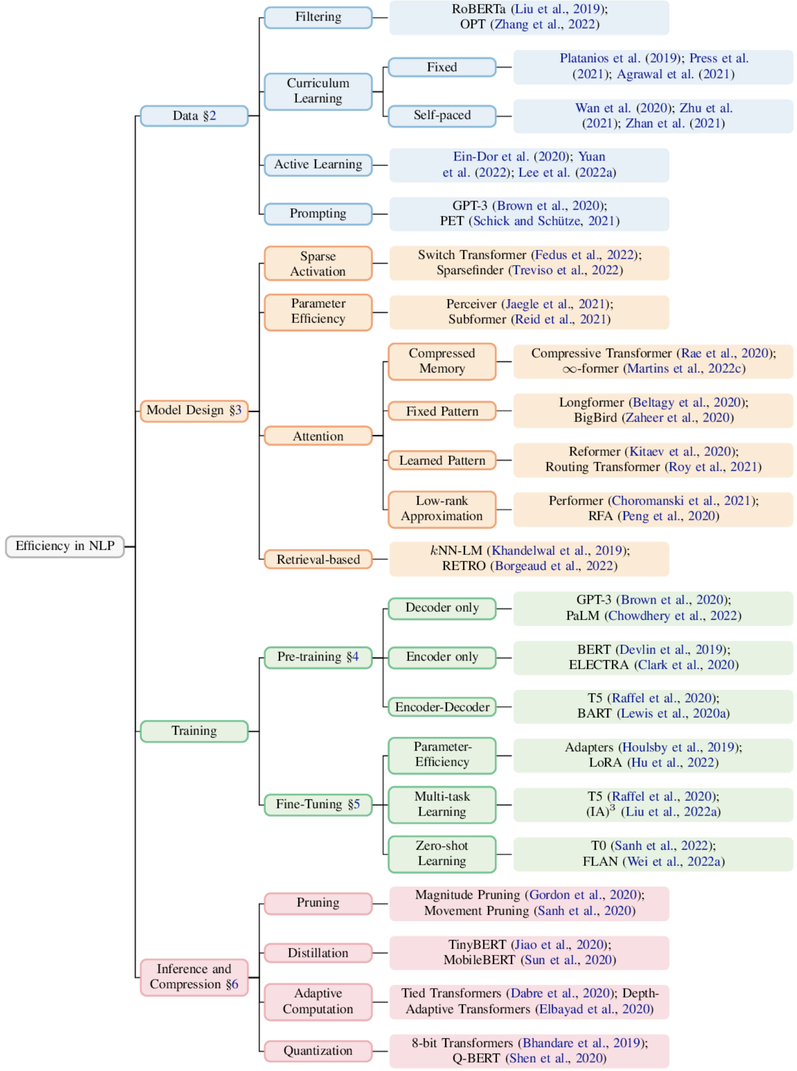
Efficient Methods for Natural Language Processing: A Survey

Marcos Treviso[10]*, Tianchu Ji[3]*, Ji-Ung Lee[7]*, Betty van Aken[8], Qingqing Cao[2], Manuel R. Ciosici[9], Michael Hassid[1], Kenneth Heafield[13], Sara Hooker[5], Pedro H. Martins[10], André F. T. Martins[10], Peter Milder[3], Colin Raffel[6], Edwin Simpson[4], Noam Slonim[12], Niranjan Balasubramanian[3], Leon Derczynski[11], Roy Schwartz[1]

[1]The Hebrew University of Jerusalem, [2]University of Washington, [3]Stony Brook University, [4]University of Bristol, [5]Cohere For AI, [6]University of North Carolina at Chapel Hill, [7]Technical University of Darmstadt, [8]Berliner Hochschule für Technik, [9]University of Southern California, [10]IST/University of Lisbon & Instituto de Telecomunicações, [11]IT University of Copenhagen, [12]IBM Research, [13]University of Edinburgh



Efficient Methods in NLP



Filtering

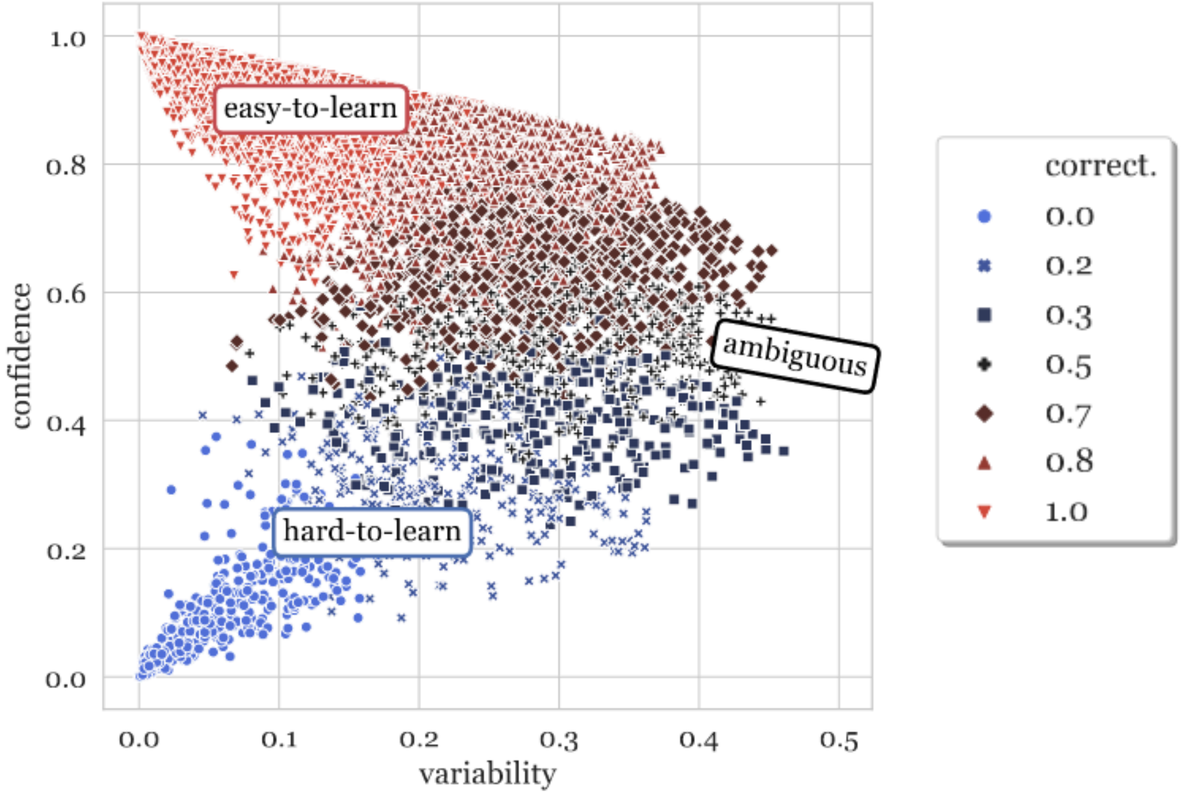
- Non-text
  - Gibrish, HTML
- Text in other languages
- Foul text
- Typically done via simple, rule-based heuristics
  - Noisy process

Data-Efficient Training

Swayamdipta, Schwartz et al. (2020)

- Not all training instances contribute the same to learning
  - Some are "*easy-to-learn*", others are more challenging

Reminder: Dataset Map



Sample-Efficient Training

WinoGrande, RoBERTa-Large

	WINOG. Val. (ID)	*WSC* (OOD)
100% train	$79.7_{0.2}$	$86.0_{0.1}$
random	$73.3_{1.3}$	$85.6_{0.4}$
ambiguous	$\mathbf{78.7}_{0.4}$	$\mathbf{87.6}_{0.6}$

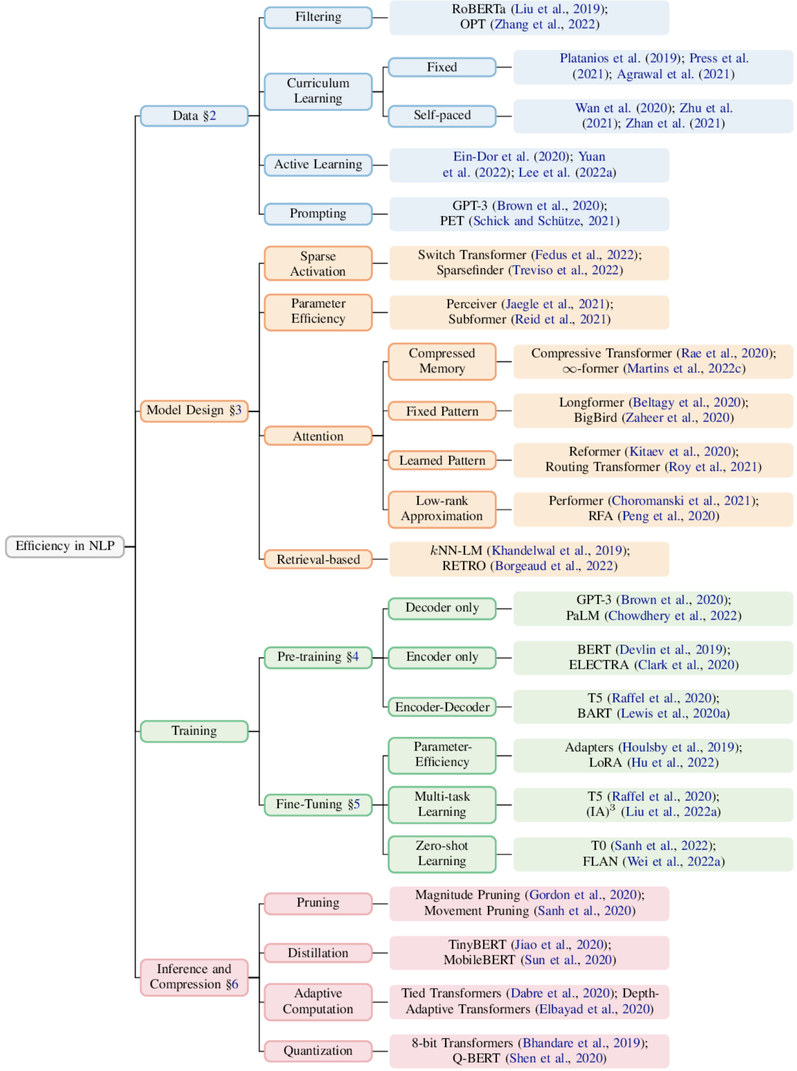
33%

Data Efficiency

Open Questions

- Do we really need massive web-scale data to train our models?
  - Can we get along with less?

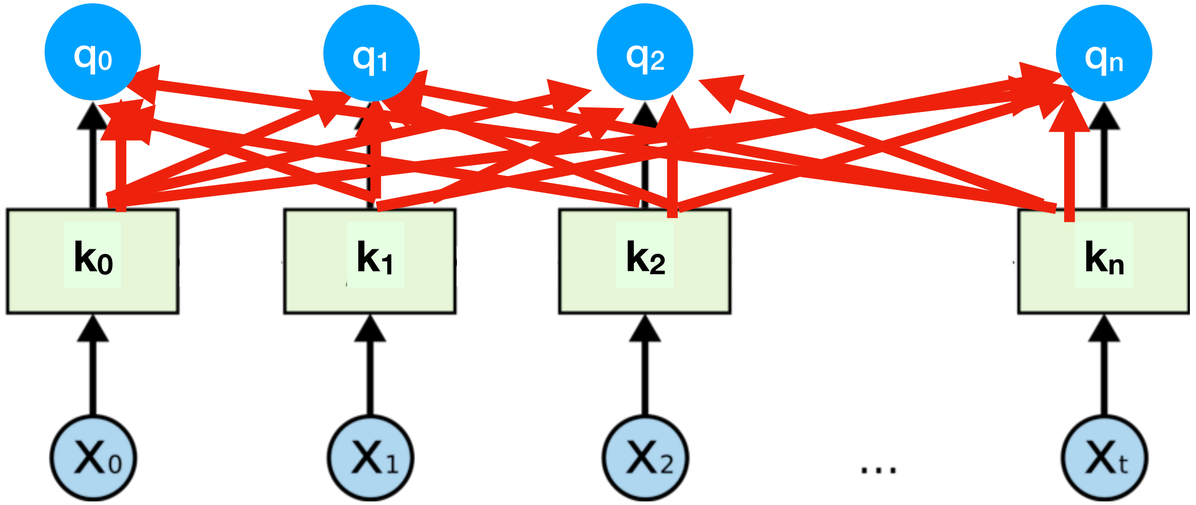
Efficient Methods in NLP



Reminder: Transformers

Vaswani et al., 2017

- Each word attends to all other words
- O(n^2) complexity in the sentence length n
- Fatal for long sequences
  - Books, articles, etc.

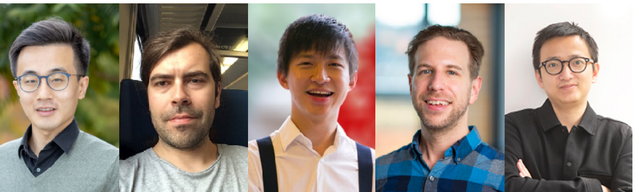


Random Feature Attention

Peng, Pappas, Yogatama, Schwartz, Smith, & Kong, 2021

- **Key idea**: approximate the attention function using random Fourier features
  - Rahimi and Recht (2007)
- Some math
- Linear runtime and memory requirements

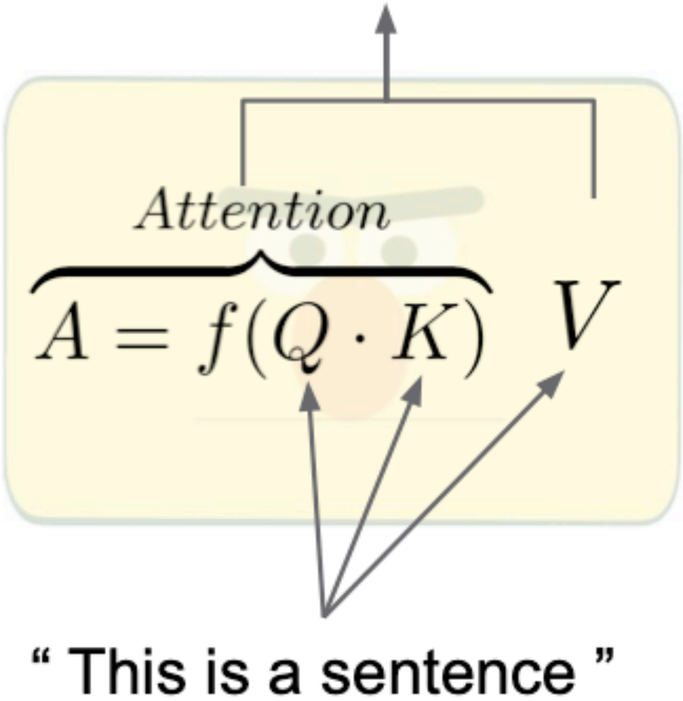
- *Define*
  - $\phi(\mathbf{x}) = \sqrt{1/D}[\sin(\mathbf{w}_1 \cdot \mathbf{x}), \ldots, \sin(\mathbf{w}_D \cdot \mathbf{x}), \cos(\mathbf{w}_1 \cdot \mathbf{x}), \ldots, \cos(\mathbf{w}_D \cdot \mathbf{x})]^\top$
  - Where $w_i \sim N(0,1)$
- Then
  - $E[\phi(\mathbf{q})^T \phi(\mathbf{k})] = \exp(\mathbf{q}^T \cdot \mathbf{k})$
- $\mathrm{attn}(\mathbf{q}_t, \{\mathbf{k}_i\}, \{\mathbf{v}_i\}) = \sum_i softmax(\mathbf{q}_t, \mathbf{k}_i)\mathbf{v}_i^\top = \sum_i \frac{\exp(\mathbf{q}_t \cdot \mathbf{k}_i)}{\sum_j \exp(\mathbf{q}_t \cdot \mathbf{k}_j)} \mathbf{v}_i^\top$
- $\approx \sum_i \frac{\phi(\mathbf{q}_t)^T \phi(\mathbf{k}_i)}{\sum_j \phi(\mathbf{q}_t)^T \phi(\mathbf{k}_j)} \mathbf{v}_i^\top$
- $= \frac{\phi(\mathbf{q}_t)^T \sum_i \phi(\mathbf{k}_i) \otimes \mathbf{v}_i^\top}{\phi(\mathbf{q}_t)^T \sum_j \phi(\mathbf{k}_j)}$

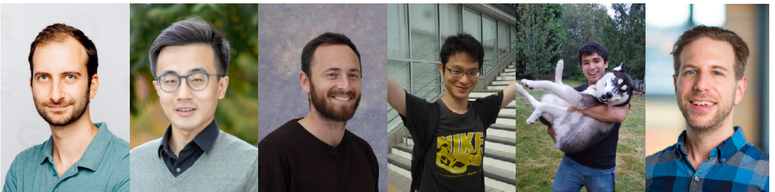


How Much Does Attention Actually Attend?

Hassid, Peng, Rotem, Kasai, Montero, Smith & Schwartz, 2022

- Model doesn't collapse
  - Average accuracy loss of 8% only
- Potential for huge savings



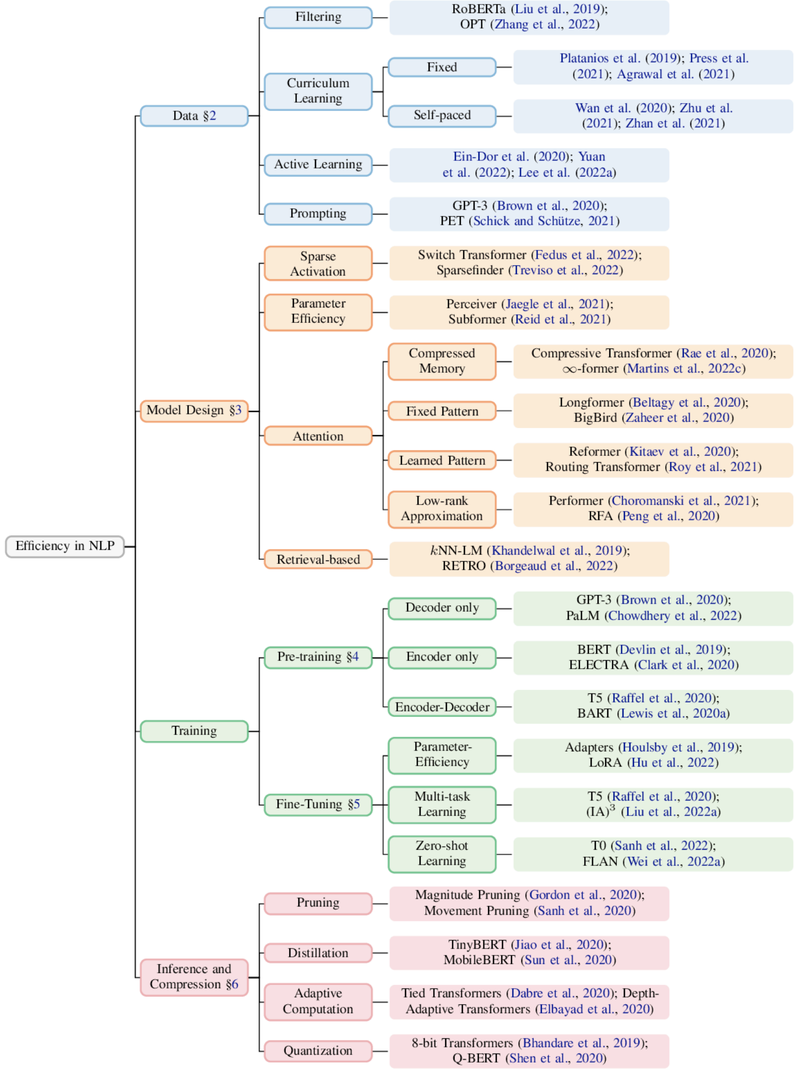


Efficient Modeling

Open Questions

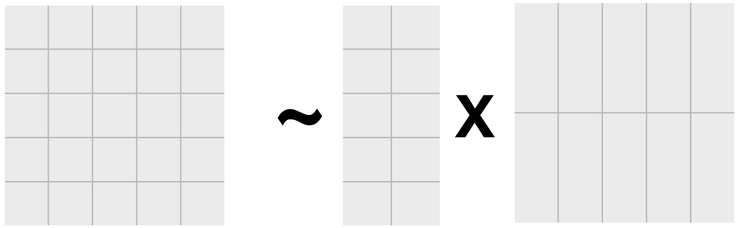
- Can we find the next generation of Transformers?
  - S4 (Gu et al., 2021), MEGA (Ma et al., 2022)
- Should we store knowledge in the model parameters?
  - Retrieval-based models
  - Gu et al (2018); Lewis et al. (2020); Borgeaud et al. (2022)

Efficient Methods in NLP

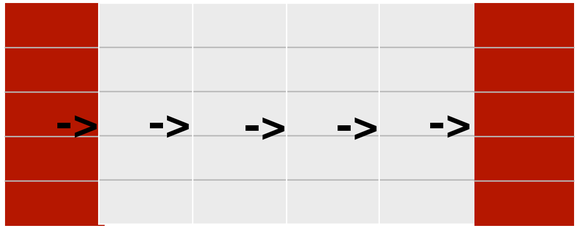


Space Efficiency

- Weight Factorization
  - Lan et al. (2019)

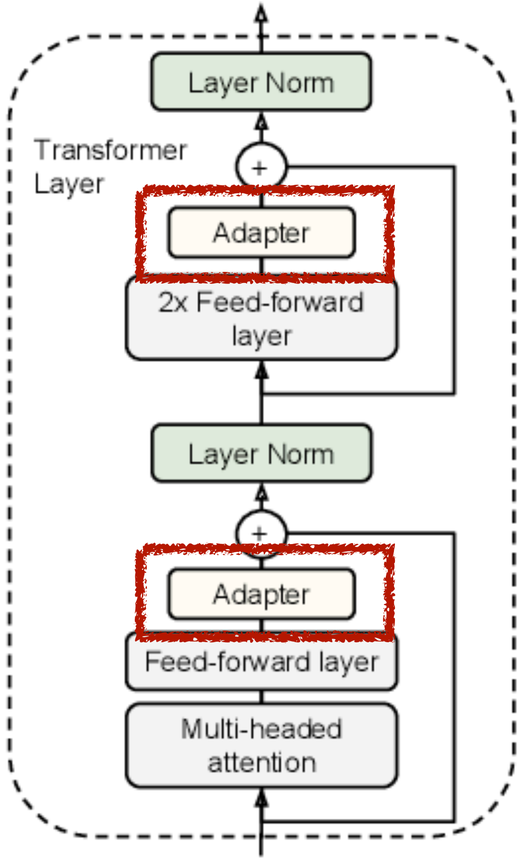


- Weight Sharing
  - Inan et al. (2016); Press & Wolf (2017)



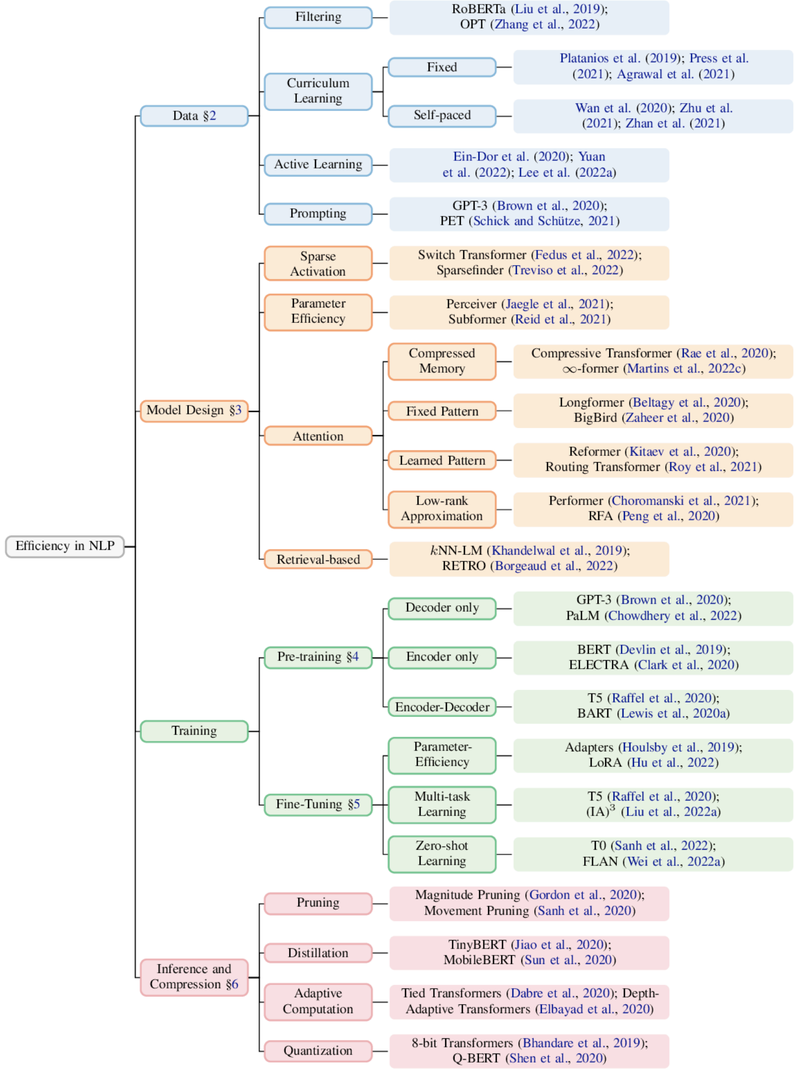
Adapters

Houlsby et al. (2019)



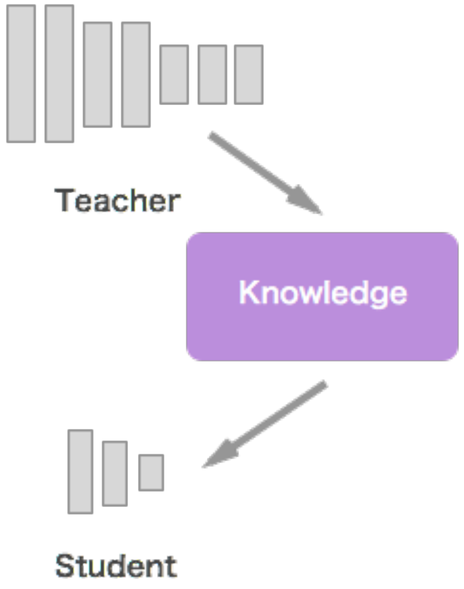
- Parameter-efficient fine-tuning
- Instead of updating all parameters during training, add <u>small</u> components (**adapters**) and only update those

Efficient Methods in NLP



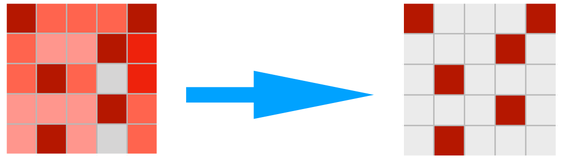
Model Distillation

- Student-teacher model
  - Hinton et al. (2015)
- Student performs on par with teacher
  - Better than a student of similar size trained from scratch



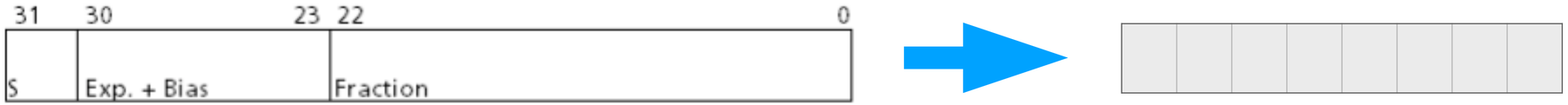
Pruning

- Learn a sparse version of a trained model
  - Lacun et al. (1989); Frankle & Corbin (2019)
- Hard to translate to speed and energy savings
  - Although see BlockSparse (Gray et al., 2017)
  - Potentially useful for fast training (Dettmers & Zettlemoyer, 2019)



Quantization

- Reduced precision
  - Zafrir et al. (2019); Shen et al. (2020); Bhandare et al. (2019)

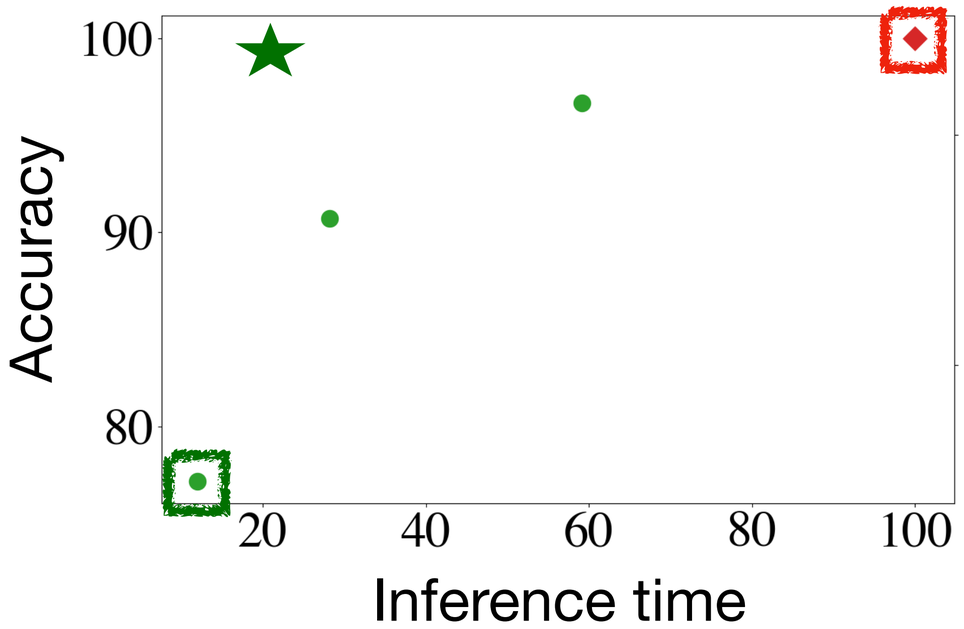


Matching Model and Instance Complexity

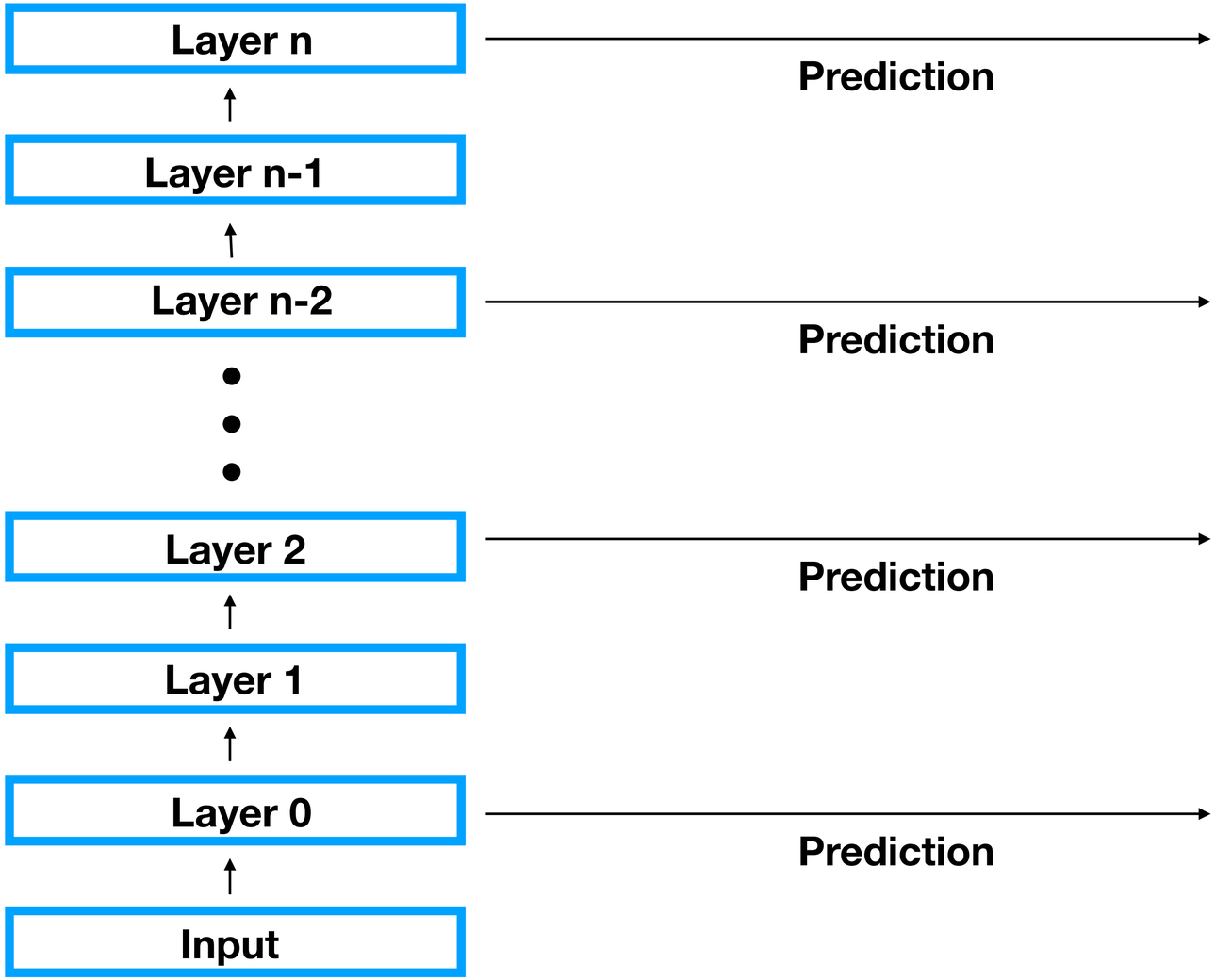
Schwartz, Stanovsky et al., 2020

Run an efficient model on "easy" instances,

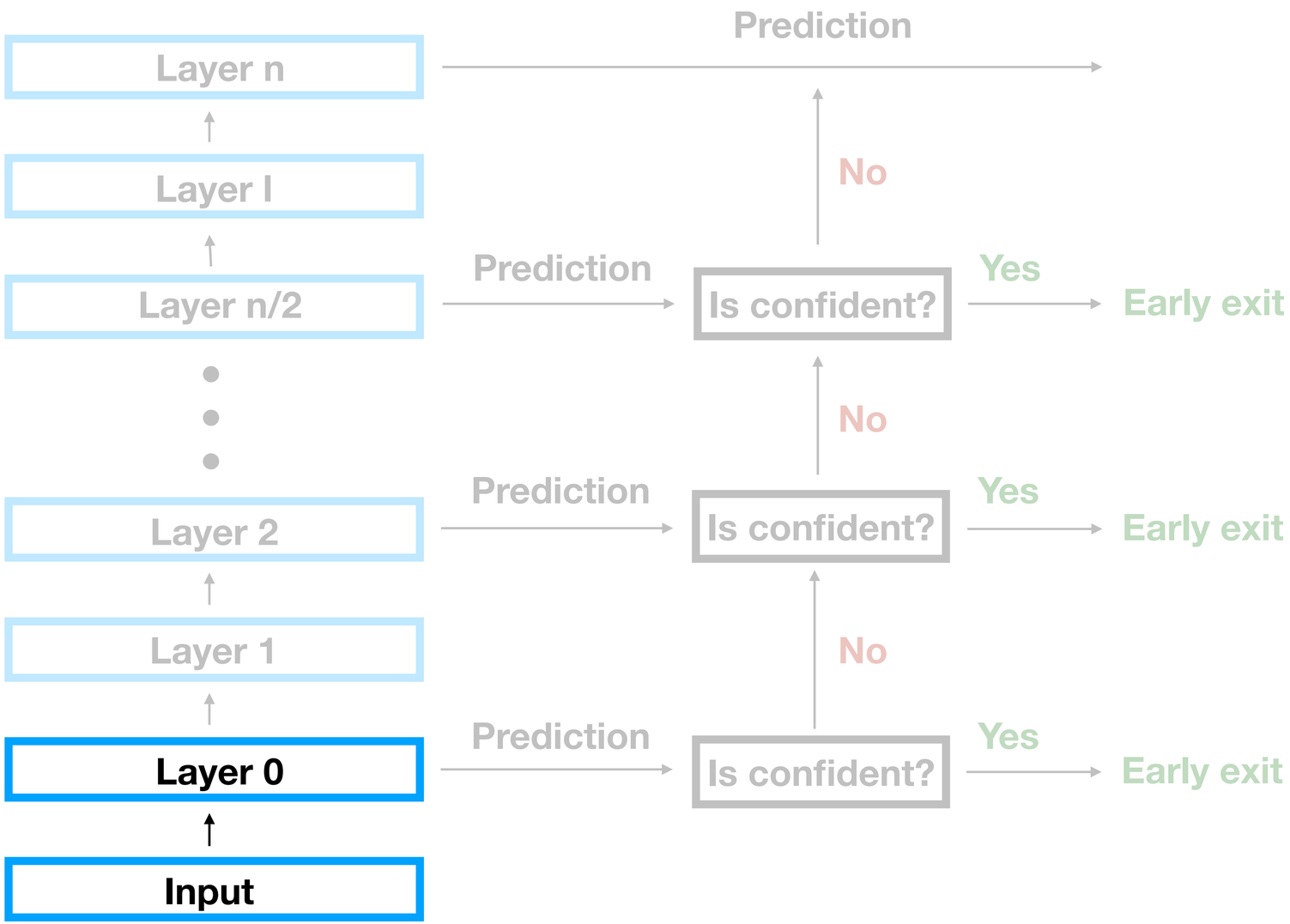
and an expensive model on "hard" instances



Our Approach: **Training** Time



Our Approach: **Test** Time

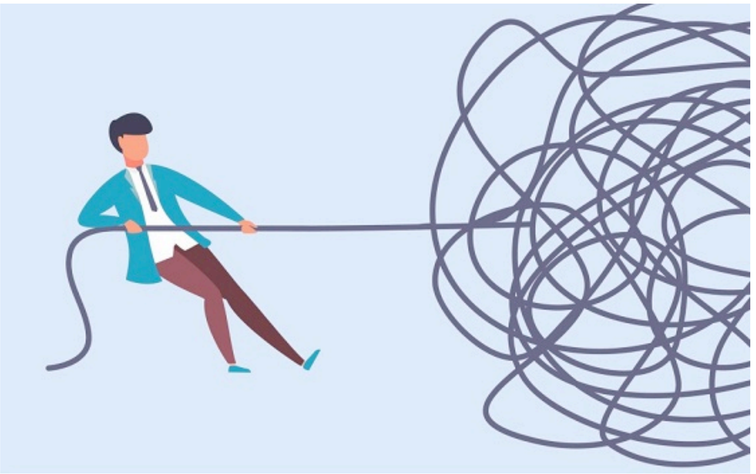


Efficiency

Open Questions

- What makes a good sparse structure?
  - Frankle and Carbin (2019)
- Combining different methods

Outline



Challenges **Opportunities** **Mitigation**

Tackling Climate Change with Machine Learning

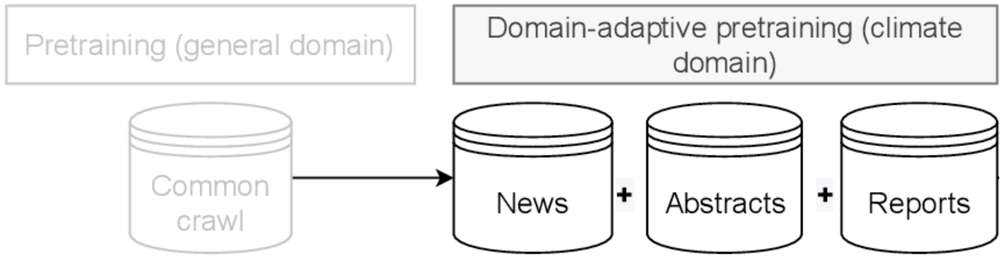
David Rolnick[1]*, Priya L. Donti[2], Lynn H. Kaack[3], Kelly Kochanski[4], Alexandre Lacoste[5], Kris Sankaran[6,7], Andrew Slavin Ross[9], Nikola Milojevic-Dupont[10,11], Natasha Jaques[12], Anna Waldman-Brown[12], Alexandra Luccioni[6,7], Tegan Maharaj[6,8], Evan D. Sherwin[2], S. Karthik Mukkavilli[6,7], Konrad P. Körding[1], Carla Gomes[13], Andrew Y. Ng[14], Demis Hassabis[15], John C. Platt[16], Felix Creutzig[10,11], Jennifer Chayes[17], Yoshua Bengio[6,7]

[1]University of Pennsylvania, [2]Carnegie Mellon University, [3]ETH Zürich, [4]University of Colorado Boulder, [5]Element AI, [6]Mila, [7]Université de Montréal, [8]École Polytechnique de Montréal, [9]Harvard University, [10]Mercator Research Institute on Global Commons and Climate Change, [11]Technische Universität Berlin, [12]Massachusetts Institute of Technology, [13]Cornell University, [14]Stanford University, [15]DeepMind, [16]Google AI, [17]Microsoft Research

	Computer vision	NLP	Time-series analysis	Unsupervised learning	RL & Control	Causal inference	Uncertainty quantification	Transfer learning	Interpretable ML	Other
Electricity Systems	1	1.1	1.1 1.2	1	1.1		1.1 1.2	1.3	1.1	1.1
Transportation	2.1 2.2 2.4		2	2.1 2.4	2	2.1 2.4	2	2.1 2.4	2	
Buildings & Cities	3.2	3.3	3	3	3.1	3.1	3.3	3		
Industry	4.1 4.3		4.3	4.3	4	4.2 4.3		4.2 4.3	4.3	
Farms & Forests	5.1 5.3 5.4				5.2			5.4		
CO_2 Removal			6.3				6.3	6.3		6.2
Climate Prediction	7.1		7				7.3		7	
Societal Impacts	8.1 8.4	8.4	8.2 8.3		8.2	8.3	8.2	8.1	8.3	
Solar Geoengineering			9.3		9.4		9.3 9.4			9.2
Tools for Individuals	10.1	10.1	10.2	10.3	10.2	10.1			10.2	10.2
Tools for Society		11.1	11.2 11.1	11.3	11.2 11.1	11.1 11.3	11.1	11	11.1	11.1 11.3
Education		12.2			12.1					
Finance		13.2	13				13.2			

General Purpose Tools

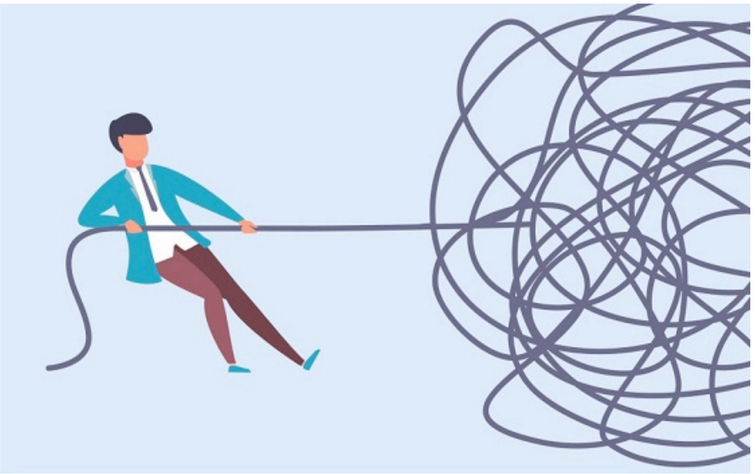
- ClimateBERT (Webersinke et al., 2022)
  - A pre-trained models trained on climate related texts



- ClimateQA (Luccioni et al., 2020)
  - A question-answering dataset focusing on climate topics

TCFD Question	Answer Passage
Does the organization describe the board's (or board committee's) oversight of climate-related risks and/or opportunities?	*The Company's Audit Committee has the delegated risk management oversight responsibility and receives updates on the risk management processes and key risk factors on a quarterly basis.*
Does the organization describe the climate-related risks or opportunities the organization has identified?	*The availability and price of these commodities are subject to factors such as changes in weather conditions, plantings, and government policies*

Summary



Challenges **Opportunities** **Mitigation**

Want to Learn More?

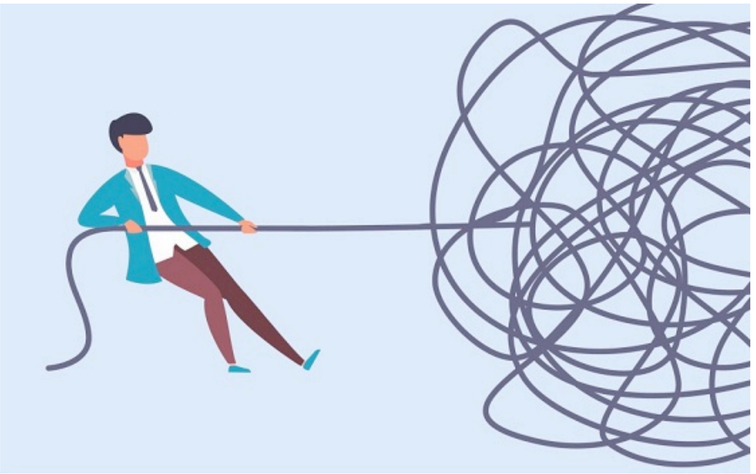
(Role Playing) Course in Efficient NLP

Fall 2023/2024, Roy Schwartz

One Last Thing….

Please fill out the teaching survey 🙏🙏🙏

Summary



Challenges **Opportunities** **Mitigation**