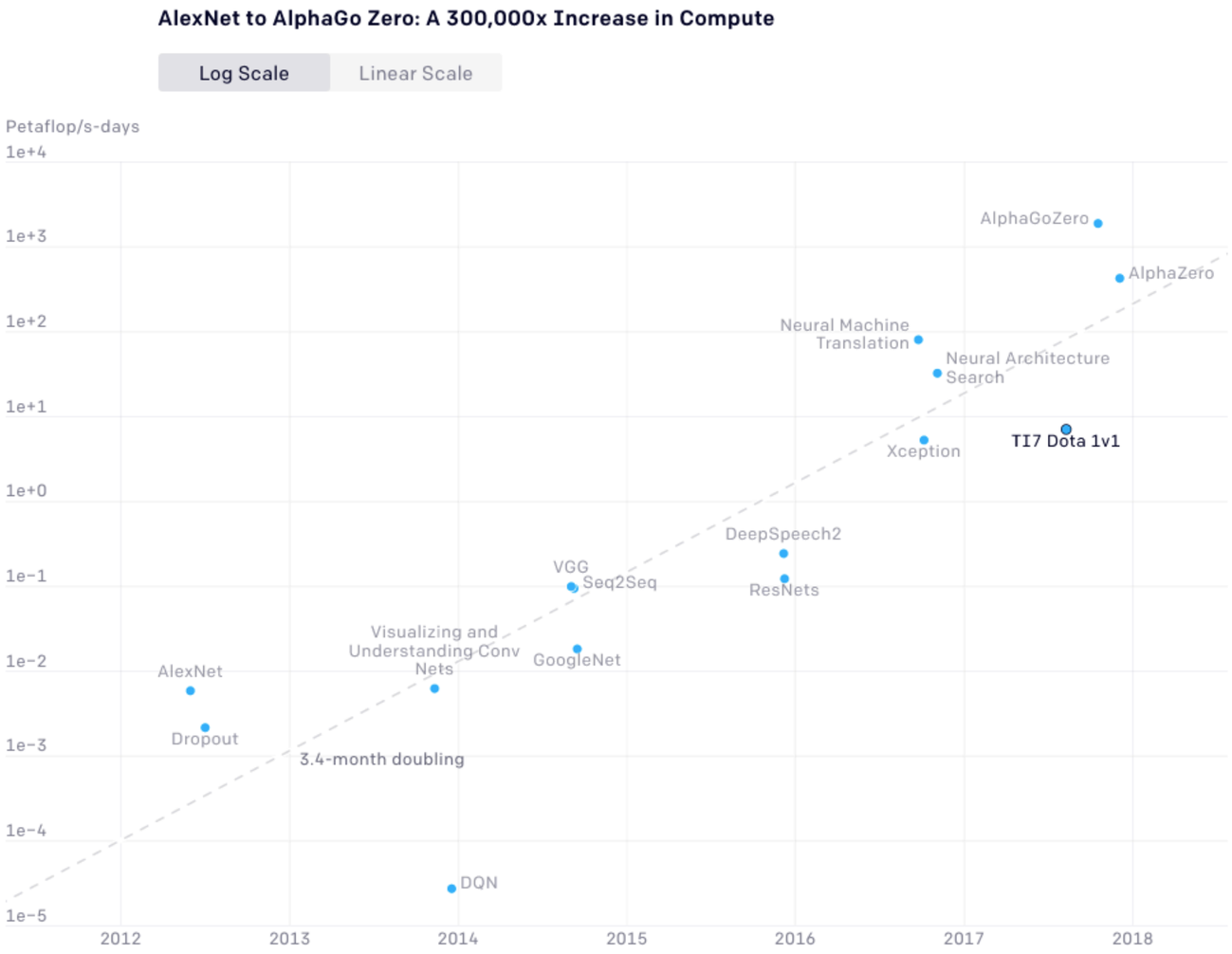
How blockers can turn into a paper: A retrospective on “Towards The Systematic Reporting of the Energy and Carbon Footprints of Machine Learning”

Peter Henderson

TOWARDS THE SYSTEMATIC REPORTING OF THE ENERGY AND CARBON FOOTPRINTS OF MACHINE LEARNING

A WORKING PAPER

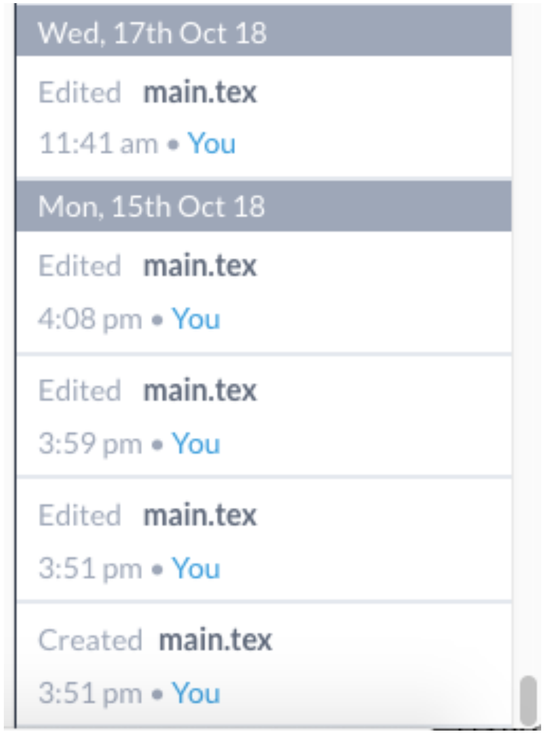
Peter Henderson[†], **Jieru Hu**[‡], **Joshua Romoff**[◇]
Emma Brunskill[†], **Dan Jurafsky**[†], **Joelle Pineau**[‡,◇]
[†]Stanford University, [‡]Facebook, [◇]Mila, McGill University



This crazy amount of compute must be having some carbon impact, how bad is it? (May 2018)

This crazy amount of compute must be having some carbon impact, how bad is it?
(May, 2018)

Start writing draft of paper
(October, 2018)



Iteration #1

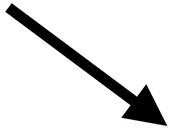
Processor	GFLOPS	W	GFLOPS/W	Release Date
NVIDIA V100 PCIe / SXM2 (NVIDIA, 2018)	14000 / 15700	250 / 300	56 / 52	June 21, 2017
NVIDIA V100 PCIe / SXM2 (Tensor) (NVIDIA, 2018)	112000 / 125000	250 / 300	448 / 417	June 21, 2017
NVIDIA GTX 1080 (NVIDIA, 2016a) / 1080Ti (NVIDIA, 2017)	8873 / 10609	180 / 250	49 / 45	May 27, 2016 / March 10, 2017
NVIDIA Titan X (NVIDIA, 2016b)	10790	250	41	August 2, 2016
NVIDIA P100 PCIe	9300	250	37.2	June 20, 2016
NVIDIA Tesla M40 (Harris, 2015)	6844	250	27.3	November 10, 2015
NVIDIA GTX 980 (NVIDIA, 2016a)	4981	165	30	September 18, 2014
NVIDIA Tesla K80 (NVIDIA, 2015)	8736	300	29.12	November 17, 2014
NVIDIA Tesla K40	4290	235	18.25	October 8, 2013
Google TPU (Tensor) (Jouppi et al., 2017a)	92000	75	1227	2015

Table 1: GFLOPS/W of some processors in terms of single precision performance unless denoted by (Tensor) in which case tensor acceleration is considered. In this case the operations are reported in terms of GOPS (Giga-Operations Per Second) since this would involve quantization for processing. We assume going forward all numbers in terms of FLOPS and that this is roughly equivalent to the amount of single-precision floating point operations needed on other hardware. Though models may change slightly for quantization changing total operations, we do not account for this. In most cases GFLOPS/W not explicitly stated, so were estimated by using GFLOPS over Thermal Design Power (TDP), denoted by W for wattage here. While the Google TPU specifications may differ by version, we were only able to reliably confirm those in Jouppi et al. (2017a).

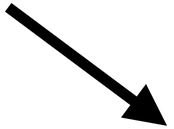
Method	PFLOPS-day	Energy (kWh)	GCPWCC	EPAWCC	ACC
[redacted]	0.20	116	88	212	132
	0.34	200	152	165	103
	0.77	499	380	911	568
	1.7	1368.6	1131	1043	706
	3.25	2860	2179	5213	3251
	4.00	5265	4012	9597	5984
	7.95	5128	3907	9346	5828
	8	6593	5024	5451	3399
	40	25806	19644	21335	13303
	85.93	55440	42245	45834	28580
	118.27	50688	38624	41905	26130
	190	122580	93406	101342	63192
Production NMT Service Rough Estimated Daily Compute	1.9×10^6	3.6×10^7	3.1×10^7	2.8×10^7	1.9×10^7
Per Capita Per Month (Puerto Rico)[2]	-	471	-	-	-
Per Minute (Entire Island of Puerto Rico)	-	35920	-	-	-

Table 2: WCC is the Worst-Case Carbon Dioxide Emissions in kg of CO2 in US Grid regions. That is the data-center regions with the worst ratio of emission-heavy to clean energy consumption. We denote GCPWCC and EPAWCC for the worst case region according to Google Cloud data and EPA data, respectively. ACC is the average case carbon emissions and is calculated according to the US average of CO2 emissions per kWh. Puerto Rico data from: https://www.eia.gov/tools/faqs/faq.php?id=97&t=3. Note, these numbers may be orders of magnitude more or less, but extrapolate based on available information using similar methods to Amodei and Hernandez (2018). In particular, the daily translation model makes many assumptions and is geared toward providing a rough estimate as described in the Appendix. We also only evaluate energy usage as calculated from the GPU, ignoring the possibly significant added energy

This crazy amount of compute must be having some carbon impact, how bad is it?
(May, 2018)



Start writing draft of paper
(October, 2018)



Estimates are inaccurate, don't want particular authors to feel attacked, decide not to submit fully written draft 2 days before the deadline.
(Nov 3, 2018)

Iteration #2

Can we estimate conference-wide distribution of compute to mask author names?

[redacted] 2017 gpu usage

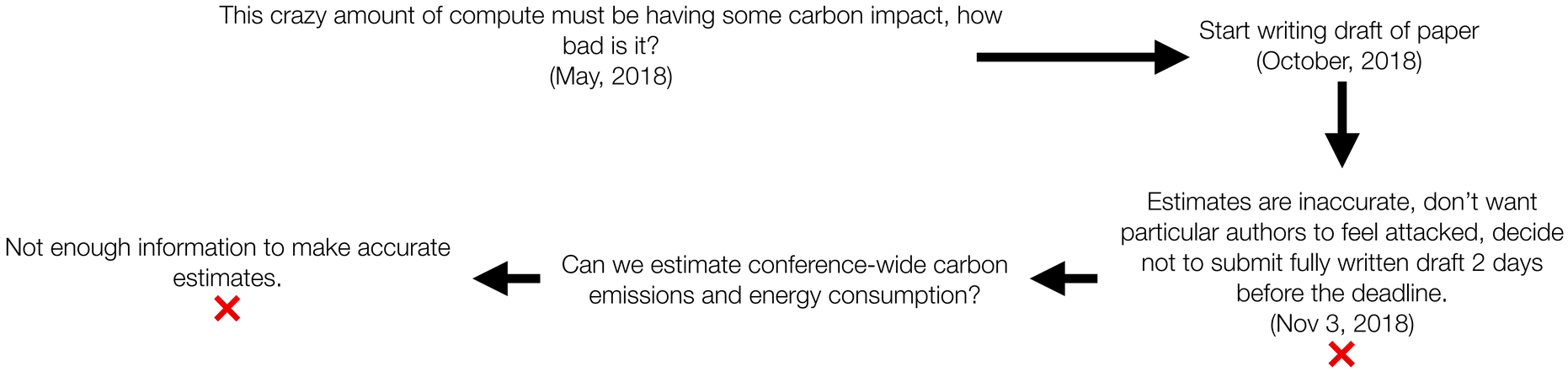
File Edit View Insert Format Data Tools Add-ons Help

Share

100% View only

Title

	A	B	C	D	E	F
1	Title	Gpu type	Pflops	gpu hours	gflop / watt	Note
2	[redacted]	Tesla K80	0.005591	150	18.64	
3		titan x				
4		p100				
5		k40m	0.004291	168	18.25	7 days total
6		no info				
7		no info				
8		1080ti				
9		titan x				
10		no info				
11		no info				
12		maybe				
13		maybe				
14		maybe				
15		not enough info				
16		not enough info (dont think they used gpus)				
17		massive experiment not easy to decipher				
18		TITAN X				
19		k80	0.005591	1224	18.64	Note: read off from figure 4 in
20		not enough info				
21		not enough info				
22		pascal titan x	0.01079	75	41	3-5 hours per method, 5 meth
23		not enough info				
24		not enough info				
25		not enough info				
26		not enough info				


This crazy amount of compute must be having some carbon impact, how bad is it?
(May, 2018)
Start writing draft of paper
(October, 2018)
Estimates are inaccurate, don't want particular authors to feel attacked, decide not to submit fully written draft 2 days before the deadline.
(Nov 3, 2018)
Can we estimate conference-wide carbon emissions and energy consumption?
Not enough information to make accurate estimates.

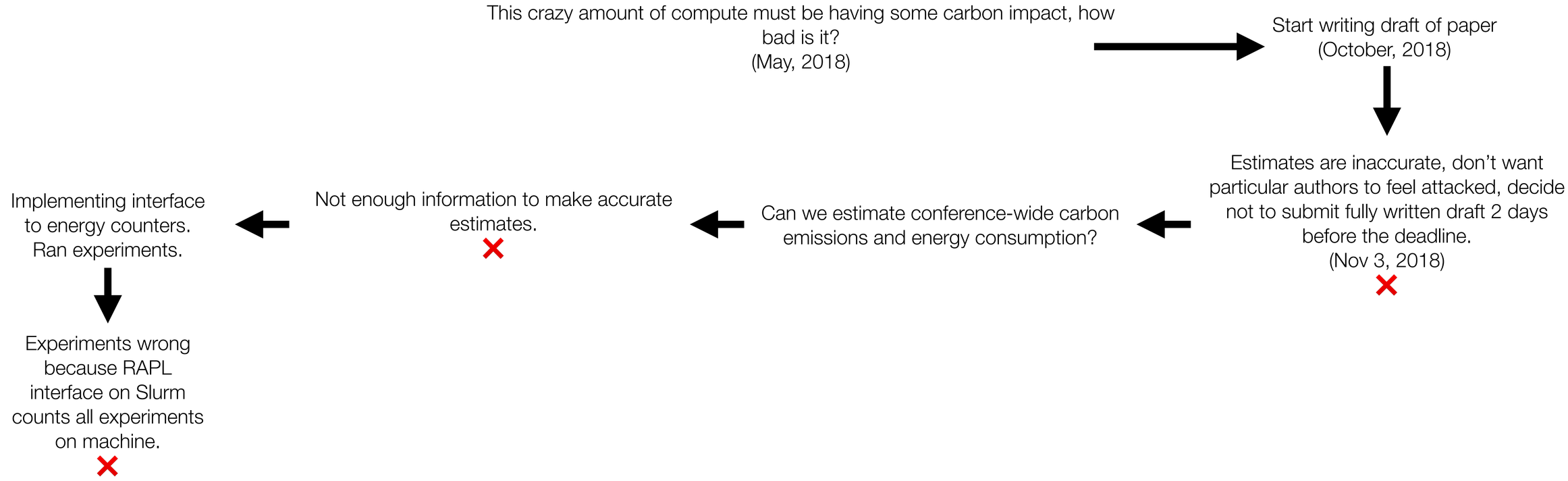
Iteration #3: How can we estimate energy more accurately?

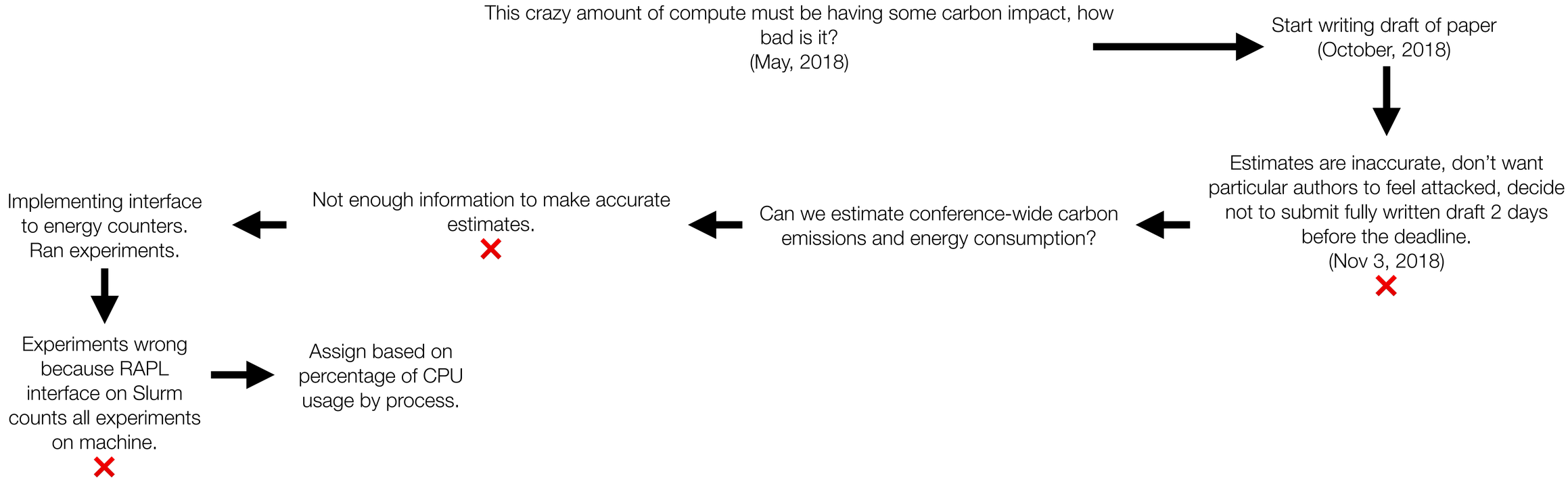
Implemented interface to different energy meters from Intel and Nvidia,
but experiments showed crazy high variance.

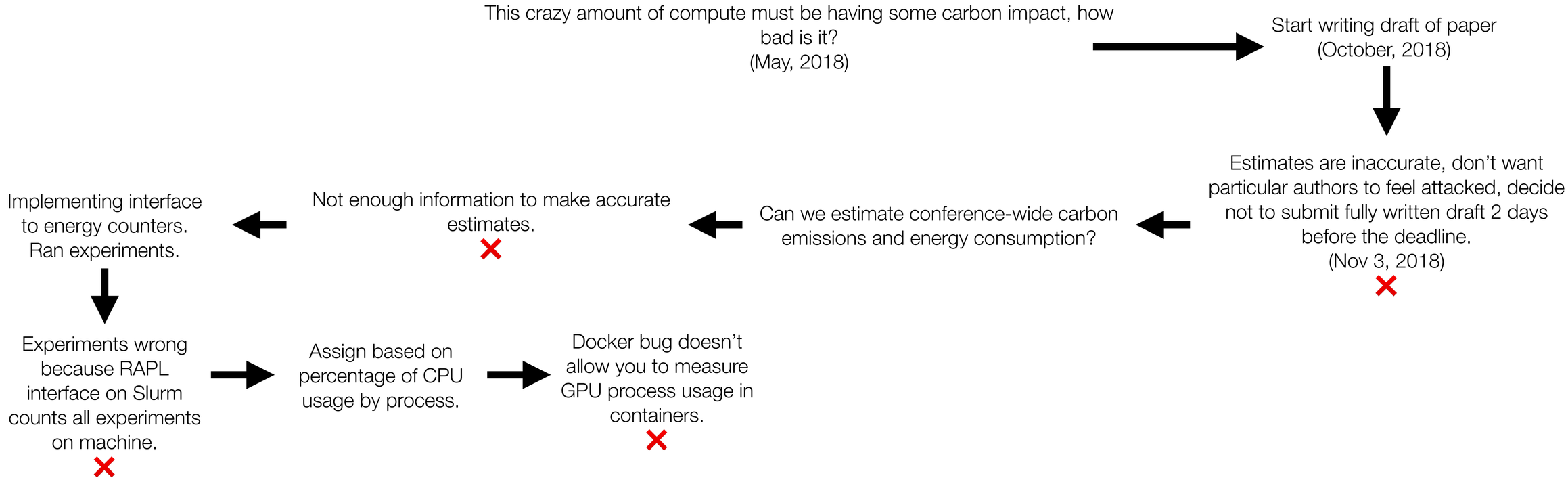
Iteration #3: How can we estimate energy more accurately?

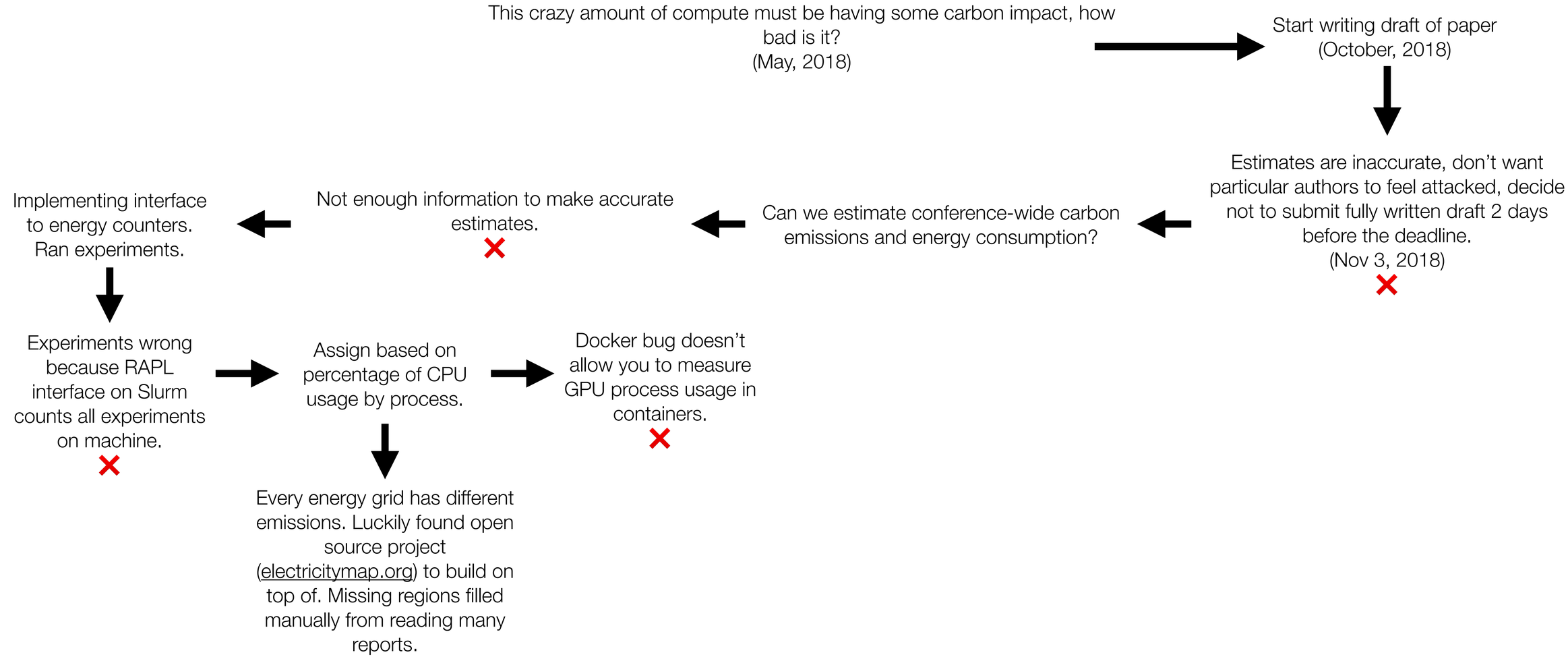
It turns out on Slurm, Intel's RAPL energy interface counts all of the energy used for every job on the worker machine!

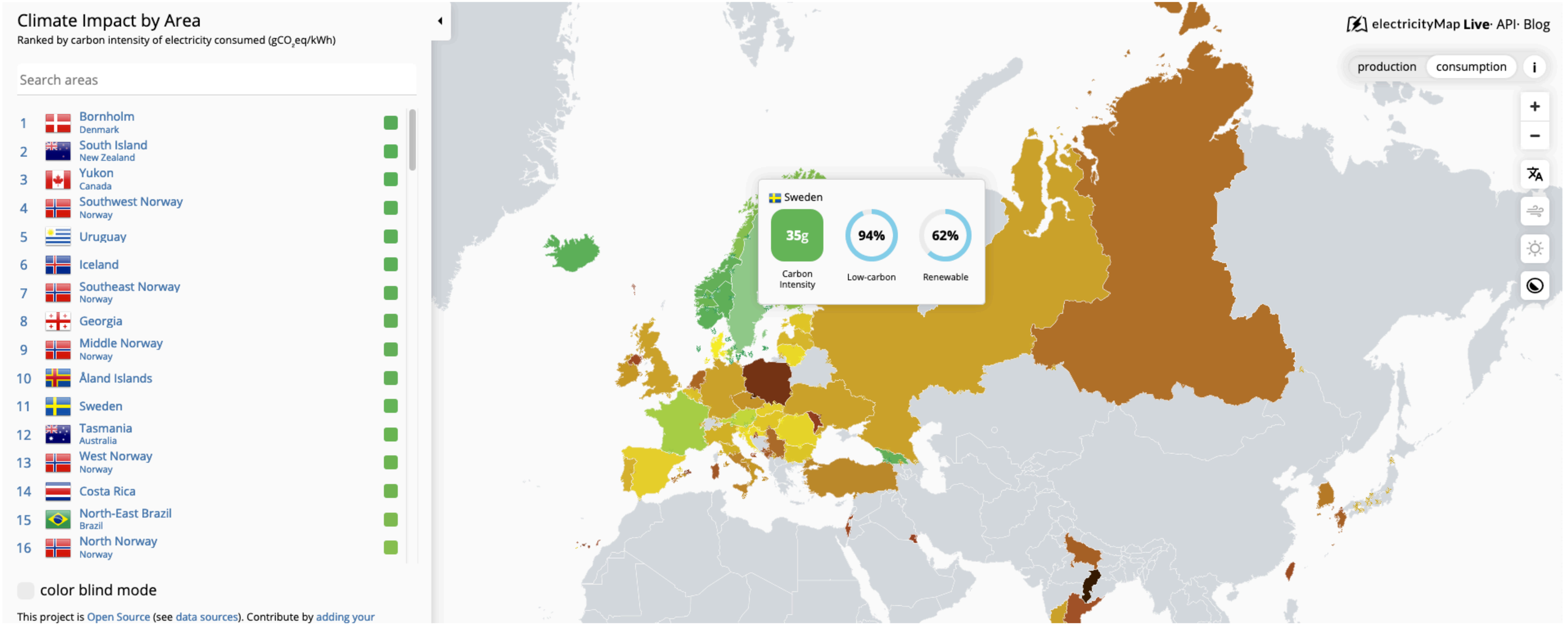
Almost moved forward with a bug.


This crazy amount of compute must be having some carbon impact, how bad is it?
(May, 2018)
Start writing draft of paper
(October, 2018)
Estimates are inaccurate, don't want particular authors to feel attacked, decide not to submit fully written draft 2 days before the deadline.
(Nov 3, 2018)
Can we estimate conference-wide carbon emissions and energy consumption?
Not enough information to make accurate estimates.
Implementing interface to energy counters. Ran experiments.
Experiments wrong because RAPL interface on Slurm counts all experiments on machine.


This crazy amount of compute must be having some carbon impact, how bad is it?
(May, 2018)
Start writing draft of paper
(October, 2018)
Estimates are inaccurate, don't want particular authors to feel attacked, decide not to submit fully written draft 2 days before the deadline.
(Nov 3, 2018)
Can we estimate conference-wide carbon emissions and energy consumption?
Not enough information to make accurate estimates.
Implementing interface to energy counters. Ran experiments.
Experiments wrong because RAPL interface on Slurm counts all experiments on machine.
Assign based on percentage of CPU usage by process.


This crazy amount of compute must be having some carbon impact, how bad is it?
(May, 2018)
Start writing draft of paper
(October, 2018)
Estimates are inaccurate, don't want particular authors to feel attacked, decide not to submit fully written draft 2 days before the deadline.
(Nov 3, 2018)
Can we estimate conference-wide carbon emissions and energy consumption?
Not enough information to make accurate estimates.
Implementing interface to energy counters. Ran experiments.
Experiments wrong because RAPL interface on Slurm counts all experiments on machine.
Assign based on percentage of CPU usage by process.
Docker bug doesn't allow you to measure GPU process usage in containers.


This crazy amount of compute must be having some carbon impact, how bad is it?
(May, 2018)
Start writing draft of paper
(October, 2018)
Estimates are inaccurate, don't want particular authors to feel attacked, decide not to submit fully written draft 2 days before the deadline.
(Nov 3, 2018)
Can we estimate conference-wide carbon emissions and energy consumption?
Not enough information to make accurate estimates.
Implementing interface to energy counters. Ran experiments.
Experiments wrong because RAPL interface on Slurm counts all experiments on machine.
Assign based on percentage of CPU usage by process.
Every energy grid has different emissions. Luckily found open source project (electricitymap.org) to build on top of. Missing regions filled manually from reading many reports.
Docker bug doesn't allow you to measure GPU process usage in containers.


Climate Impact by Area
Ranked by carbon intensity of electricity consumed (gCO2eq/kWh)
Search areas
1 Bornholm Denmark
2 South Island New Zealand
3 Yukon Canada
4 Southwest Norway Norway
5 Uruguay
6 Iceland
7 Southeast Norway Norway
8 Georgia
9 Middle Norway Norway
10 Åland Islands
11 Sweden
12 Tasmania Australia
13 West Norway Norway
14 Costa Rica
15 North-East Brazil Brazil
16 North Norway Norway
color blind mode
This project is Open Source (see data sources). Contribute by adding your
electricityMap Live API Blog
production
consumption
Sweden
35g
Carbon Intensity
94%
Low-carbon
62%
Renewable

Energy and Policy Considerations for Deep Learning in NLP

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

Awesome paper!

Awesome paper!

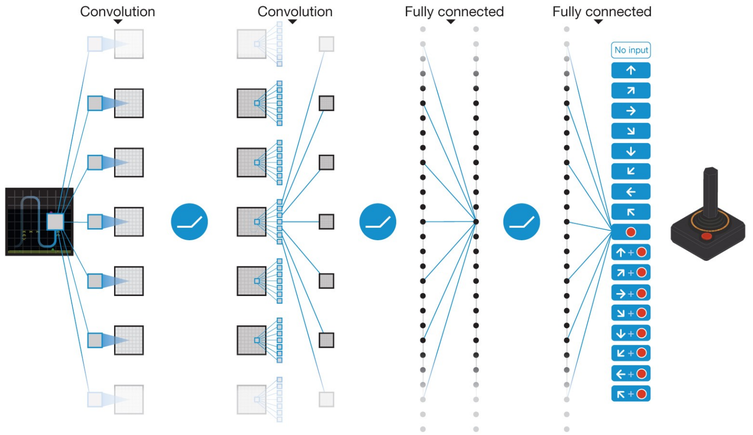
But beat us to the punch… should we even bother continuing?

Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
$\text{Transformer}_{base}$	P100x8	1415.78	12	27	26	$41–$140
Transformer_{big}	P100x8	1515.43	84	201	192	$289–$981
ELMo	P100x3	517.66	336	275	262	$433–$1472
BERT_{base}	V100x64	12,041.51	79	1507	1438	$3751–$12,571
BERT_{base}	TPUv2x16	—	96	—	—	$2074–$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	$942,973–$3,201,722
NAS	TPUv2x1	—	32,623	—	—	$44,055–$146,848
GPT-2	TPUv3x32	—	168	—	—	$12,902–$43,008

Table 3: Estimated cost of training a model in terms of CO_2 emissions (lbs) and cloud compute cost (USD).[7] Power and carbon footprint are omitted for TPUs due to lack of public information on power draw for this hardware.

Iteration #4: Publish lessons learned with new efficient RL models

RL models use ConvNets for Atari games, what if we use a mobile-optimized architecture to make them more efficient?



From: Human-level control through deep reinforcement learning

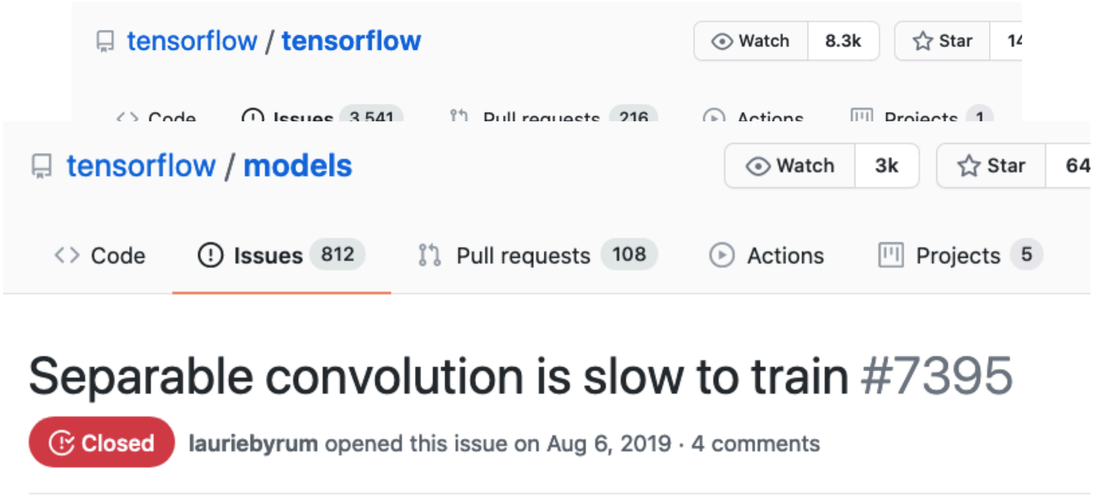
But wait… our experiments showed that using mobile architectures with less Floating Point Operations used the same or MORE energy?!

Iteration #4: Publish lessons learned with new efficient RL models

Result: mobile architectures slower and more energy hungry??

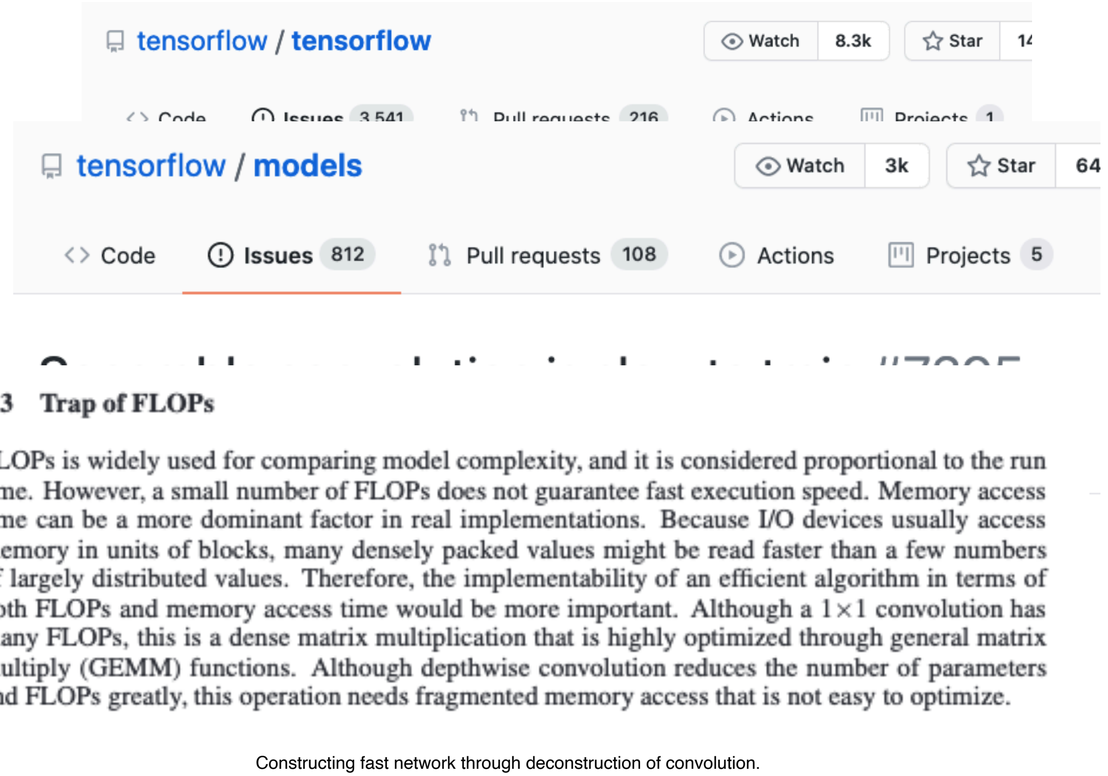
Iteration #4: Publish lessons learned with new efficient RL models

Result: mobile architectures slower and more energy hungry??



Iteration #4: Publish lessons learned with new efficient RL models

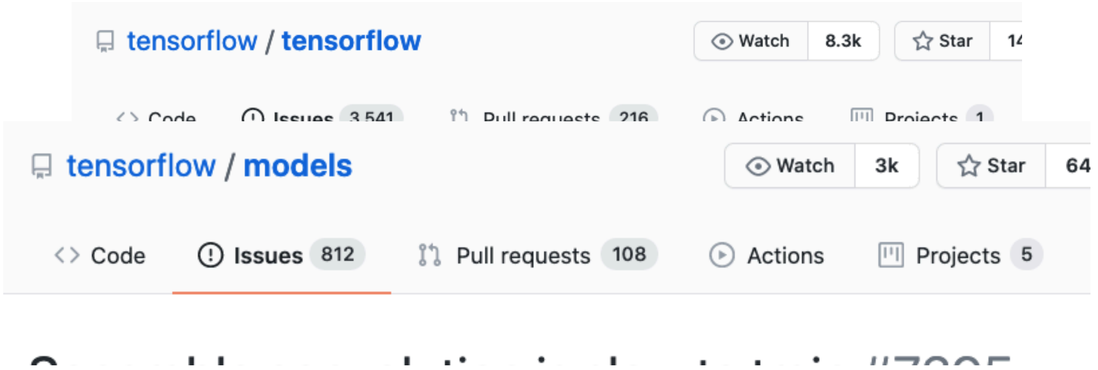
Result: mobile architectures slower and more energy hungry??



Constructing fast network through deconstruction of convolution.

Iteration #4: Publish lessons learned with new efficient RL models

Result: mobile architectures slower and more energy hungry??



3.3 Trap of FLOPs

FLOPs is widely used for comparing model complexity, and it is considered proportional to the run time. However, a small number of FLOPs does not guarantee fast execution speed. Memory access

TABLE I

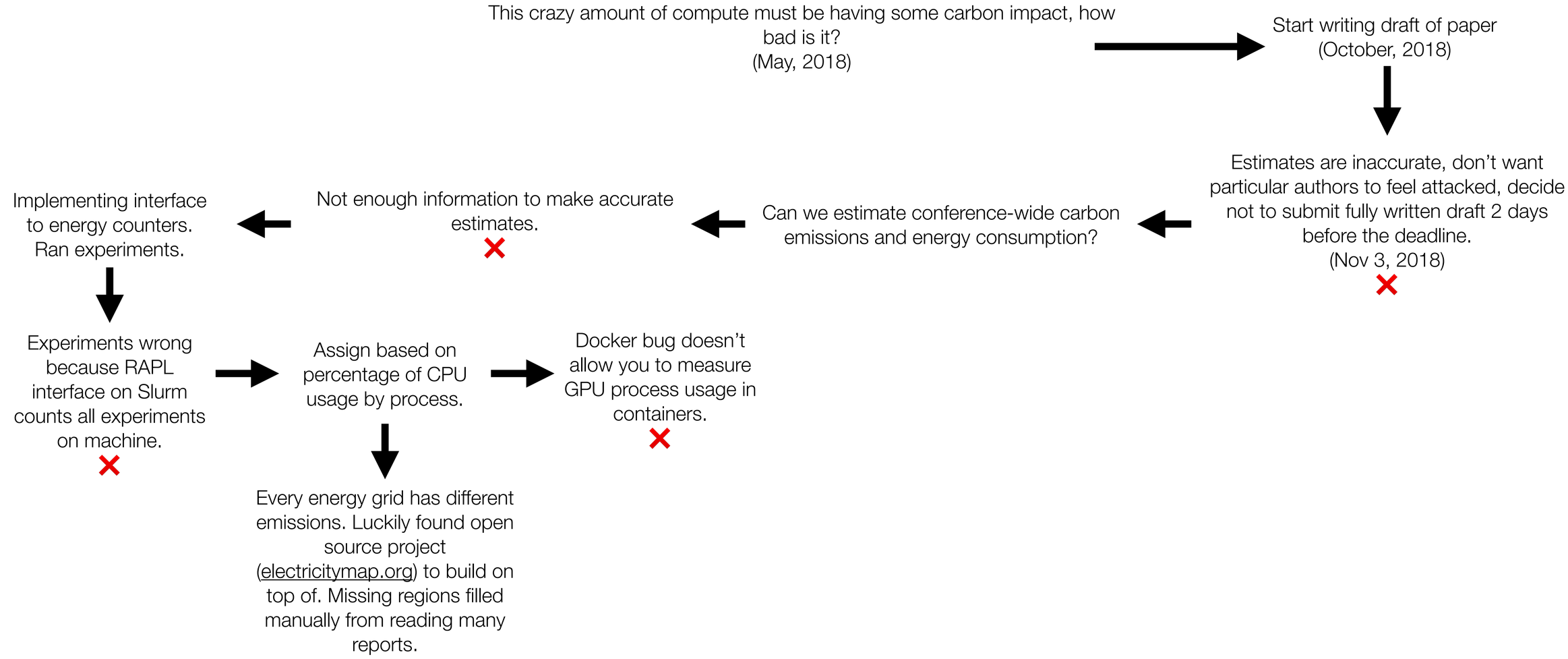
RATIOS OF MULT-ADDS, PARAMETERS, AND TRAINING TIME OF DIFFERENT LAYER TYPES FOR MOBILENETS ON CAFFE.

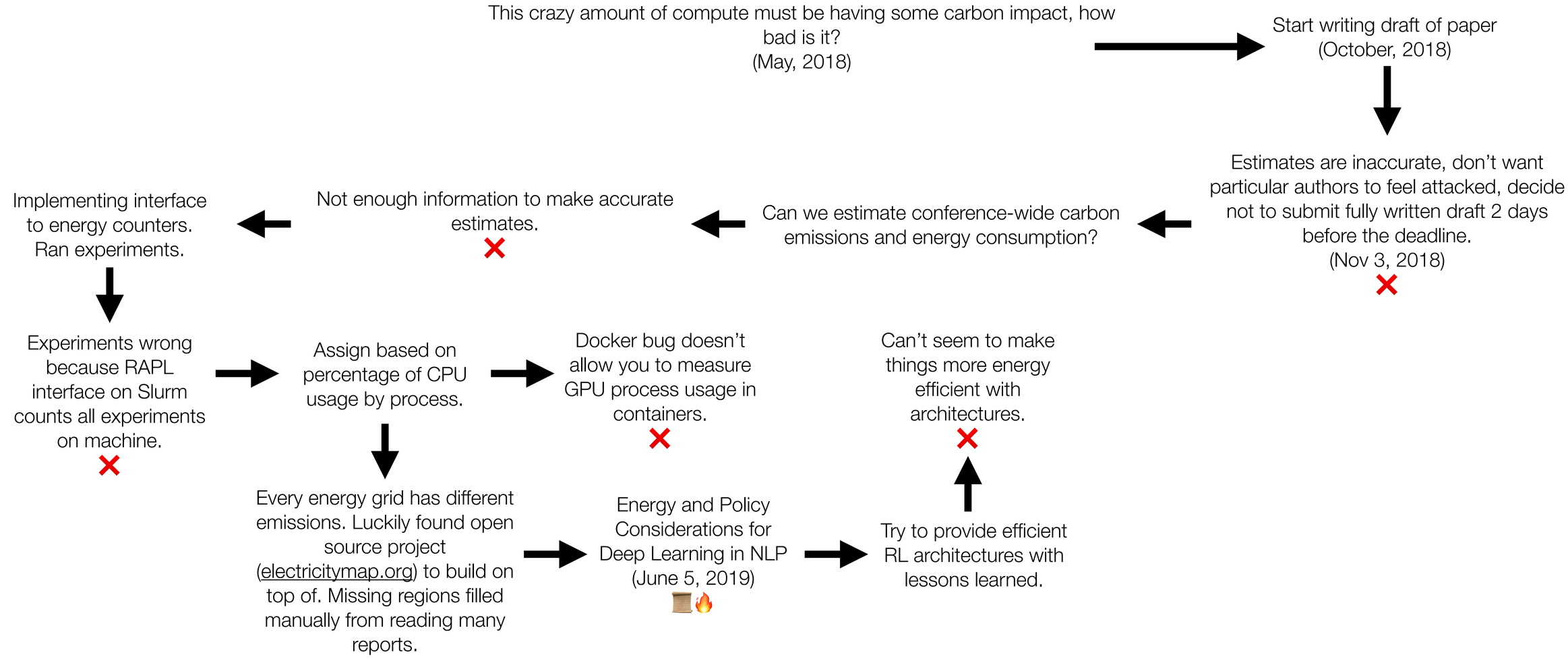
Type	Mult-Adds	Parameters	Training Time
Conv 1×1	94.86%	74.59%	16.39%
Conv DW 3×3	3.06%	1.06%	82.86%
Conv 3×3	1.19%	0.02%	0.72%
Fully Connected	0.18%	24.33%	0.03%

Conv DW: depthwise convolution layer.

se I/O devices usually access
ad faster than a few numbers
fficient algorithm in terms of
hough a 1×1 convolution has
mized through general matrix
es the number of parameters
at is not easy to optimize.

Diagonalwise Refactorization: An EfficientTraining Method for Depthwise Convolutions


This crazy amount of compute must be having some carbon impact, how bad is it?
(May, 2018)
Start writing draft of paper
(October, 2018)
Estimates are inaccurate, don't want particular authors to feel attacked, decide not to submit fully written draft 2 days before the deadline.
(Nov 3, 2018)
Can we estimate conference-wide carbon emissions and energy consumption?
Not enough information to make accurate estimates.
Implementing interface to energy counters. Ran experiments.
Experiments wrong because RAPL interface on Slurm counts all experiments on machine.
Assign based on percentage of CPU usage by process.
Every energy grid has different emissions. Luckily found open source project (electricitymap.org) to build on top of. Missing regions filled manually from reading many reports.
Docker bug doesn't allow you to measure GPU process usage in containers.


This crazy amount of compute must be having some carbon impact, how bad is it?
(May, 2018)
Start writing draft of paper
(October, 2018)
Estimates are inaccurate, don't want particular authors to feel attacked, decide not to submit fully written draft 2 days before the deadline.
(Nov 3, 2018)
Can we estimate conference-wide carbon emissions and energy consumption?
Not enough information to make accurate estimates.
Implementing interface to energy counters. Ran experiments.
Experiments wrong because RAPL interface on Slurm counts all experiments on machine.
Assign based on percentage of CPU usage by process.
Docker bug doesn't allow you to measure GPU process usage in containers.
Every energy grid has different emissions. Luckily found open source project (electricitymap.org) to build on top of. Missing regions filled manually from reading many reports.
Energy and Policy Considerations for Deep Learning in NLP
(June 5, 2019)
Try to provide efficient RL architectures with lessons learned.
Can't seem to make things more energy efficient with architectures.

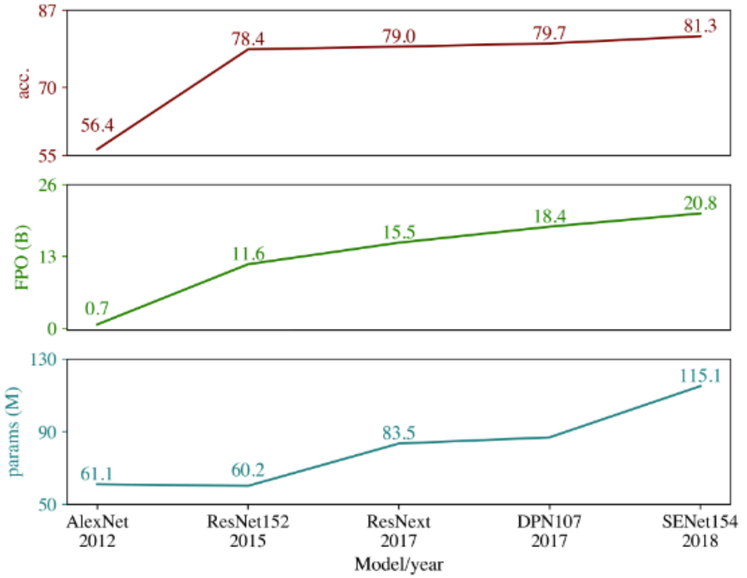
Green AI

Another great paper!
Do we have anything left to add?

Green AI

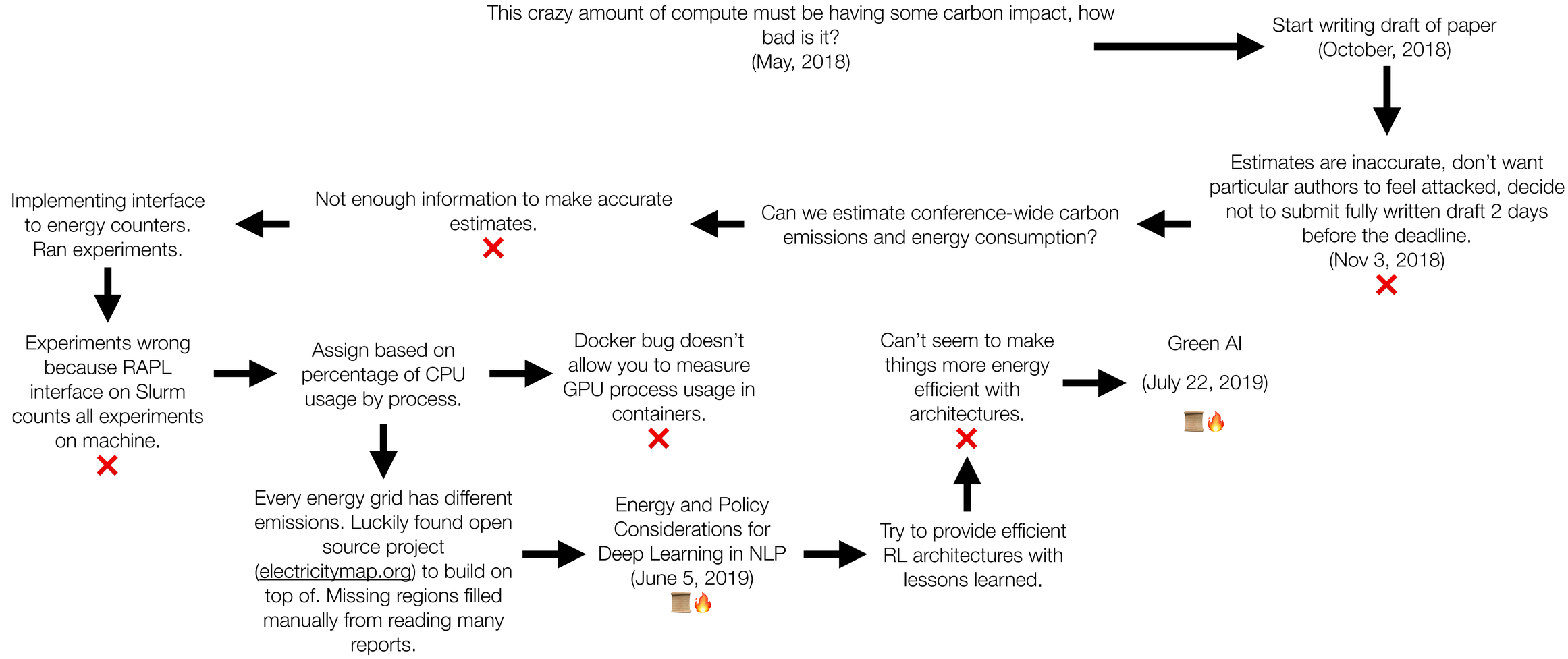
Roy Schwartz*[◇] Jesse Dodge*[◇♣] Noah A. Smith[◇♡] Oren Etzioni[◇]

[◇]Allen Institute for AI, Seattle, Washington, USA
[♣] Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[♡] University of Washington, Seattle, Washington, USA

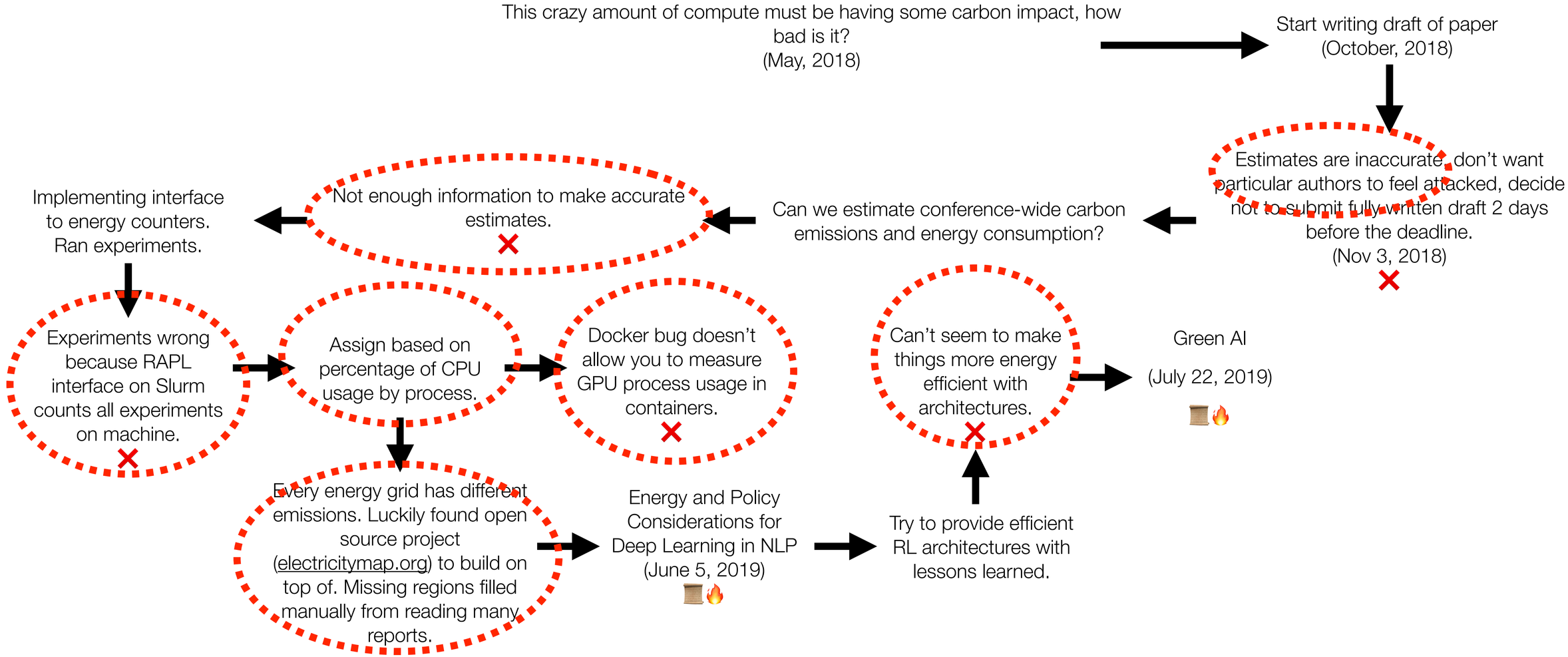


(a) Different models.

Figure 4: Increase in FPO results in diminishing return


This crazy amount of compute must be having some carbon impact, how bad is it?
(May, 2018)
Start writing draft of paper
(October, 2018)
Estimates are inaccurate, don't want particular authors to feel attacked, decide not to submit fully written draft 2 days before the deadline.
(Nov 3, 2018)
Can we estimate conference-wide carbon emissions and energy consumption?
Not enough information to make accurate estimates.
Implementing interface to energy counters. Ran experiments.
Experiments wrong because RAPL interface on Slurm counts all experiments on machine.
Assign based on percentage of CPU usage by process.
Docker bug doesn't allow you to measure GPU process usage in containers.
Every energy grid has different emissions. Luckily found open source project (electricitymap.org) to build on top of. Missing regions filled manually from reading many reports.
Energy and Policy Considerations for Deep Learning in NLP
(June 5, 2019)
Try to provide efficient RL architectures with lessons learned.
Can't seem to make things more energy efficient with architectures.
Green AI
(July 22, 2019)

But we built useful tools and learned a lot that might be helpful for the community!


This crazy amount of compute must be having some carbon impact, how bad is it?
(May, 2018)
Start writing draft of paper
(October, 2018)
Estimates are inaccurate, don't want particular authors to feel attacked, decide not to submit fully-written draft 2 days before the deadline.
(Nov 3, 2018)
Can we estimate conference-wide carbon emissions and energy consumption?
Not enough information to make accurate estimates.
Implementing interface to energy counters. Ran experiments.
Experiments wrong because RAPL interface on Slurm counts all experiments on machine.
Assign based on percentage of CPU usage by process.
Docker bug doesn't allow you to measure GPU process usage in containers.
Every energy grid has different emissions. Luckily found open source project (electricitymap.org) to build on top of. Missing regions filled manually from reading many reports.
Energy and Policy Considerations for Deep Learning in NLP
(June 5, 2019)
Try to provide efficient RL architectures with lessons learned.
Can't seem to make things more energy efficient with architectures.
Green AI
(July 22, 2019)

Iteration #5: Blockers turn into useful information

This crazy amount of compute must be having some carbon impact, how bad is it?
(May, 2018)

Start writing draft of paper
(October, 2018)

Estimates are inaccurate, don't want particular authors to feel attacked, decide not to submit fully written draft 2 days before the deadline.
(Nov 3, 2018)

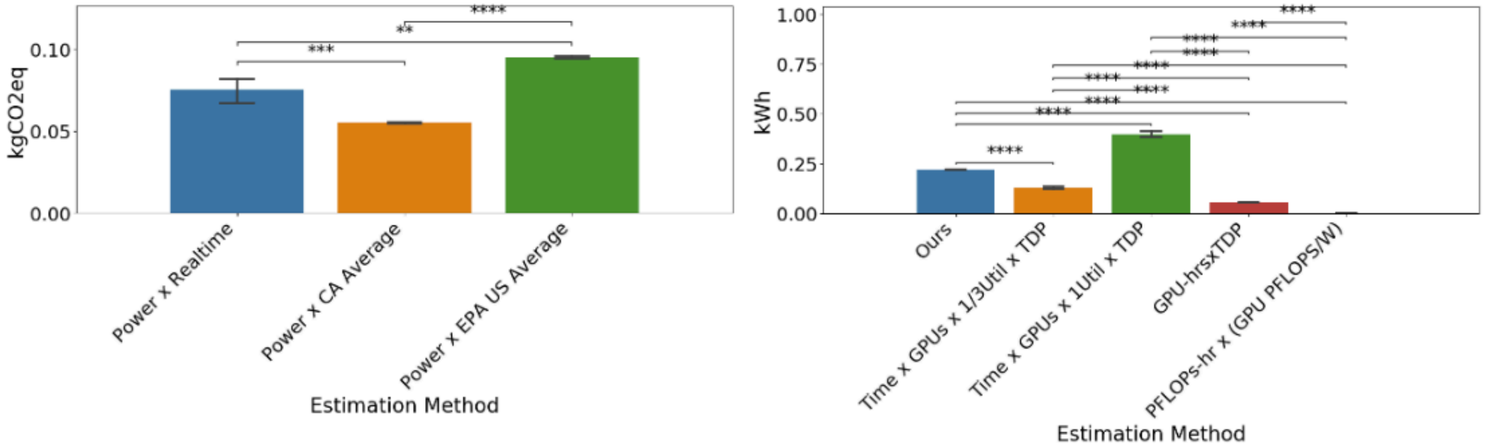


Figure 4: We compare carbon emissions (left) and kWh (right) of our Pong PPO experiment (see Appendix E for more details) by using different estimation methods. By only using country wide or even regional average estimates, carbon emissions may be over or under-estimated (respectively). Similarly, by using partial information to estimate energy usage (right, for more information about the estimation methods see Appendix E), estimates significantly differ from when collecting all data in real time (as in our method). Clearly, without detailed accounting, it is easy to over- or under-estimate carbon or energy emissions in a number of situations. Stars indicate level of significance: * $p < .05$, ** $p < .01$, *** $p < .001$, **** $p < .0001$. Annotation provided via: https://github.com/webermarcolivier/statannot.

Iteration #5: Blockers turn into useful information

This crazy amount of compute must be having some carbon impact, how bad is it?
(May, 2018)

Start writing draft of paper
(October, 2018)

Estimates are inaccurate, don't want particular authors to feel attacked, decide not to submit fully written draft 2 days before the deadline.
(Nov 3, 2018)
✅

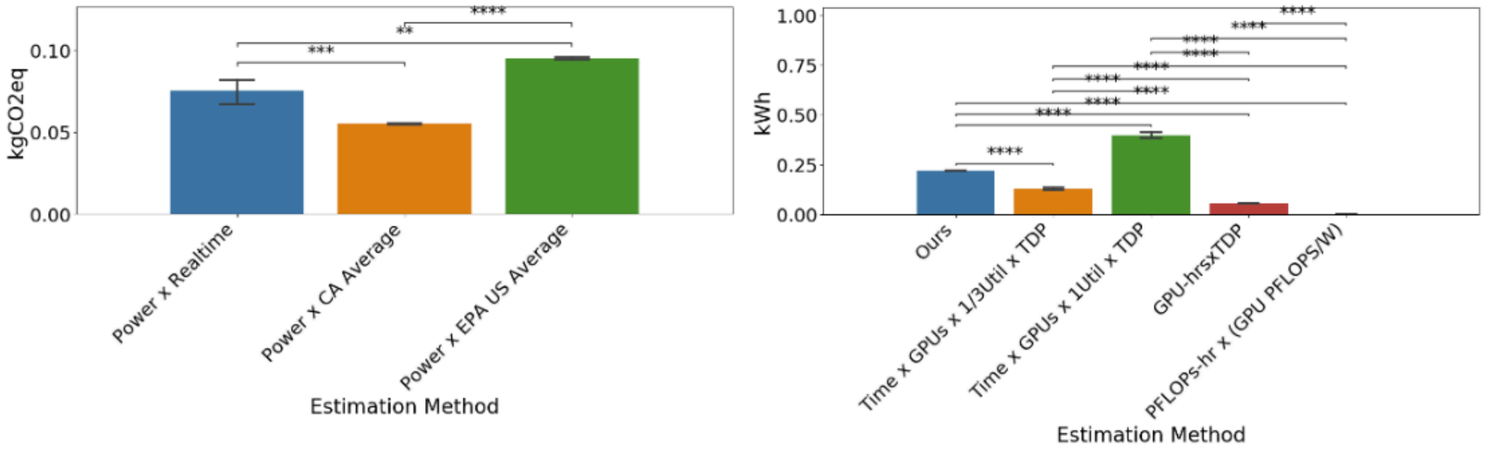
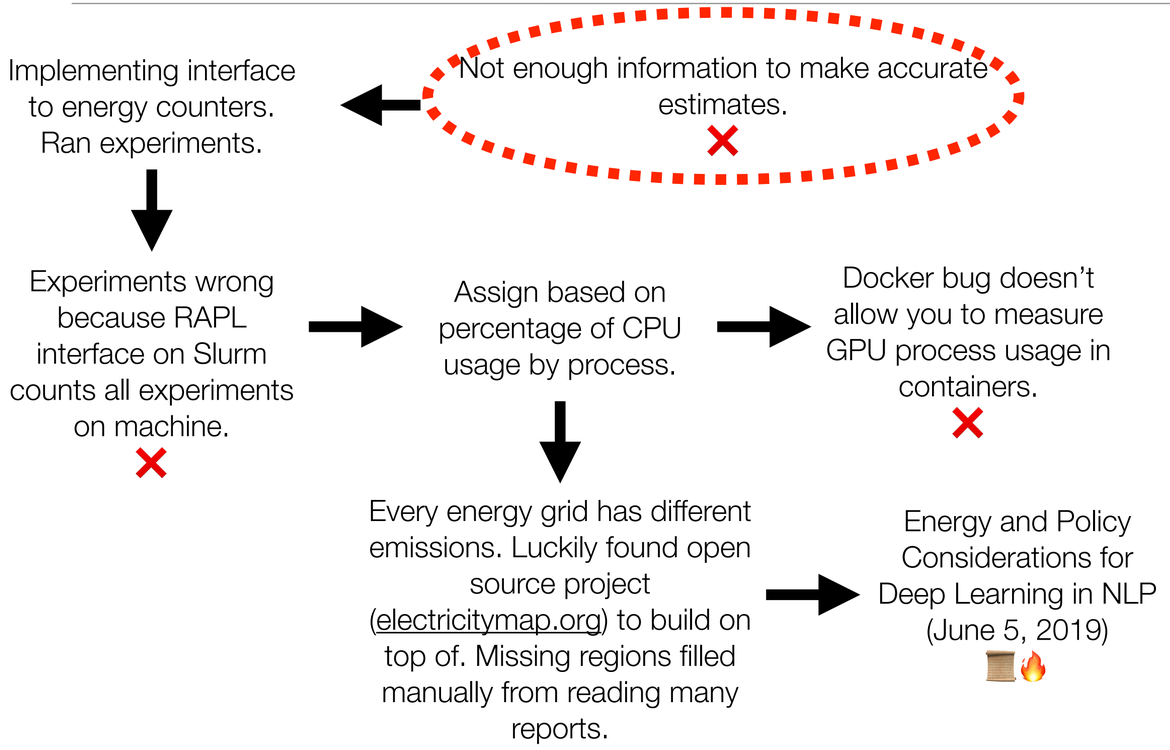


Figure 4: We compare carbon emissions (left) and kWh (right) of our Pong PPO experiment (see Appendix E for more details) by using different estimation methods. By only using country wide or even regional average estimates, carbon emissions may be over or under-estimated (respectively). Similarly, by using partial information to estimate energy usage (right, for more information about the estimation methods see Appendix E), estimates significantly differ from when collecting all data in real time (as in our method). Clearly, without detailed accounting, it is easy to over- or under-estimate carbon or energy emissions in a number of situations. Stars indicate level of significance: * $p < .05$, ** $p < .01$, *** $p < .001$, **** $p < .0001$. Annotation provided via: https://github.com/webermarcolivier/statannot.

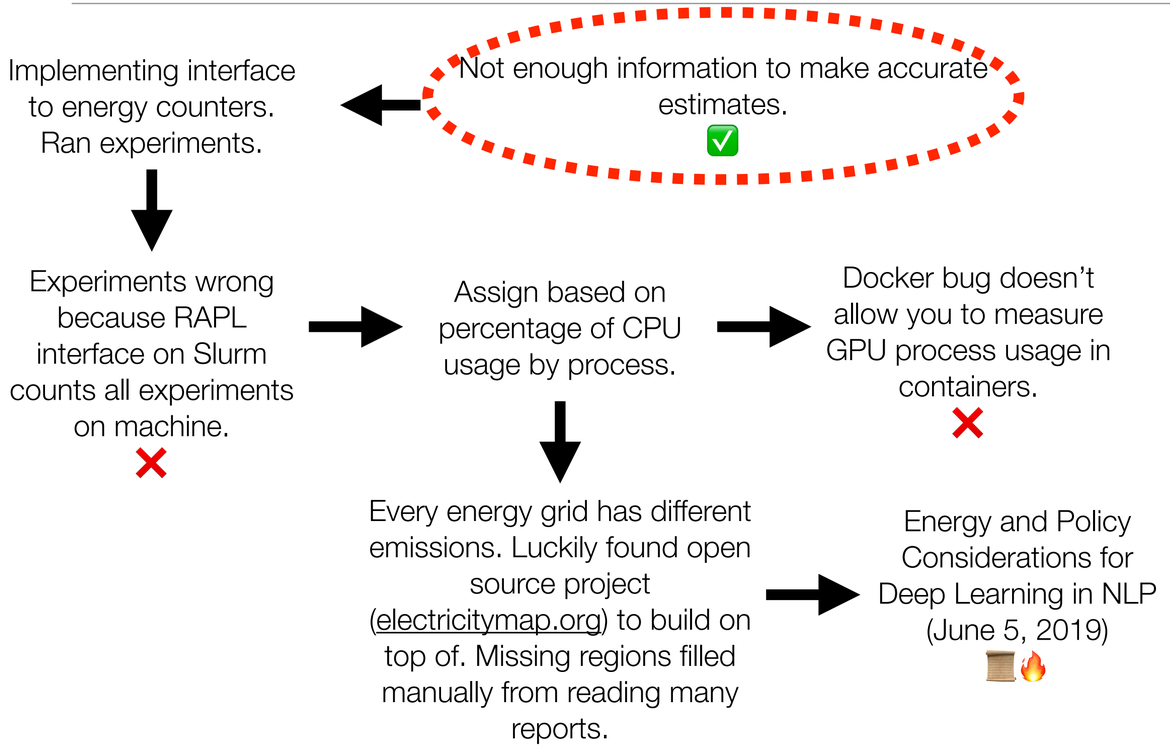
Iteration #5: Blockers turn into useful information



Carbon Impact Statement

This work contributed 8.021 kg of CO_{2eq} to the atmosphere and used 24.344 kWh of electricity, having a USA-specific social cost of carbon of $0.38 ($0.00, $0.95). Carbon accounting information can be found here: https://breakend.github.io/ClimateChangeFromMachineLearningResearch/measuring_and_mitigating_energy_and_carbon_footprints_in_machine_learning/ and https://breakend.github.io/RL-Energy-Leaderboard/reinforcement_learning_energy_leaderboard/index.html. The social cost of carbon uses models from (Ricke et al., 2018). This statement and carbon emissions information was generated using *experiment-impact-tracker* described in this paper.

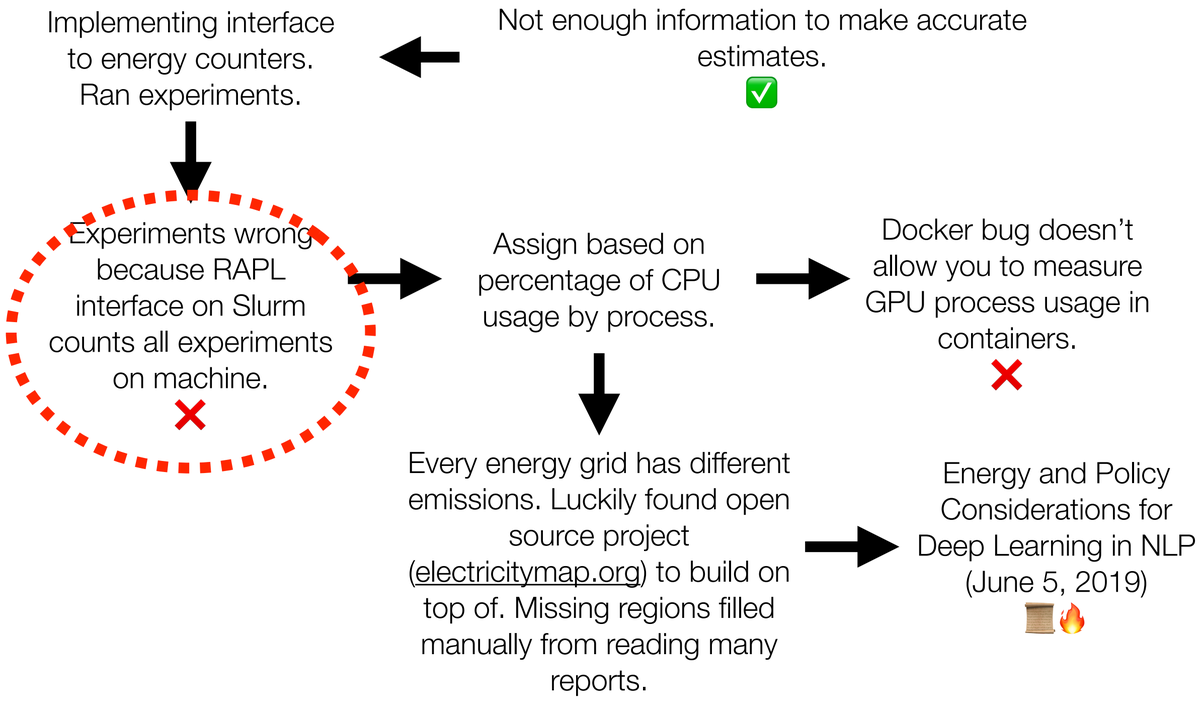
Iteration #5: Blockers turn into useful information



Carbon Impact Statement

This work contributed 8.021 kg of CO_{2eq} to the atmosphere and used 24.344 kWh of electricity, having a USA-specific social cost of carbon of $0.38 ($0.00, $0.95). Carbon accounting information can be found here: https://breakend.github.io/ClimateChangeFromMachineLearningResearch/measuring_and_mitigating_energy_and_carbon_footprints_in_machine_learning/ and https://breakend.github.io/RL-Energy-Leaderboard/reinforcement_learning_energy_leaderboard/index.html. The social cost of carbon uses models from (Ricke et al., 2018). This statement and carbon emissions information was generated using *experiment-impact-tracker* described in this paper.

Iteration #5: Blockers turn into useful information

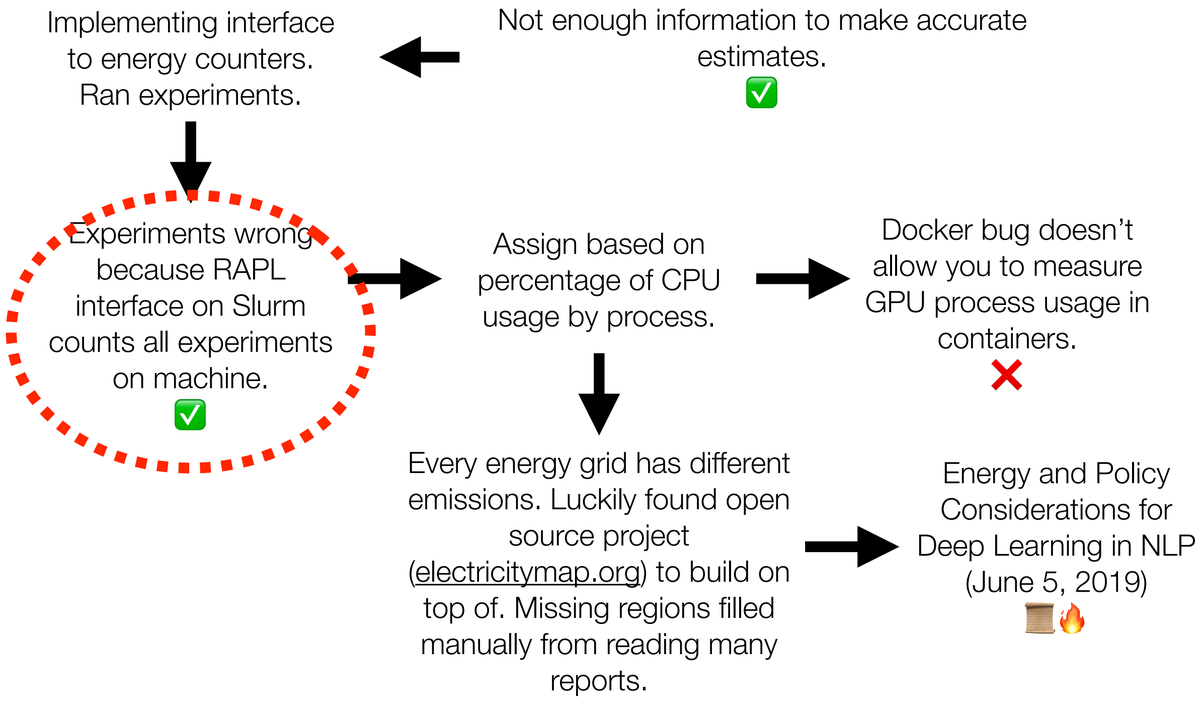


We calculate total energy as:

$$e_{\text{total}} = \text{PUE} \sum_{p} (\text{p}_{\text{dram}} e_{\text{dram}} + \text{p}_{\text{cpu}} e_{\text{cpu}} + \text{p}_{\text{gpu}} e_{\text{gpu}}), \quad (1)$$

where p_{resource} are the percentages of each system resource used by the attributable processes relative to the total in-use resources and e_{resource} is the energy usage of that resource. This is the per-process equivalent of the method which Strubell et al. (2019) use. We assume the same constant power usage effectiveness (PUE) as Strubell et al. (2019). This value compensates for excess energy from cooling or heating the data-center.

Iteration #5: Blockers turn into useful information

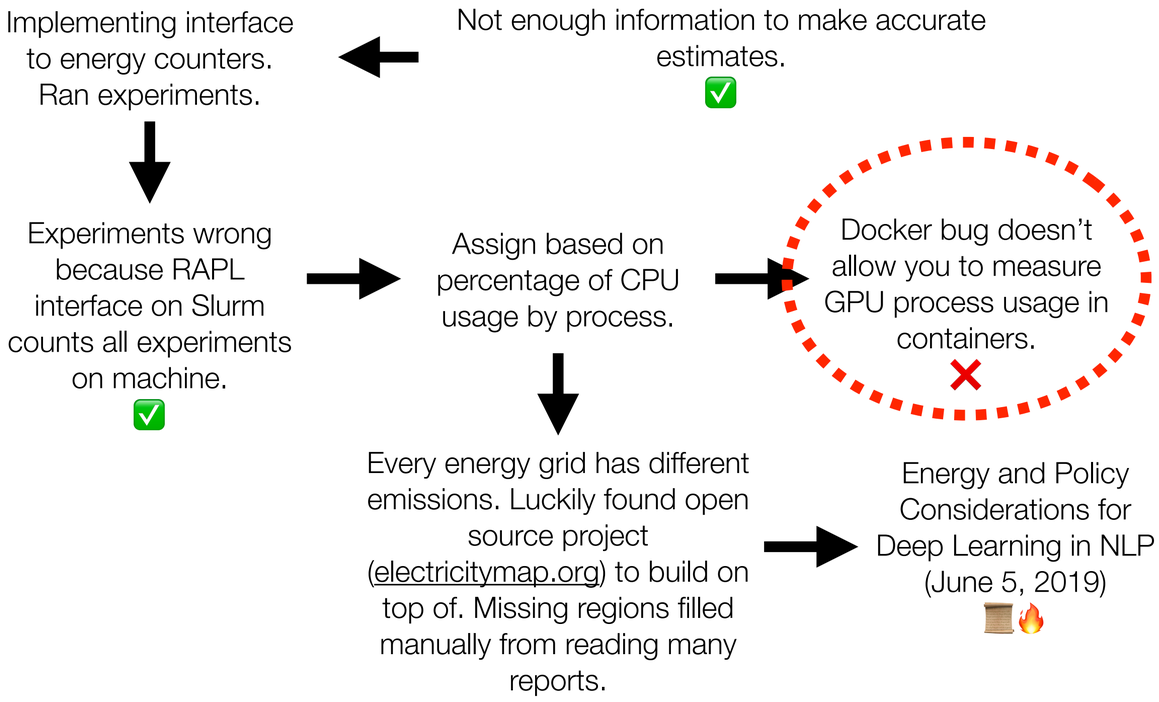


We calculate total energy as:

$$e_{\text{total}} = \text{PUE} \sum_p (\text{p}_{\text{dram}} e_{\text{dram}} + \text{p}_{\text{cpu}} e_{\text{cpu}} + \text{p}_{\text{gpu}} e_{\text{gpu}}), \quad (1)$$

where p_{resource} are the percentages of each system resource used by the attributable processes relative to the total in-use resources and e_{resource} is the energy usage of that resource. This is the per-process equivalent of the method which Strubell et al. (2019) use. We assume the same constant power usage effectiveness (PUE) as Strubell et al. (2019). This value compensates for excess energy from cooling or heating the data-center.

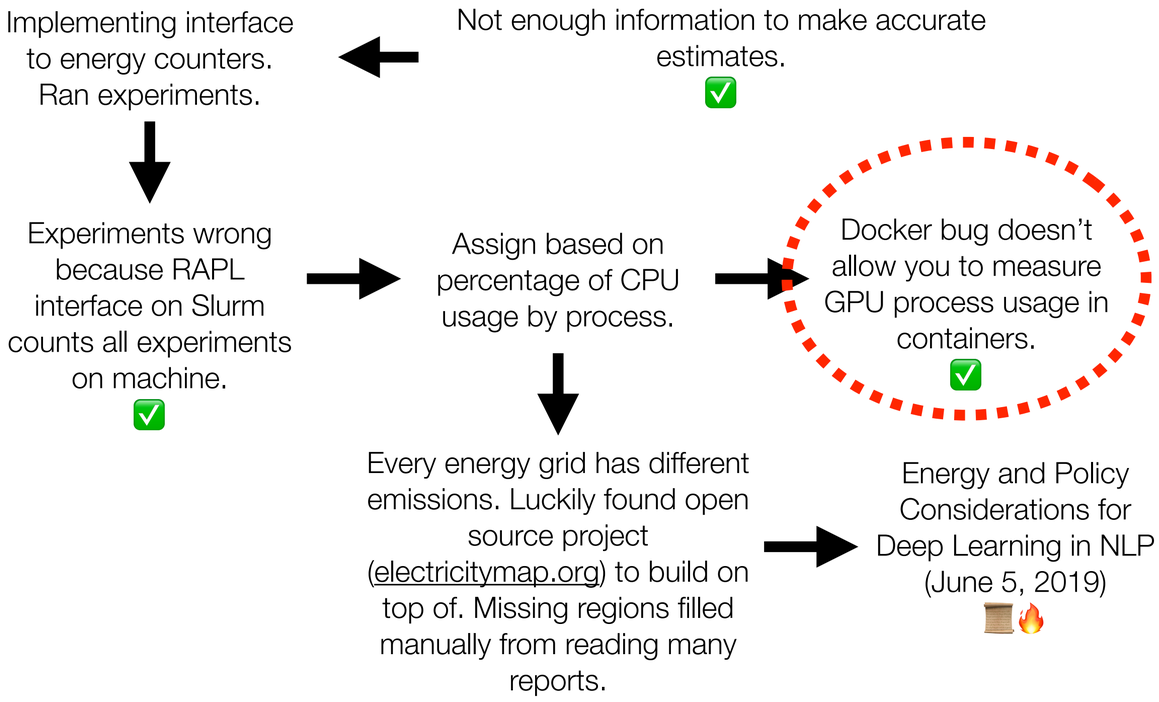
Iteration #5: Blockers turn into useful information



7.7 Driver and Implementation Difficulties

The *experiment-impact-tracker* framework abstracts away many of the previously mentioned difficulties in estimating carbon and energy impacts: it handles routing to appropriate tools for collecting information, aggregates information across tools to handle carbon calculations, finds carbon intensity information automatically, and corrects for multiple processes on one machine. Yet, a few other challenges may be hidden by using the framework which remain difficult to circumvent.

Iteration #5: Blockers turn into useful information



7.7 Driver and Implementation Difficulties

The *experiment-impact-tracker* framework abstracts away many of the previously mentioned difficulties in estimating carbon and energy impacts: it handles routing to appropriate tools for collecting information, aggregates information across tools to handle carbon calculations, finds carbon intensity information automatically, and corrects for multiple processes on one machine. Yet, a few other challenges may be hidden by using the framework which remain difficult to circumvent.

Iteration #5: Blockers turn into useful information

Implementing interface to energy counters. Ran experiments.

Not enough information to make accurate estimates. ✅

Experiments wrong because RAPL interface on Slurm counts all experiments on machine. ✅

Assign based on percentage of CPU usage by process.

Docker bug doesn't allow you to measure GPU process usage in containers. ✅

Every energy grid has different emissions. Luckily found open source project (electricitymap.org) to build on top of. Missing regions filled manually from reading many reports.

Energy and Policy Considerations for Deep Learning in NLP (June 5, 2019) 📜🔥

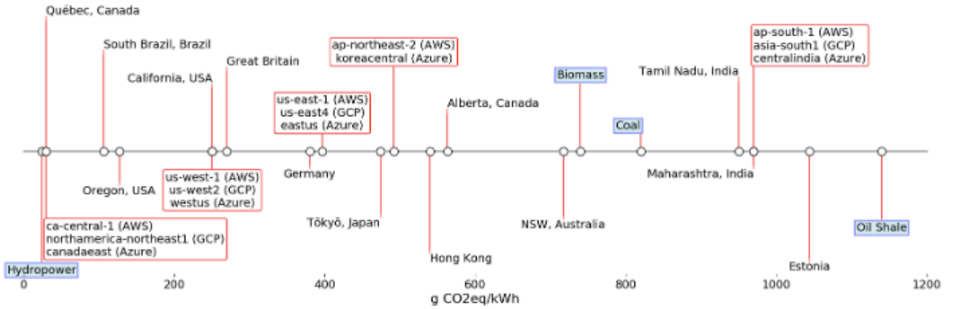


Figure 6: Carbon Intensity (gCO_{2eq}/kWh) of selected energy grid regions is shown from least carbon emissions (left) to most carbon emissions (right). Red/unshaded boxes indicate carbon intensities of cloud provider regions. Blue/shaded boxes indicate carbon intensities of various generation methods. Oil shale is the most carbon emitting method of energy production in the Figure. Estonia is powered mainly by oil shale and thus is close to it in carbon intensity. Similarly, Québec is mostly powered by hydroelectric methods and is close to it in carbon intensity. Cloud provider carbon intensities are based on the regional energy grid in which they are located. Thus, us-west-1, located in California, has the same carbon intensity as the state. See `https://github.com/Breakend/experiment-impact-tracker/` for data sources of regional information. Energy source information from Krey et al. (2014); International Energy Agency (2015).

Iteration #5: Blockers turn into useful information

Implementing interface to energy counters. Ran experiments.

Not enough information to make accurate estimates. ✅

Experiments wrong because RAPL interface on Slurm counts all experiments on machine. ✅

Assign based on percentage of CPU usage by process.

Docker bug doesn't allow you to measure GPU process usage in containers. ✅

Every energy grid has different emissions. Luckily found open source project (electricitymap.org) to build on top of. Missing regions filled manually from reading many reports. ✅

Energy and Policy Considerations for Deep Learning in NLP (June 5, 2019) 📜🔥

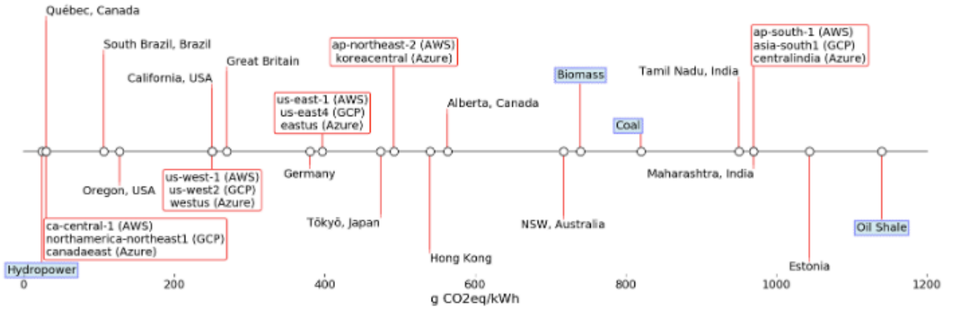


Figure 6: Carbon Intensity (gCO_{2eq}/kWh) of selected energy grid regions is shown from least carbon emissions (left) to most carbon emissions (right). Red/unshaded boxes indicate carbon intensities of cloud provider regions. Blue/shaded boxes indicate carbon intensities of various generation methods. Oil shale is the most carbon emitting method of energy production in the Figure. Estonia is powered mainly by oil shale and thus is close to it in carbon intensity. Similarly, Québec is mostly powered by hydroelectric methods and is close to it in carbon intensity. Cloud provider carbon intensities are based on the regional energy grid in which they are located. Thus, us-west-1, located in California, has the same carbon intensity as the state. See `https://github.com/Breakend/experiment-impact-tracker/` for data sources of regional information. Energy source information from Krey et al. (2014); International Energy Agency (2015).

Iteration #5: Blockers turn into useful information

Docker bug doesn't allow you to measure GPU process usage in containers.

❌

Can't seem to make things more energy efficient with architectures.

❌

Green AI

(July 22, 2019)

📜🔥

Energy and Policy Considerations for Deep Learning in NLP (June 5, 2019)

→

Try to provide efficient RL architectures with lessons learned.

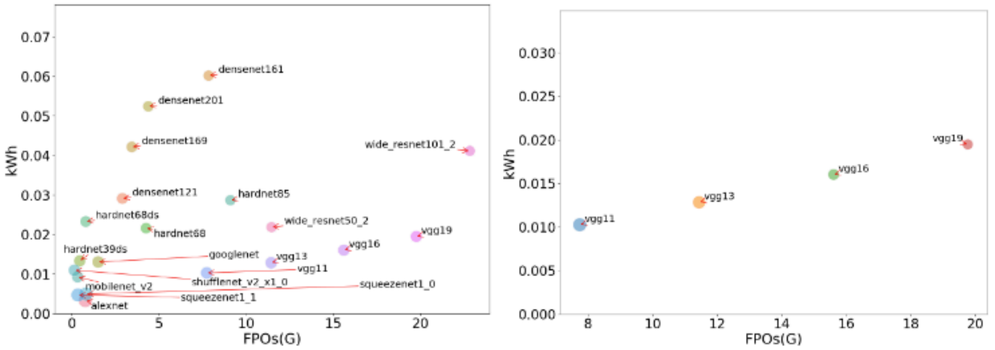


Figure 3: We run 50,000 rounds of inference on a single sampled image through pre-trained image classification models and record kWh, experiment time, FPOs, and number of parameters (repeating 4 times on different random seeds). References for models, code, and expanded experiment details can be found in Appendix D. We run a similar analysis to Canziani et al. (2016) and find (left) that FPOs are not strongly correlated with energy consumption ($R^2 = 0.083$, Pearson 0.289) nor with time ($R^2 = 0.005$, Pearson -0.074) when measured across different architectures. However, within an architecture (right) correlations are much stronger. Only considering different versions of VGG, FPOs are strongly correlated with energy ($R^2 = .999$, Pearson 1.0) and time ($R^2 = .998$, Pearson .999). Comparing parameters against energy yields similar results (see Appendix D for these results and plots against experiment runtime).

Iteration #5: Blockers turn into useful information

Docker bug doesn't allow you to measure GPU process usage in containers.

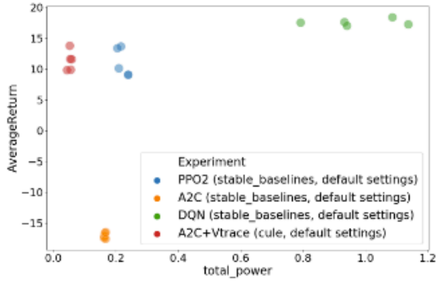
Can't seem to make things more energy efficient with architectures.

Green AI

(July 22, 2019)

Energy and Policy Considerations for Deep Learning in NLP (June 5, 2019)

Try to provide efficient RL architectures with lessons learned.



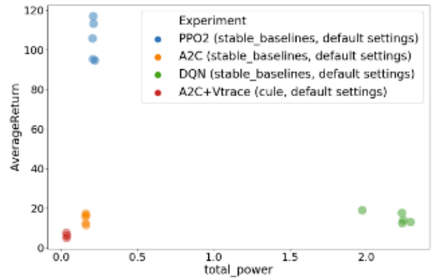


Figure 5: We evaluate A2C, PPO, DQN, and A2C+VTraces on PongNoFrameskip-v4 (left) and BreakoutNoFrameskip-v4 (right), two common evaluation environments included in OpenAI Gym. We train for only 5M timesteps, less than prior work, to encourage energy efficiency and evaluate for 25 episodes every 250k timesteps. We show the Average Return across all evaluations throughout training (giving some measure of both ability and speed of convergence of an algorithm) as compared to the total energy in kWh. Weighted rankings of Average Return per kWh place A2C+Vtrace first on Pong and PPO first on Breakout. Using PPO versus DQN can yield significant energy savings, while retaining performance on both environments (in the 5M samples regime). See Appendix F for more details and results in terms of asymptotic performance.

Iteration #5: Blockers turn into useful information

Docker bug doesn't allow you to measure GPU process usage in containers.

❌

Can't seem to make things more energy efficient with architectures.

✅

Green AI

(July 22, 2019)

Energy and Policy Considerations for Deep Learning in NLP (June 5, 2019)

Try to provide efficient RL architectures with lessons learned.

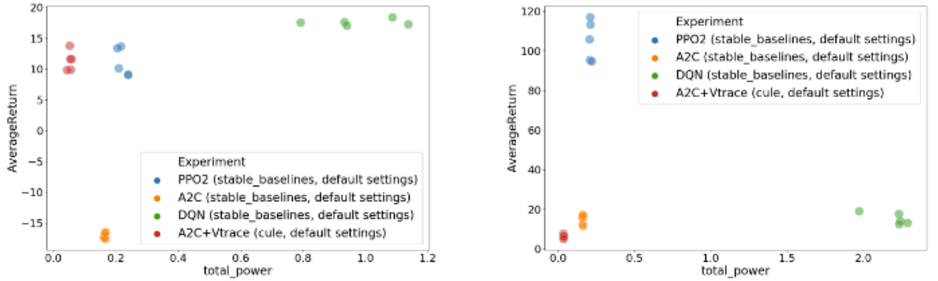
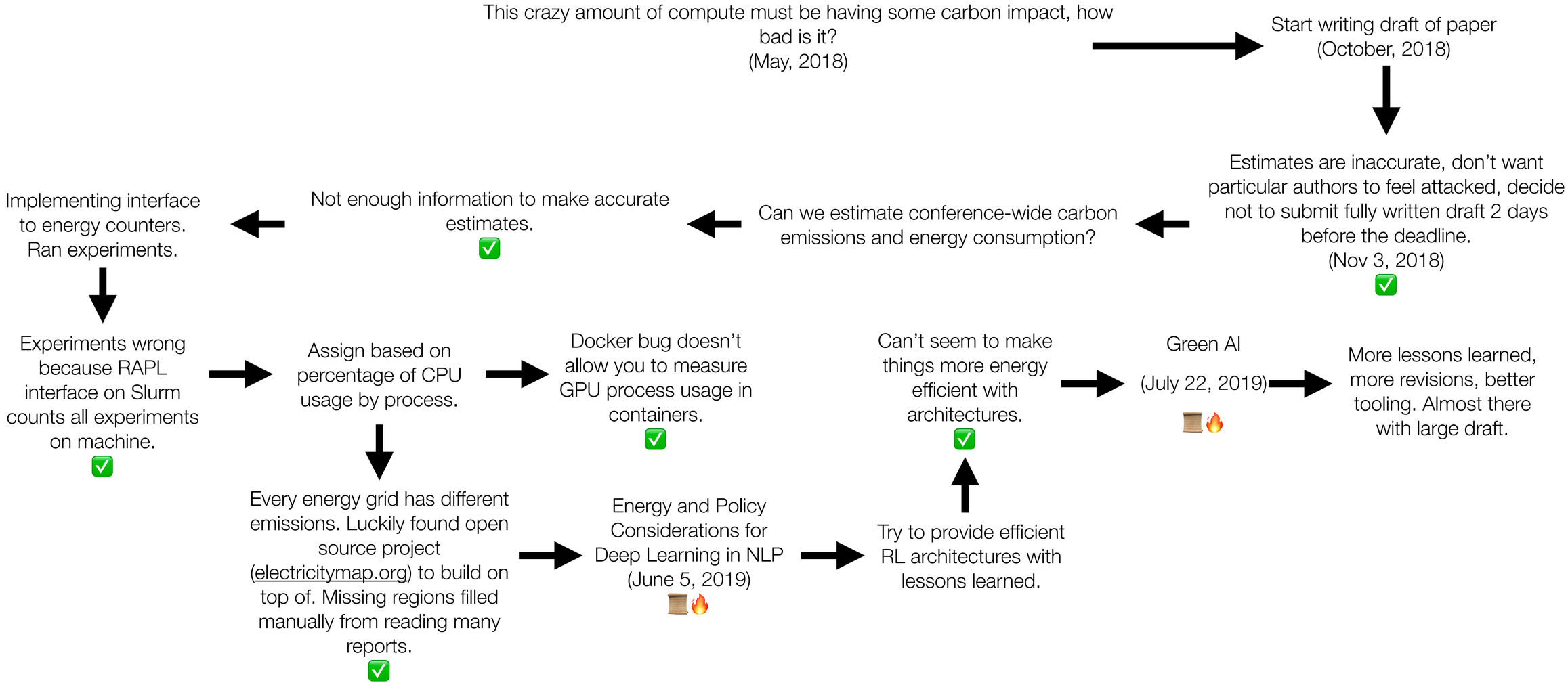
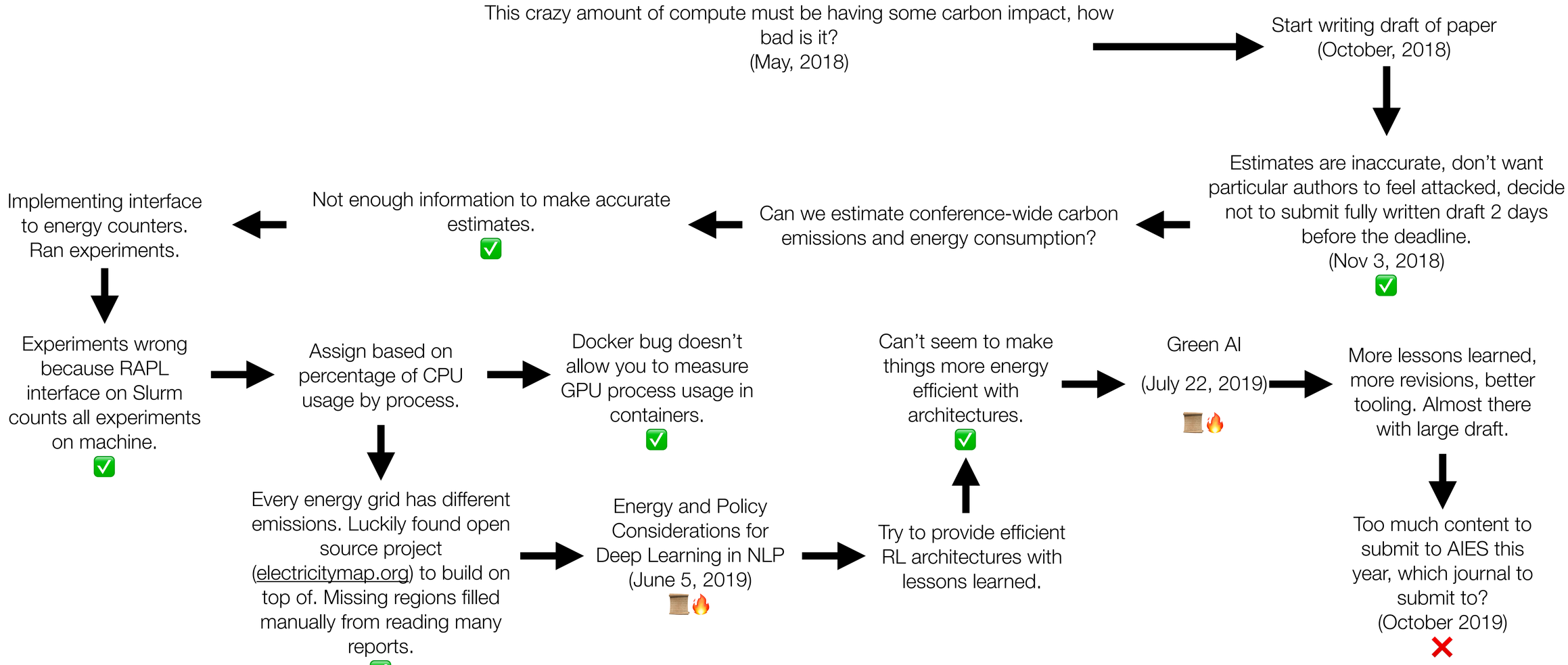


Figure 5: We evaluate A2C, PPO, DQN, and A2C+VTraces on PongNoFrameskip-v4 (left) and BreakoutNoFrameskip-v4 (right), two common evaluation environments included in OpenAI Gym. We train for only 5M timesteps, less than prior work, to encourage energy efficiency and evaluate for 25 episodes every 250k timesteps. We show the Average Return across all evaluations throughout training (giving some measure of both ability and speed of convergence of an algorithm) as compared to the total energy in kWh. Weighted rankings of Average Return per kWh place A2C+Vtrace first on Pong and PPO first on Breakout. Using PPO versus DQN can yield significant energy savings, while retaining performance on both environments (in the 5M samples regime). See Appendix F for more details and results in terms of asymptotic performance.


This crazy amount of compute must be having some carbon impact, how bad is it?
(May, 2018)
Start writing draft of paper
(October, 2018)
Estimates are inaccurate, don't want particular authors to feel attacked, decide not to submit fully written draft 2 days before the deadline.
(Nov 3, 2018)
✅
Can we estimate conference-wide carbon emissions and energy consumption?
Not enough information to make accurate estimates.
✅
Implementing interface to energy counters. Ran experiments.
Experiments wrong because RAPL interface on Slurm counts all experiments on machine.
✅
Assign based on percentage of CPU usage by process.
Docker bug doesn't allow you to measure GPU process usage in containers.
✅
Every energy grid has different emissions. Luckily found open source project (electricitymap.org) to build on top of. Missing regions filled manually from reading many reports.
✅
Energy and Policy Considerations for Deep Learning in NLP
(June 5, 2019)
📜🔥
Try to provide efficient RL architectures with lessons learned.
Can't seem to make things more energy efficient with architectures.
✅
Green AI
(July 22, 2019)
📜🔥
More lessons learned, more revisions, better tooling. Almost there with large draft.


This crazy amount of compute must be having some carbon impact, how bad is it?
(May, 2018)
Start writing draft of paper
(October, 2018)
Estimates are inaccurate, don't want particular authors to feel attacked, decide not to submit fully written draft 2 days before the deadline.
(Nov 3, 2018)
✅
Can we estimate conference-wide carbon emissions and energy consumption?
Not enough information to make accurate estimates.
✅
Implementing interface to energy counters. Ran experiments.
Experiments wrong because RAPL interface on Slurm counts all experiments on machine.
✅
Assign based on percentage of CPU usage by process.
Docker bug doesn't allow you to measure GPU process usage in containers.
✅
Every energy grid has different emissions. Luckily found open source project (electricitymap.org) to build on top of. Missing regions filled manually from reading many reports.
✅
Energy and Policy Considerations for Deep Learning in NLP
(June 5, 2019)
📜🔥
Try to provide efficient RL architectures with lessons learned.
Can't seem to make things more energy efficient with architectures.
✅
Green AI
(July 22, 2019)
📜🔥
More lessons learned, more revisions, better tooling. Almost there with large draft.
Too much content to submit to AIES this year, which journal to submit to?
(October 2019)
❌

Quantifying the Carbon Emissions of Machine Learning

Another great paper!

Quantifying the Carbon Emissions of Machine Learning

Alexandre Lacoste*
Element AI
allac@elementai.com

Alexandra Luccioni*
Mila, Université de Montréal
luccionis@mila.quebec

Victor Schmidt*
Mila, Université de Montréal
schmidtv@mila.quebec

Thomas Dandres
Polytechnique Montréal, CIRAIG
thomas.dandres@polymtl.ca

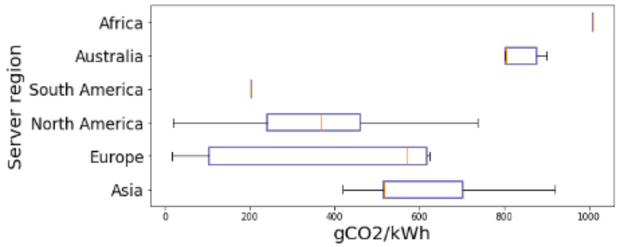
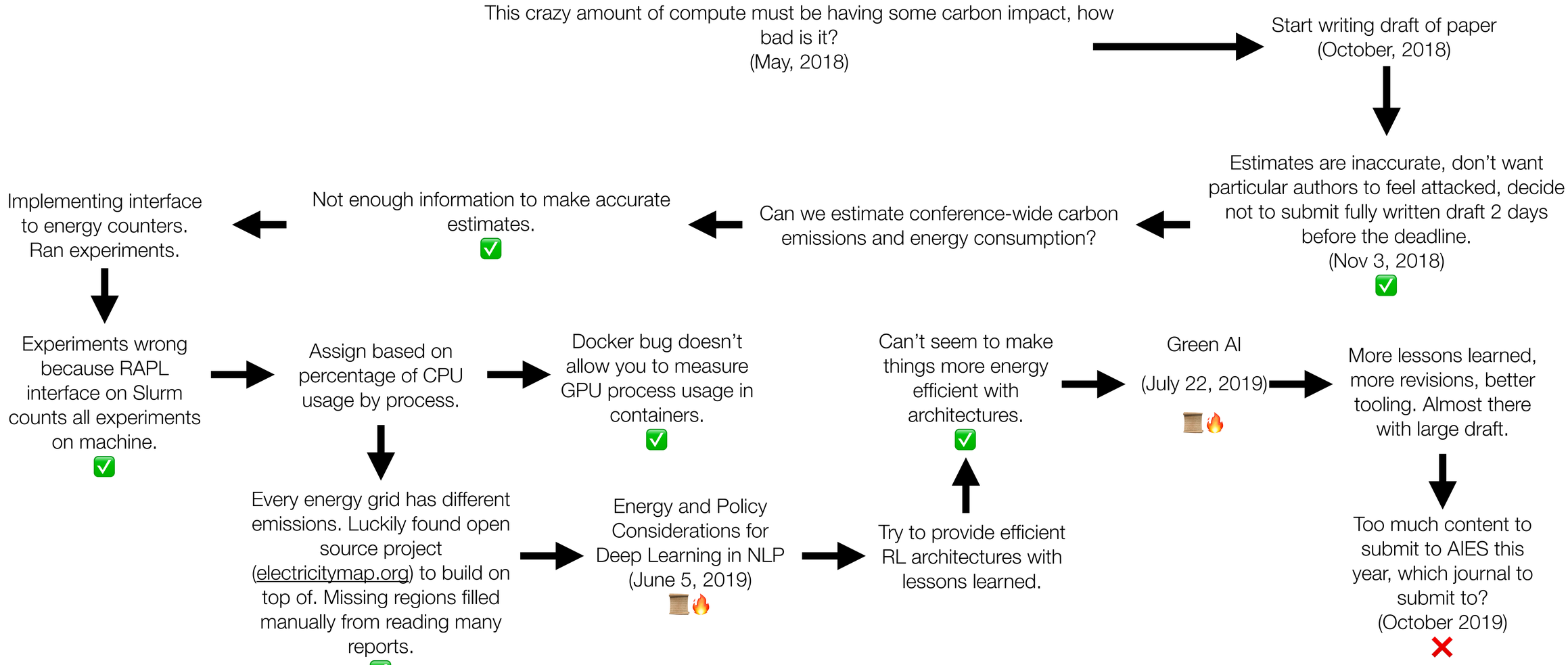
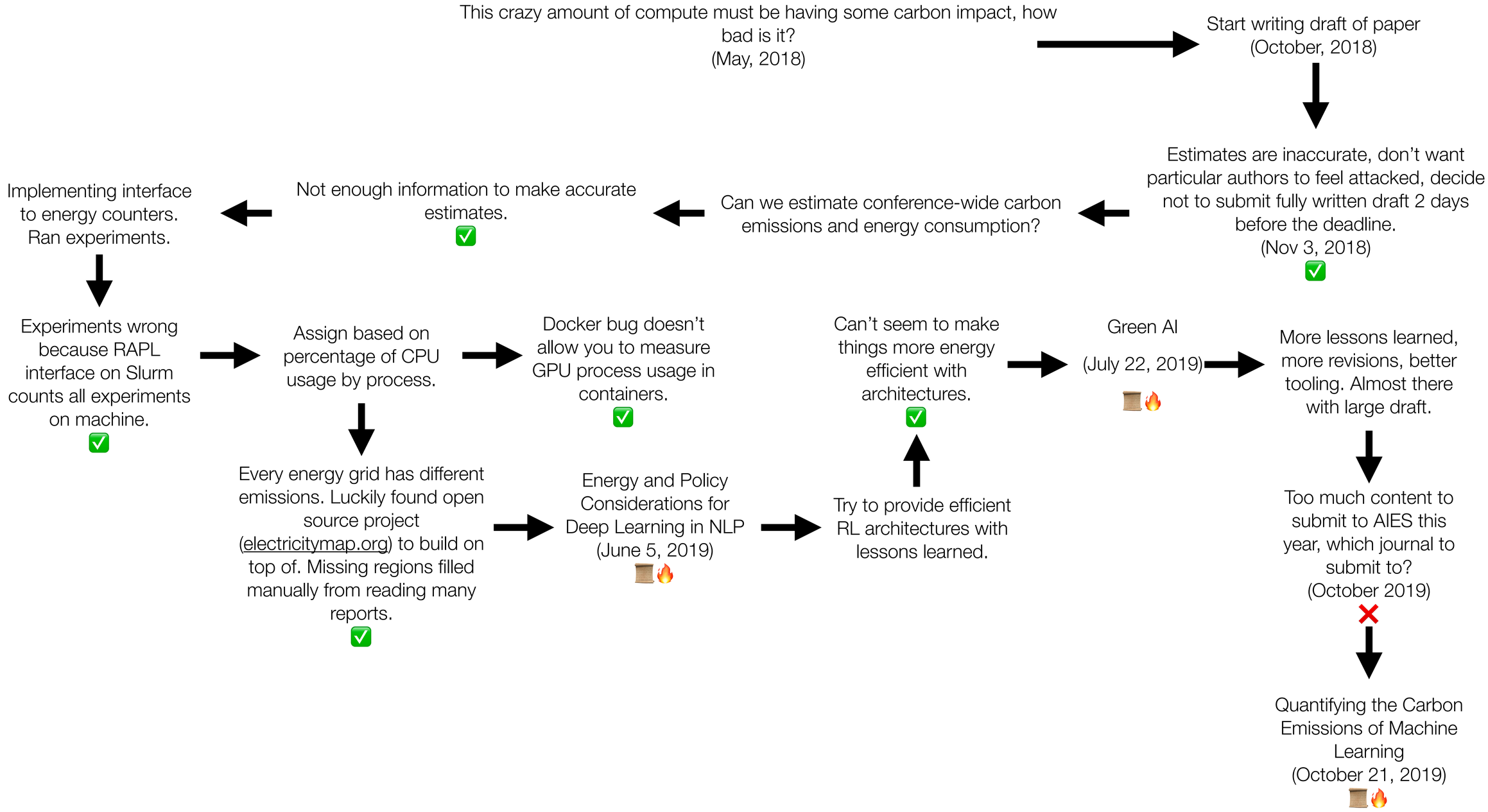
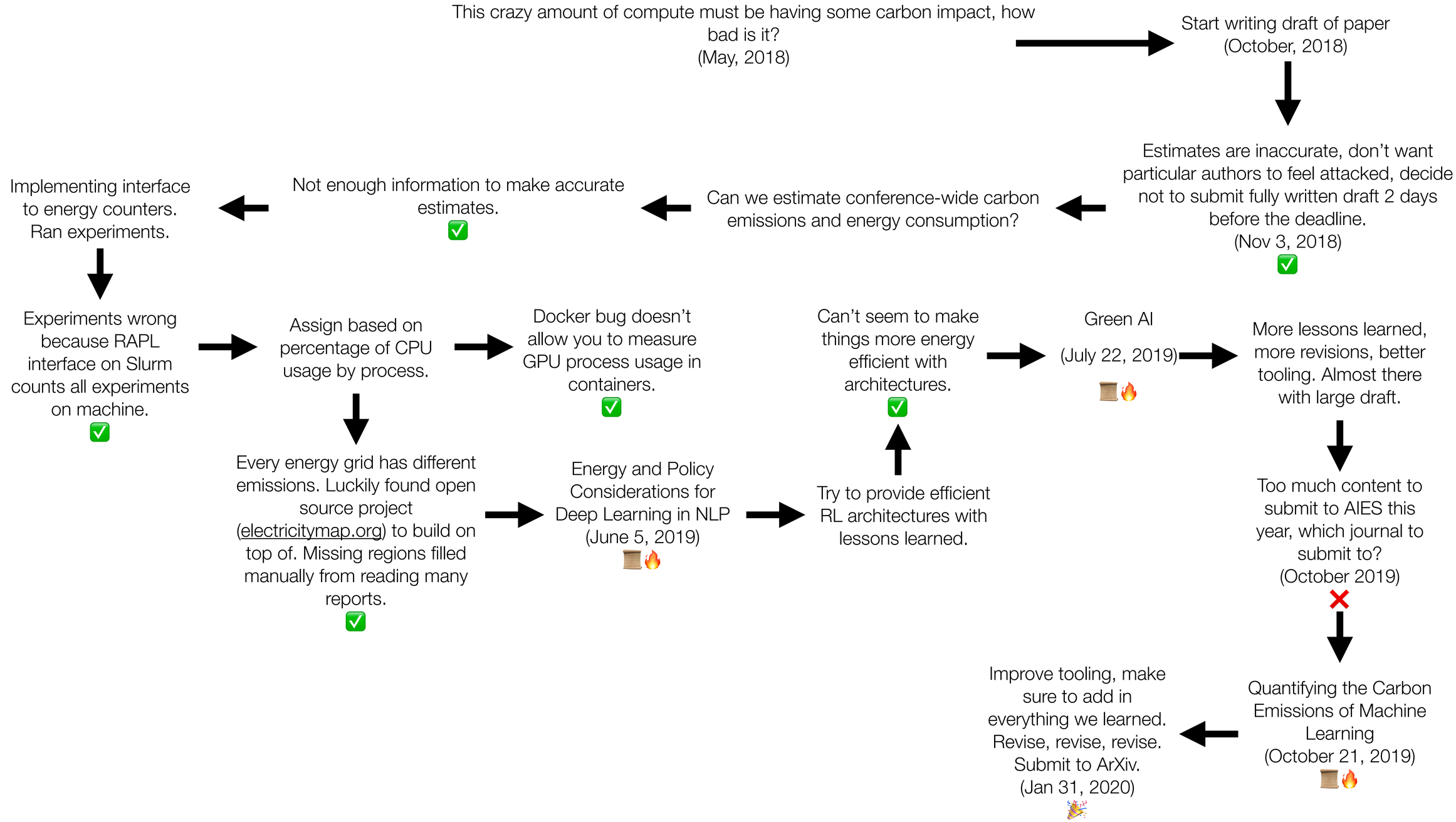
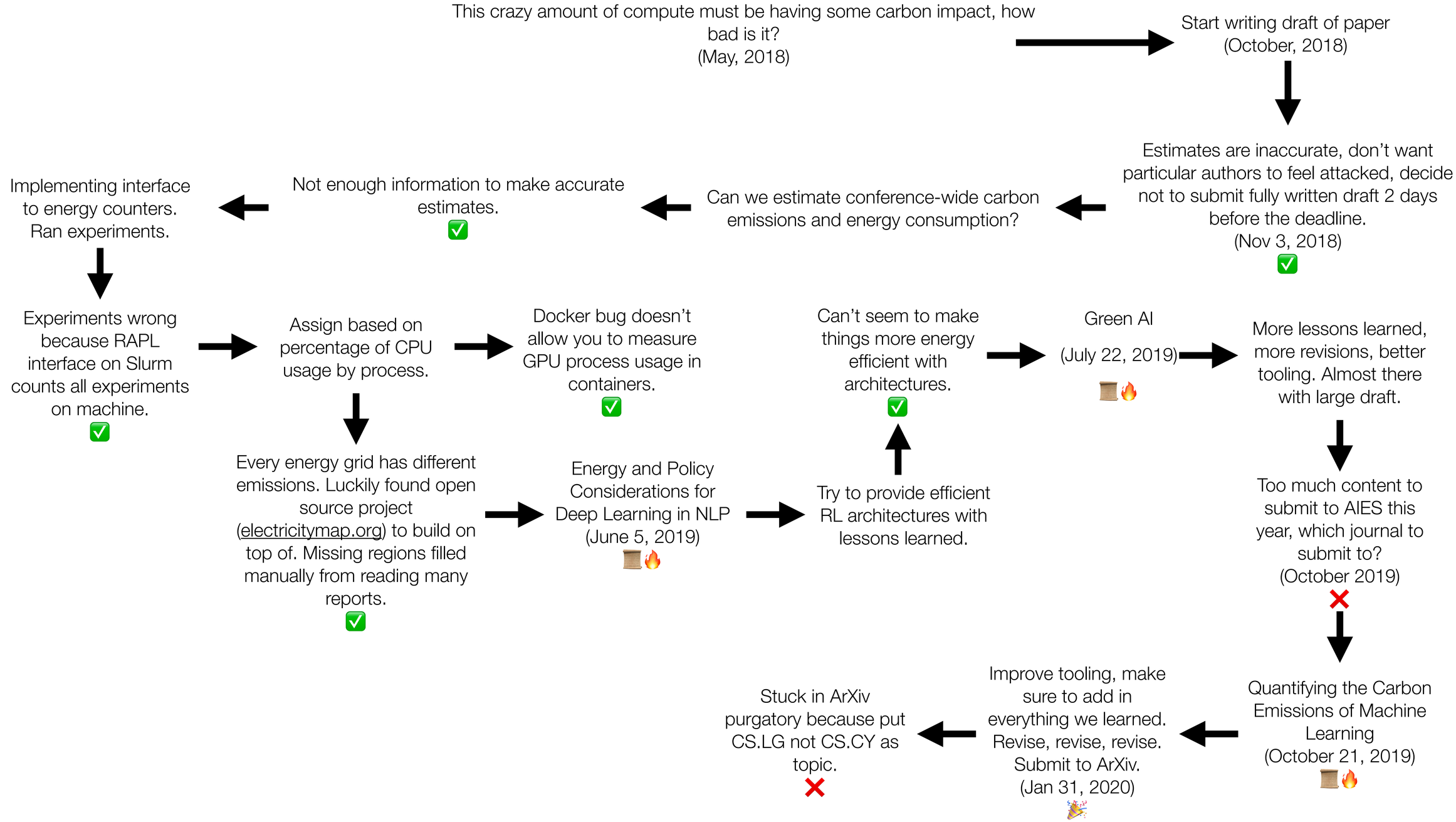


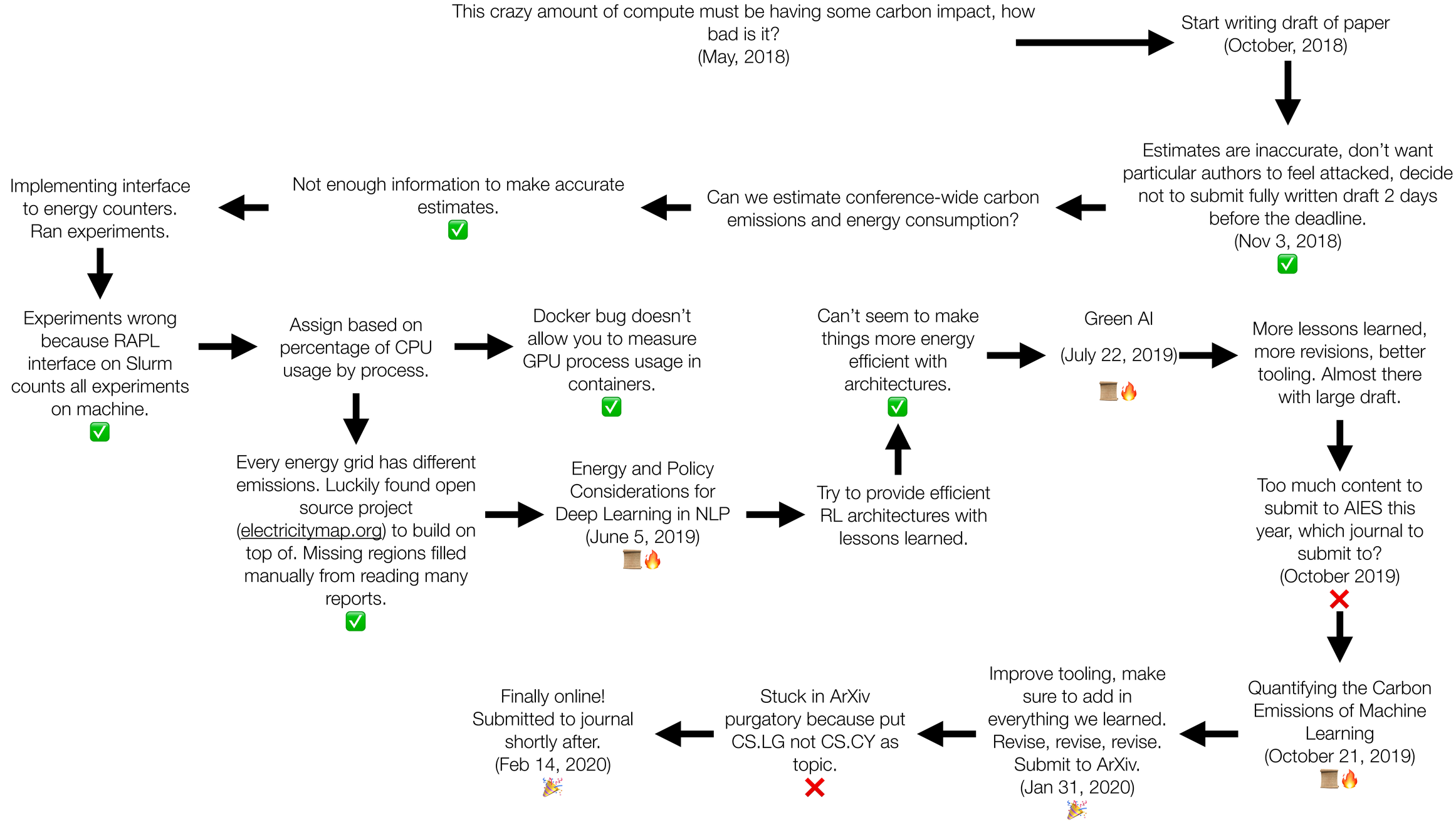
Figure 1: Variation of the Average Carbon Intensity of Servers Worldwide, by Region. (Vertical bars represent regions with a single available data point.)

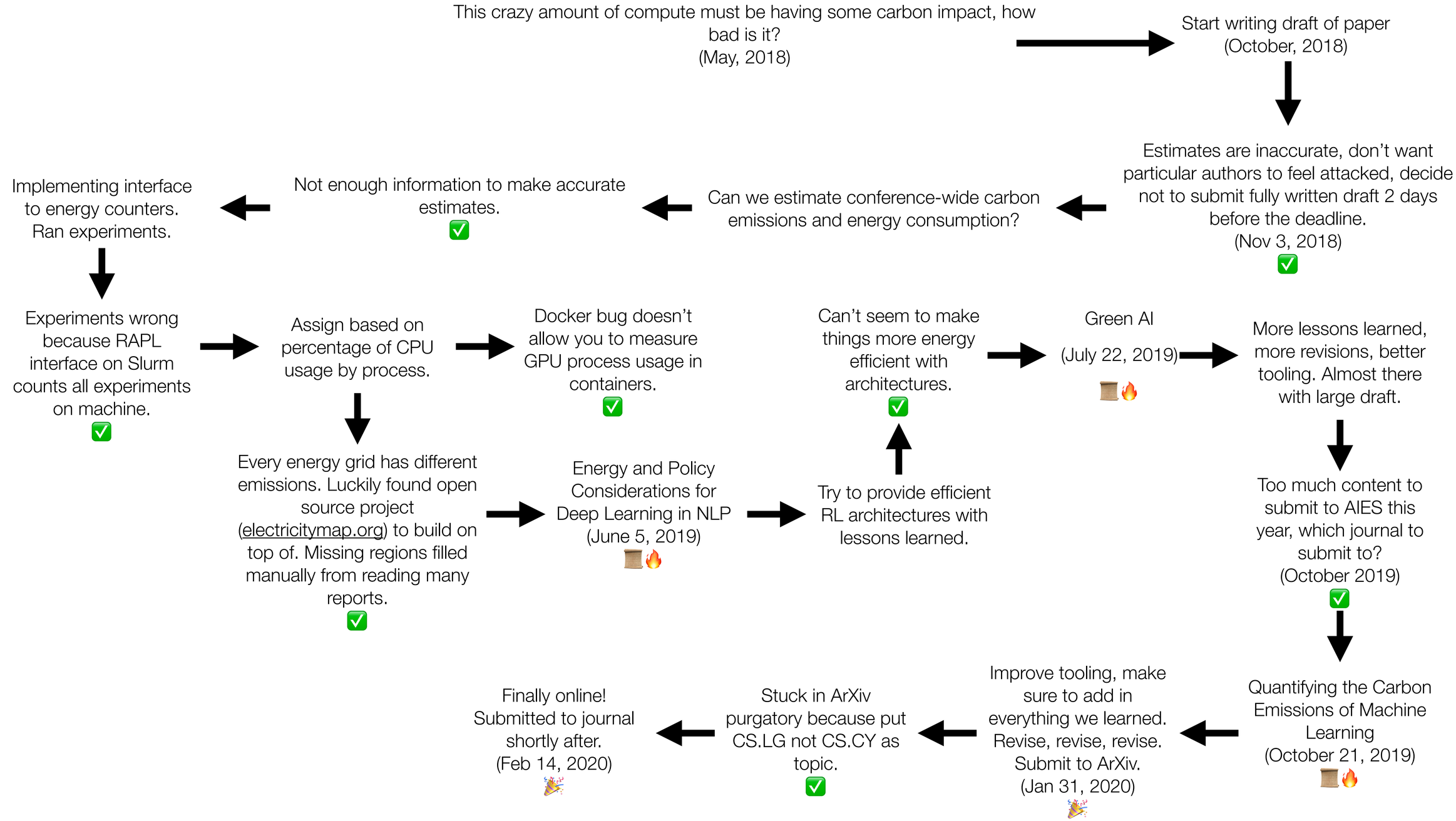

This crazy amount of compute must be having some carbon impact, how bad is it?
(May, 2018)
Start writing draft of paper
(October, 2018)
Estimates are inaccurate, don't want particular authors to feel attacked, decide not to submit fully written draft 2 days before the deadline.
(Nov 3, 2018)
✅
Can we estimate conference-wide carbon emissions and energy consumption?
Not enough information to make accurate estimates.
✅
Implementing interface to energy counters. Ran experiments.
Experiments wrong because RAPL interface on Slurm counts all experiments on machine.
✅
Assign based on percentage of CPU usage by process.
Docker bug doesn't allow you to measure GPU process usage in containers.
✅
Every energy grid has different emissions. Luckily found open source project (electricitymap.org) to build on top of. Missing regions filled manually from reading many reports.
✅
Energy and Policy Considerations for Deep Learning in NLP
(June 5, 2019)
📜🔥
Try to provide efficient RL architectures with lessons learned.
Can't seem to make things more energy efficient with architectures.
✅
Green AI
(July 22, 2019)
📜🔥
More lessons learned, more revisions, better tooling. Almost there with large draft.
Too much content to submit to AIES this year, which journal to submit to?
(October 2019)
❌


This crazy amount of compute must be having some carbon impact, how bad is it?
(May, 2018)
Start writing draft of paper
(October, 2018)
Estimates are inaccurate, don't want particular authors to feel attacked, decide not to submit fully written draft 2 days before the deadline.
(Nov 3, 2018)
✅
Can we estimate conference-wide carbon emissions and energy consumption?
Not enough information to make accurate estimates.
✅
Implementing interface to energy counters. Ran experiments.
Experiments wrong because RAPL interface on Slurm counts all experiments on machine.
✅
Assign based on percentage of CPU usage by process.
Docker bug doesn't allow you to measure GPU process usage in containers.
✅
Every energy grid has different emissions. Luckily found open source project (electricitymap.org) to build on top of. Missing regions filled manually from reading many reports.
✅
Energy and Policy Considerations for Deep Learning in NLP
(June 5, 2019)
📜🔥
Try to provide efficient RL architectures with lessons learned.
Can't seem to make things more energy efficient with architectures.
✅
Green AI
(July 22, 2019)
📜🔥
More lessons learned, more revisions, better tooling. Almost there with large draft.
Too much content to submit to AIES this year, which journal to submit to?
(October 2019)
❌
Quantifying the Carbon Emissions of Machine Learning
(October 21, 2019)
📜🔥


This crazy amount of compute must be having some carbon impact, how bad is it?
(May, 2018)
Start writing draft of paper
(October, 2018)
Estimates are inaccurate, don't want particular authors to feel attacked, decide not to submit fully written draft 2 days before the deadline.
(Nov 3, 2018)
✅
Can we estimate conference-wide carbon emissions and energy consumption?
Not enough information to make accurate estimates.
✅
Implementing interface to energy counters. Ran experiments.
Experiments wrong because RAPL interface on Slurm counts all experiments on machine.
✅
Assign based on percentage of CPU usage by process.
Docker bug doesn't allow you to measure GPU process usage in containers.
✅
Every energy grid has different emissions. Luckily found open source project (electricitymap.org) to build on top of. Missing regions filled manually from reading many reports.
✅
Energy and Policy Considerations for Deep Learning in NLP
(June 5, 2019)
📜🔥
Try to provide efficient RL architectures with lessons learned.
Can't seem to make things more energy efficient with architectures.
✅
Green AI
(July 22, 2019)
📜🔥
More lessons learned, more revisions, better tooling. Almost there with large draft.
Too much content to submit to AIES this year, which journal to submit to?
(October 2019)
❌
Quantifying the Carbon Emissions of Machine Learning
(October 21, 2019)
📜🔥
Improve tooling, make sure to add in everything we learned. Revise, revise, revise. Submit to ArXiv.
(Jan 31, 2020)
🎉


This crazy amount of compute must be having some carbon impact, how bad is it?
(May, 2018)
Start writing draft of paper
(October, 2018)
Estimates are inaccurate, don't want particular authors to feel attacked, decide not to submit fully written draft 2 days before the deadline.
(Nov 3, 2018)
✅
Can we estimate conference-wide carbon emissions and energy consumption?
Not enough information to make accurate estimates.
✅
Implementing interface to energy counters. Ran experiments.
Experiments wrong because RAPL interface on Slurm counts all experiments on machine.
✅
Assign based on percentage of CPU usage by process.
Docker bug doesn't allow you to measure GPU process usage in containers.
✅
Every energy grid has different emissions. Luckily found open source project (electricitymap.org) to build on top of. Missing regions filled manually from reading many reports.
✅
Energy and Policy Considerations for Deep Learning in NLP
(June 5, 2019)
📜🔥
Try to provide efficient RL architectures with lessons learned.
Can't seem to make things more energy efficient with architectures.
✅
Green AI
(July 22, 2019)
📜🔥
More lessons learned, more revisions, better tooling. Almost there with large draft.
Too much content to submit to AIES this year, which journal to submit to?
(October 2019)
❌
Quantifying the Carbon Emissions of Machine Learning
(October 21, 2019)
📜🔥
Improve tooling, make sure to add in everything we learned. Revise, revise, revise. Submit to ArXiv.
(Jan 31, 2020)
🎉
Stuck in ArXiv purgatory because put CS.LG not CS.CY as topic.
❌


This crazy amount of compute must be having some carbon impact, how bad is it?
(May, 2018)
Start writing draft of paper
(October, 2018)
Estimates are inaccurate, don't want particular authors to feel attacked, decide not to submit fully written draft 2 days before the deadline.
(Nov 3, 2018)
✅
Can we estimate conference-wide carbon emissions and energy consumption?
Not enough information to make accurate estimates.
✅
Implementing interface to energy counters. Ran experiments.
Experiments wrong because RAPL interface on Slurm counts all experiments on machine.
✅
Assign based on percentage of CPU usage by process.
Docker bug doesn't allow you to measure GPU process usage in containers.
✅
Every energy grid has different emissions. Luckily found open source project (electricitymap.org) to build on top of. Missing regions filled manually from reading many reports.
✅
Energy and Policy Considerations for Deep Learning in NLP
(June 5, 2019)
Try to provide efficient RL architectures with lessons learned.
Can't seem to make things more energy efficient with architectures.
✅
Green AI
(July 22, 2019)
More lessons learned, more revisions, better tooling. Almost there with large draft.
Too much content to submit to AIES this year, which journal to submit to?
(October 2019)
❌
Quantifying the Carbon Emissions of Machine Learning
(October 21, 2019)
Improve tooling, make sure to add in everything we learned. Revise, revise, revise. Submit to ArXiv.
(Jan 31, 2020)
Stuck in ArXiv purgatory because put CS.LG not CS.CY as topic.
❌
Finally online! Submitted to journal shortly after.
(Feb 14, 2020)


This crazy amount of compute must be having some carbon impact, how bad is it?
(May, 2018)
Start writing draft of paper
(October, 2018)
Estimates are inaccurate, don't want particular authors to feel attacked, decide not to submit fully written draft 2 days before the deadline.
(Nov 3, 2018)
✅
Can we estimate conference-wide carbon emissions and energy consumption?
Not enough information to make accurate estimates.
✅
Implementing interface to energy counters. Ran experiments.
Experiments wrong because RAPL interface on Slurm counts all experiments on machine.
✅
Assign based on percentage of CPU usage by process.
Docker bug doesn't allow you to measure GPU process usage in containers.
✅
Every energy grid has different emissions. Luckily found open source project (electricitymap.org) to build on top of. Missing regions filled manually from reading many reports.
✅
Energy and Policy Considerations for Deep Learning in NLP
(June 5, 2019)
📜🔥
Try to provide efficient RL architectures with lessons learned.
Can't seem to make things more energy efficient with architectures.
✅
Green AI
(July 22, 2019)
📜🔥
More lessons learned, more revisions, better tooling. Almost there with large draft.
Too much content to submit to AIES this year, which journal to submit to?
(October 2019)
✅
Quantifying the Carbon Emissions of Machine Learning
(October 21, 2019)
📜🔥
Improve tooling, make sure to add in everything we learned. Revise, revise, revise. Submit to ArXiv.
(Jan 31, 2020)
🎉
Stuck in ArXiv purgatory because put CS.LG not CS.CY as topic.
✅
Finally online! Submitted to journal shortly after.
(Feb 14, 2020)
🎉

Lessons Learned

#1 Diving deeper leads to insights, but you might get scooped in the meantime.

Lessons Learned

#2 Getting scooped is stressful, but those deeper insights might be valuable by themselves. Still worth publishing (or at least putting online).

Lessons Learned

#3 Journal-length papers give you more room to share all the nitty gritty lessons learned, but seems like not as many people actually read through the whole thing.

Lessons Learned

#4 We should be more pro-active about sharing lessons learned, hence this workshop!

Lessons Learned

#5 We need to share more details as a community, either through appendices or journal-length follow-ups. Simply not enough information for post-hoc meta-analyses. In our case, we couldn't estimate emissions which is why we ended up down this year and a half long journey.

Thanks for listening!

If you want to chat with me about this or anything else,
feel free to reach out to me.