Some Damaging Delusions of DL Practice (and How to Avoid Them)

Arun Kumar

UC San Diego
JACOBS SCHOOL OF ENGINEERING
Computer Science and Engineering

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

ACM KDD Deep Learning Day

August 18, 2021

My Research

New abstractions, algorithms, and software systems to "***democratize***" ML/AI-based data analytics from a data management/systems standpoint

My Research

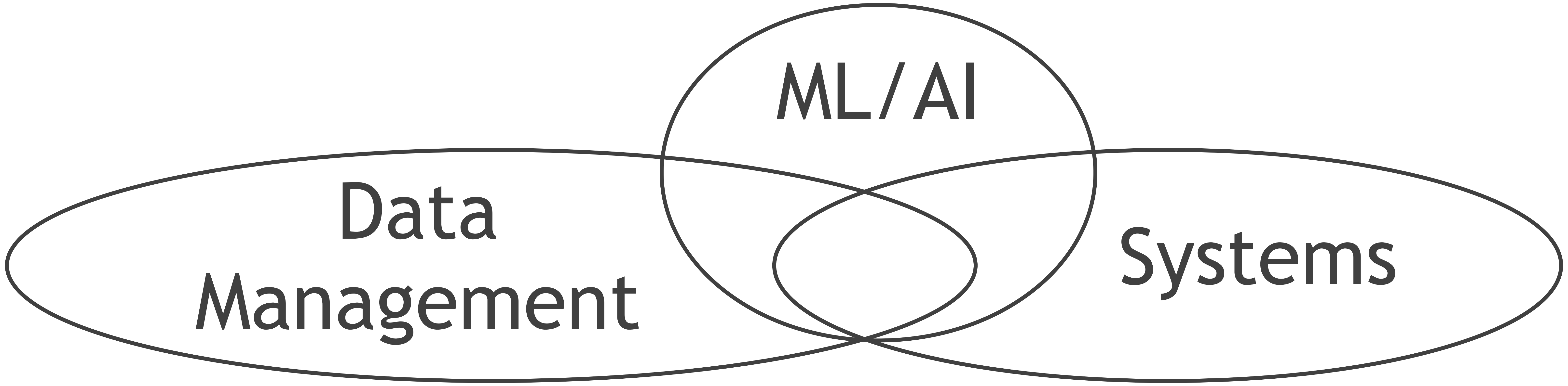
New abstractions, algorithms, and software systems to "***democratize***" ML/AI-based data analytics from a data management/systems standpoint

Democratization = System Efficiency (Reduce costs) + Human Efficiency (Improve productivity)

My Research

New abstractions, algorithms, and software systems to “***democratize***” ML/AI-based data analytics from a data management/systems standpoint

Democratization = System Efficiency (Reduce costs) + Human Efficiency (Improve productivity)



My Research

New abstractions, algorithms, and software systems to "***democratize***" ML/AI-based data analytics from a data management/systems standpoint

Democratization = **System Efficiency (Reduce costs)** + **Human Efficiency (Improve productivity)**

Practical and scalable data systems for ML/AI analytics

Inspired by *relational database systems* principles

Exploit insights from *learning theory* and *optimization theory*

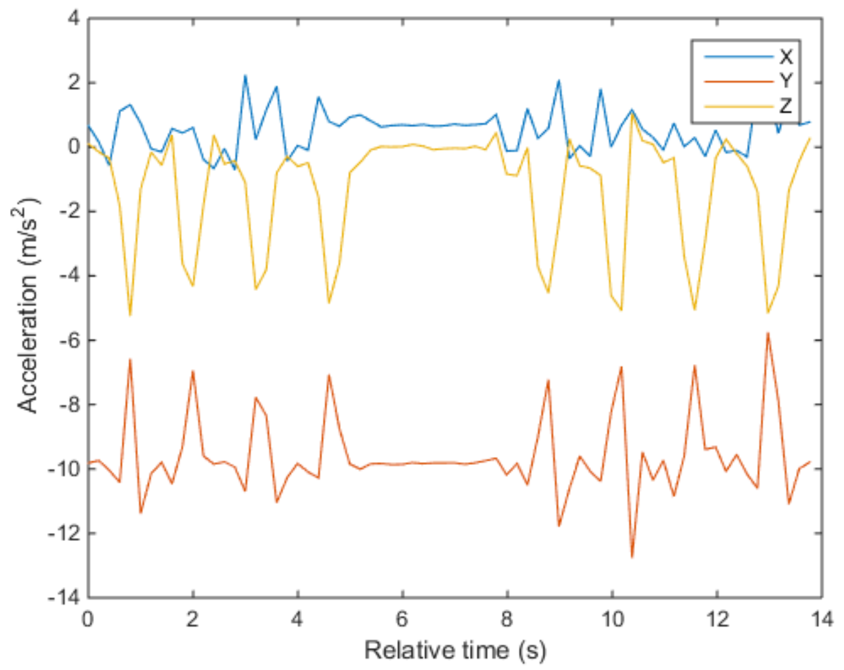
Outline

- Why am I here to speak?
- Modeling-related DL Delusions
- Systems-related DL Delusions

Large-Scale DL for Public Health

Example:

Predict sit vs not sit using ~1 TB of accelerometer data

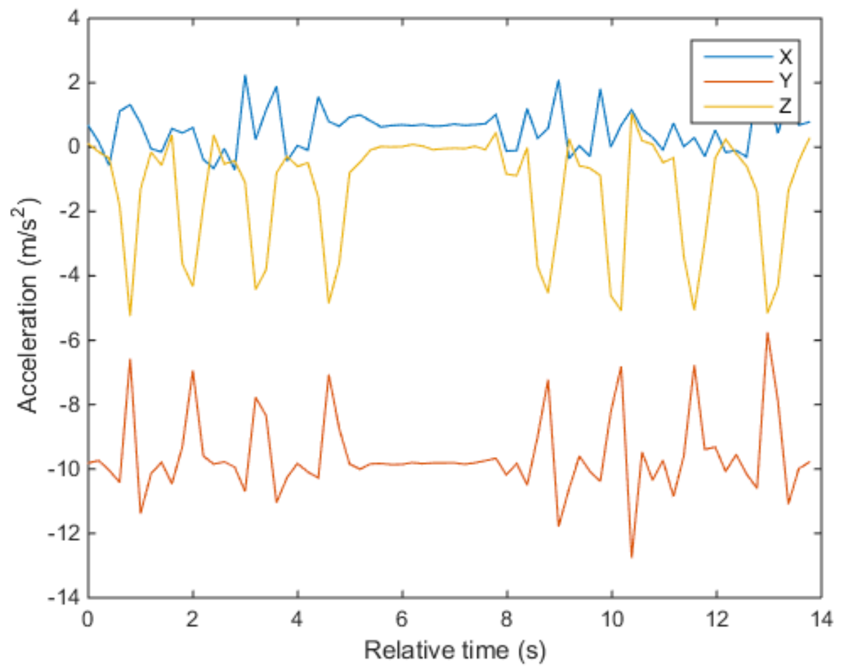


UC San Diego
SCHOOL OF MEDICINE

THE HERBERT WERTHEIM SCHOOL OF PUBLIC HEALTH
AND HUMAN LONGEVITY SCIENCE

Large-Scale DL for Public Health

Example:
Predict sit vs not sit using
~1 TB of accelerometer data



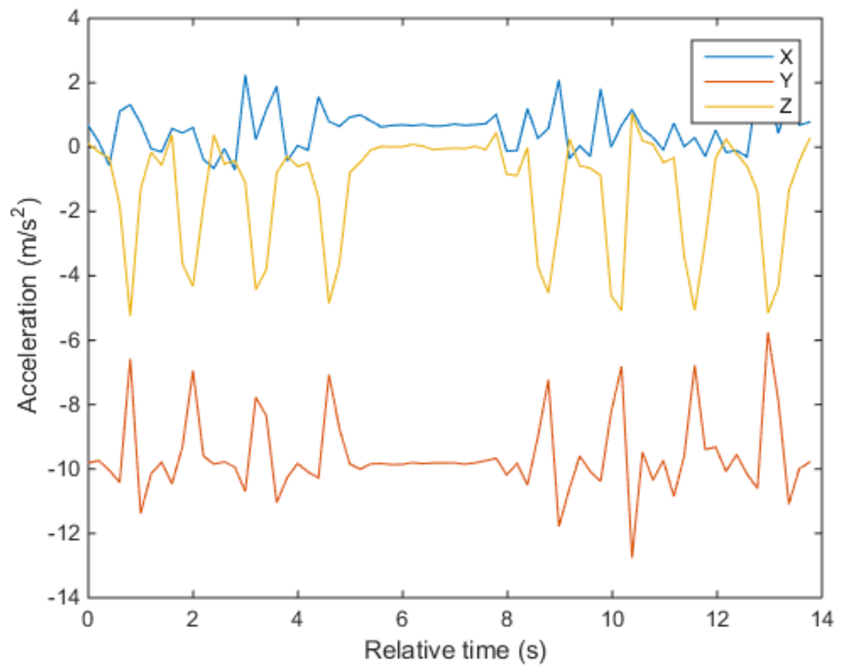
UC San Diego
SCHOOL OF MEDICINE

THE HERBERT WERTHEIM SCHOOL OF PUBLIC HEALTH
AND HUMAN LONGEVITY SCIENCE

Their prior hand-tuned physics-based features + RandomForest: 76%
Our best 1-D CNN-LSTM: 92%!

Large-Scale DL for Public Health

Example:
Predict sit vs not sit using
~1 TB of accelerometer data



UC San Diego
SCHOOL OF MEDICINE

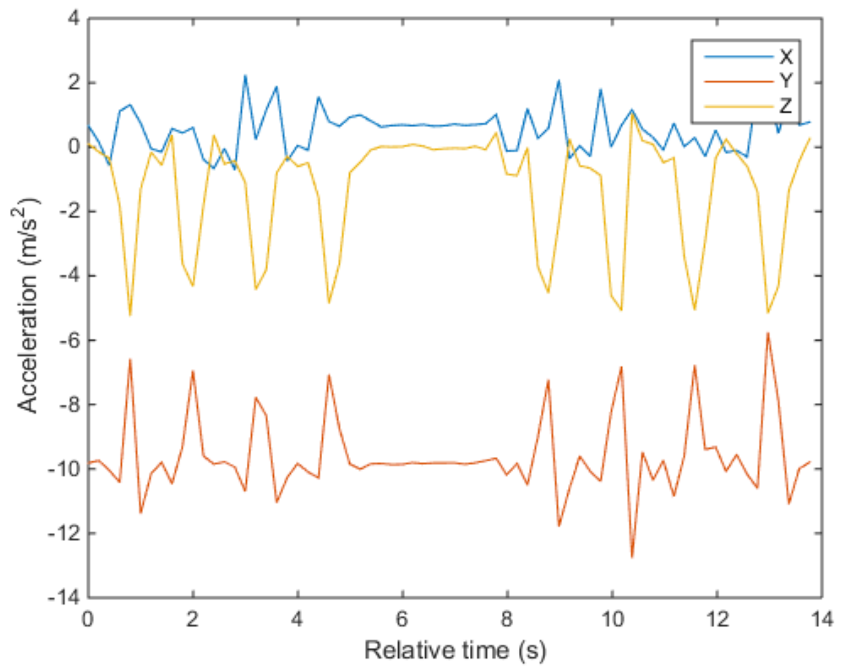
THE HERBERT WERTHEIM SCHOOL OF PUBLIC HEALTH
AND HUMAN LONGEVITY SCIENCE

Their prior hand-tuned physics-based features + RandomForest: 76%
Our best 1-D CNN-LSTM: 92%!

***Q:** How did we achieve such a high lift?*

Large-Scale DL for Public Health

Example:
Predict sit vs not sit using
~1 TB of accelerometer data



UC San Diego
SCHOOL OF MEDICINE

THE HERBERT WERTHEIM SCHOOL OF PUBLIC HEALTH
AND HUMAN LONGEVITY SCIENCE

Their prior hand-tuned physics-based features + RandomForest: 76%
Our best 1-D CNN-LSTM: 92%!

Q: How did we achieve such a high lift?

Secret Sauce:
Model selection exploration throughput
Existing DL systems' parallelism was a poor fit!

My friends, the reason I am here today.
Is to bust many DL delusions and to slay.
DL practices so abysmal.
DL systems so dismal.
They even turned my hair gray!

Outline

- Why am I here to speak?
- Modeling-related DL Delusions
- Systems-related DL Delusions

Background: B-V-N Tradeoff

ML (Test) Error = Bias + Variance + Bayes Noise

Complexity of feature space & Model complexity

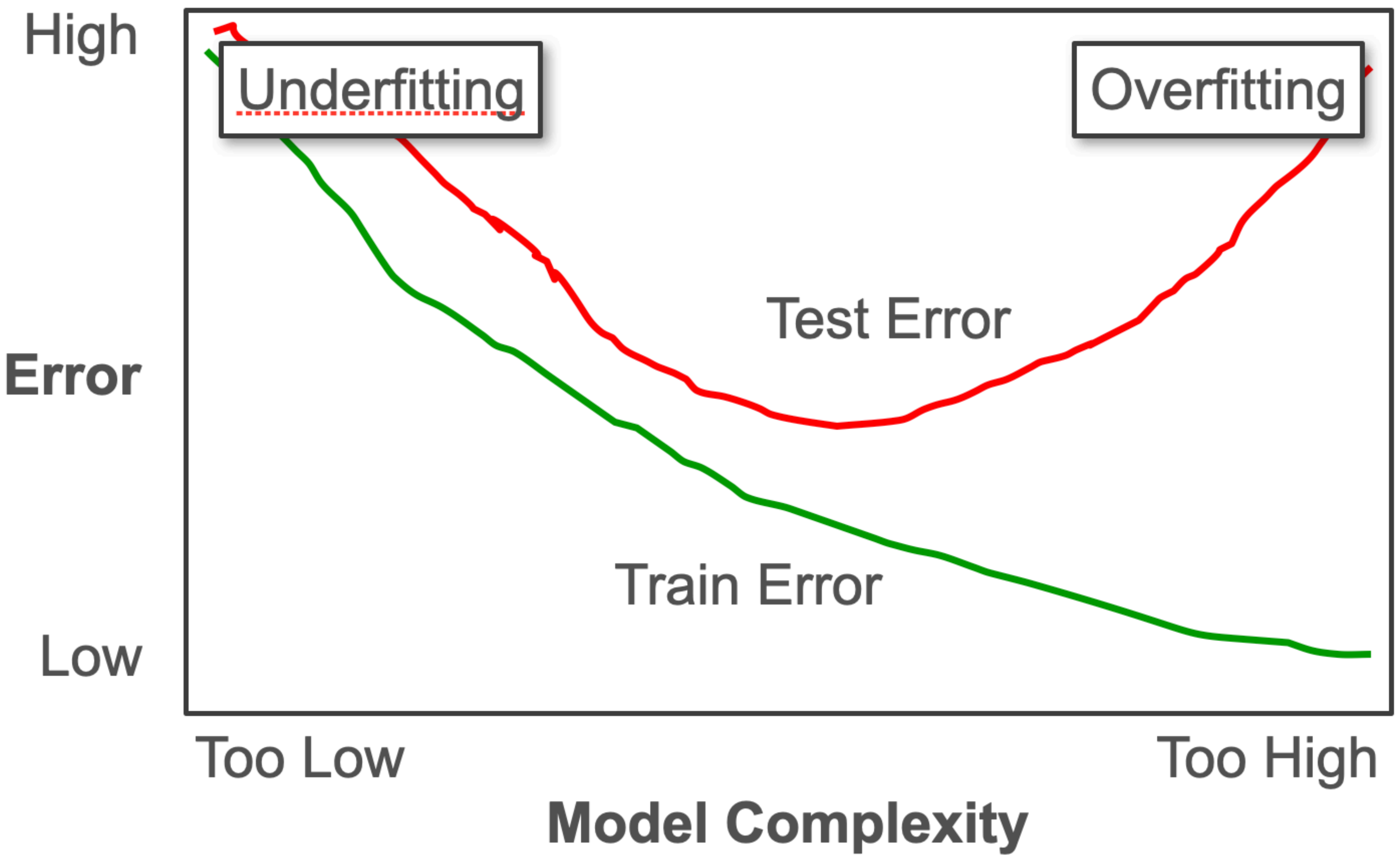
Discriminability of examples

Background: B-V-N Tradeoff

ML (Test) Error = Bias + Variance + Bayes Noise

Complexity of feature space & Model complexity

Discriminability of examples

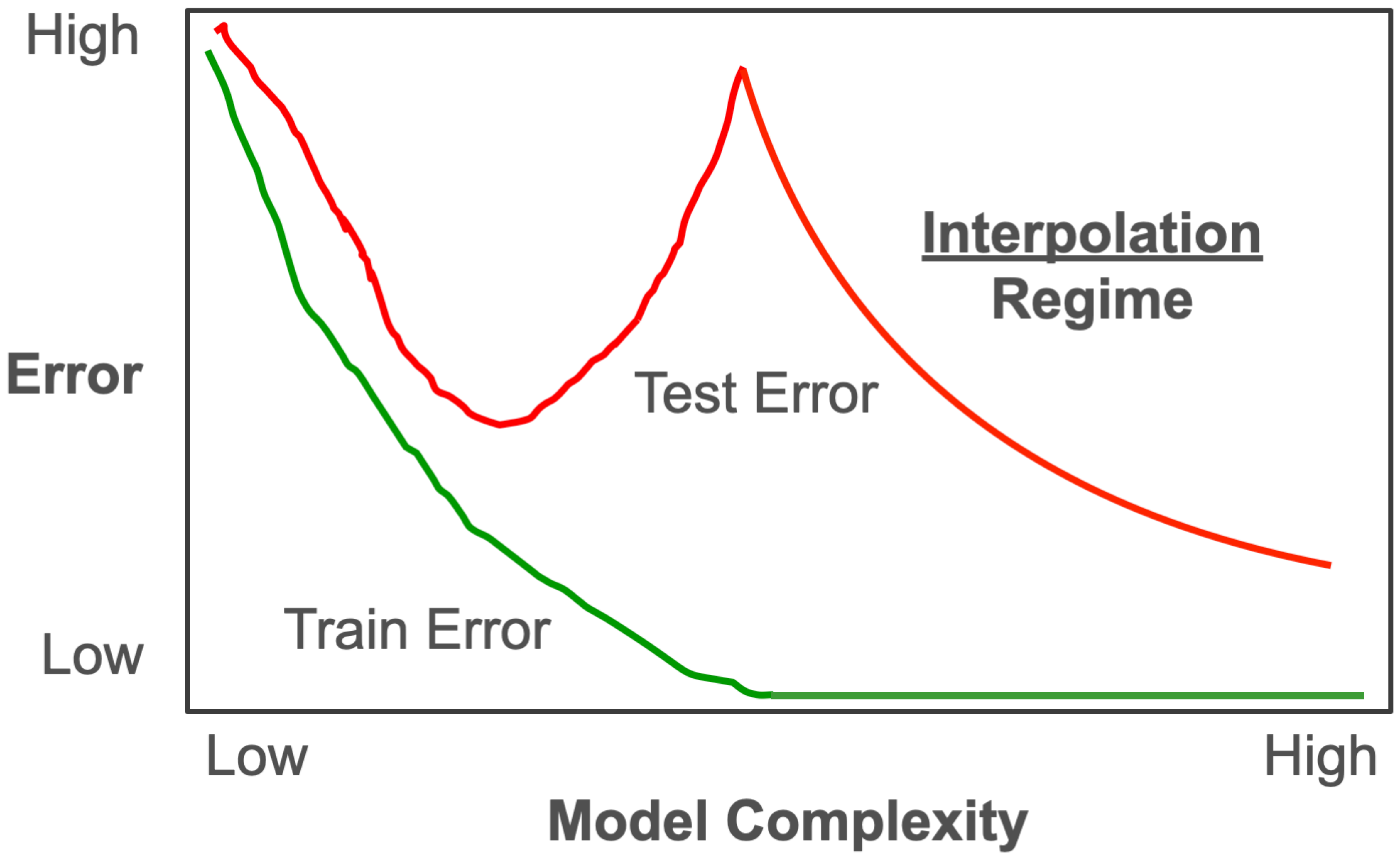


Background: B-V-N Tradeoff

ML (Test) Error = Bias + Variance + Bayes Noise

Complexity of feature space & Model complexity

Discriminability of examples

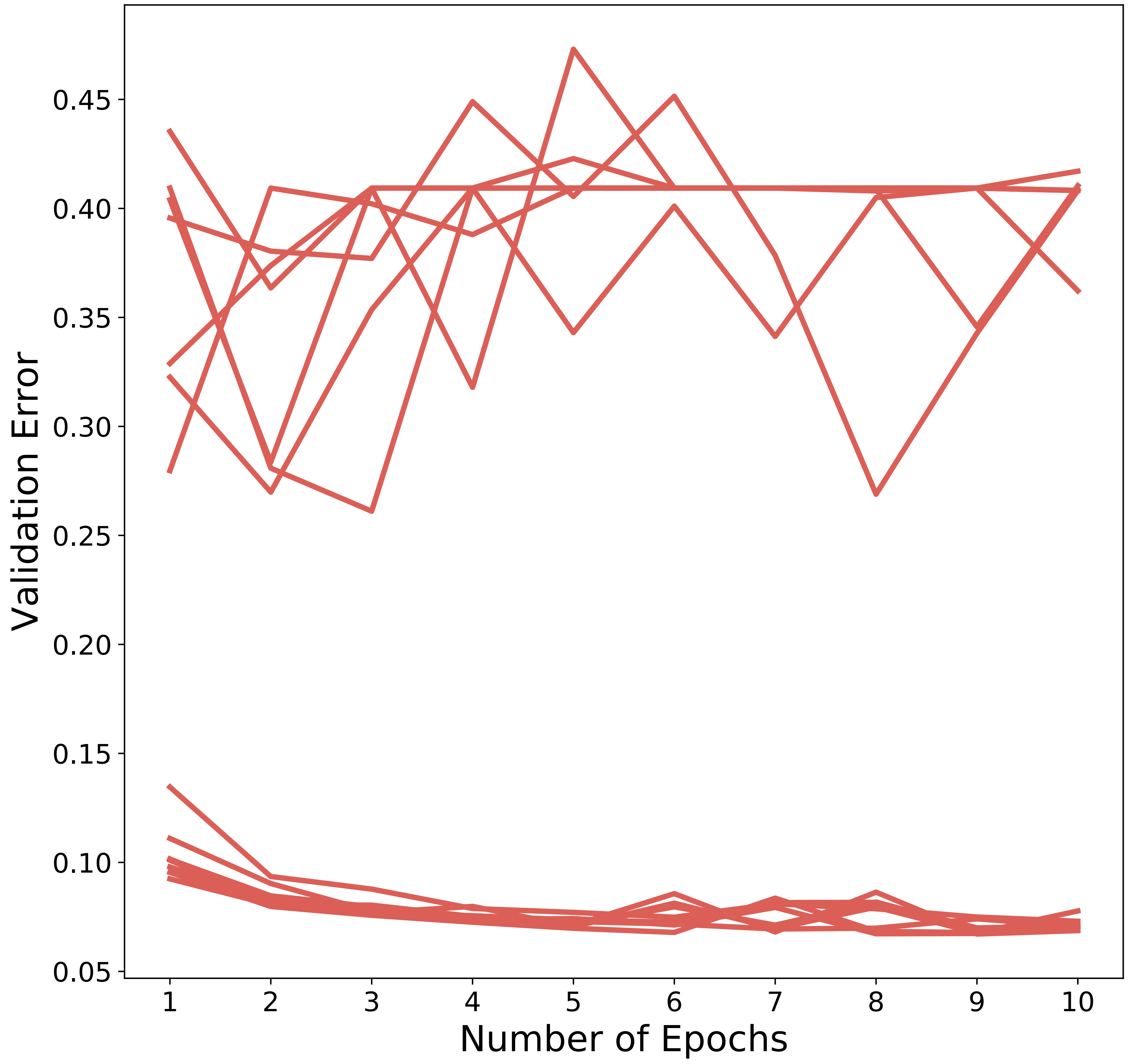


Model selection is **inevitable**!

Configuring data *representation*, neural *architecture*, and *hyper-parameters* is how one navigates Bias-Variance-Noise tradeoff space

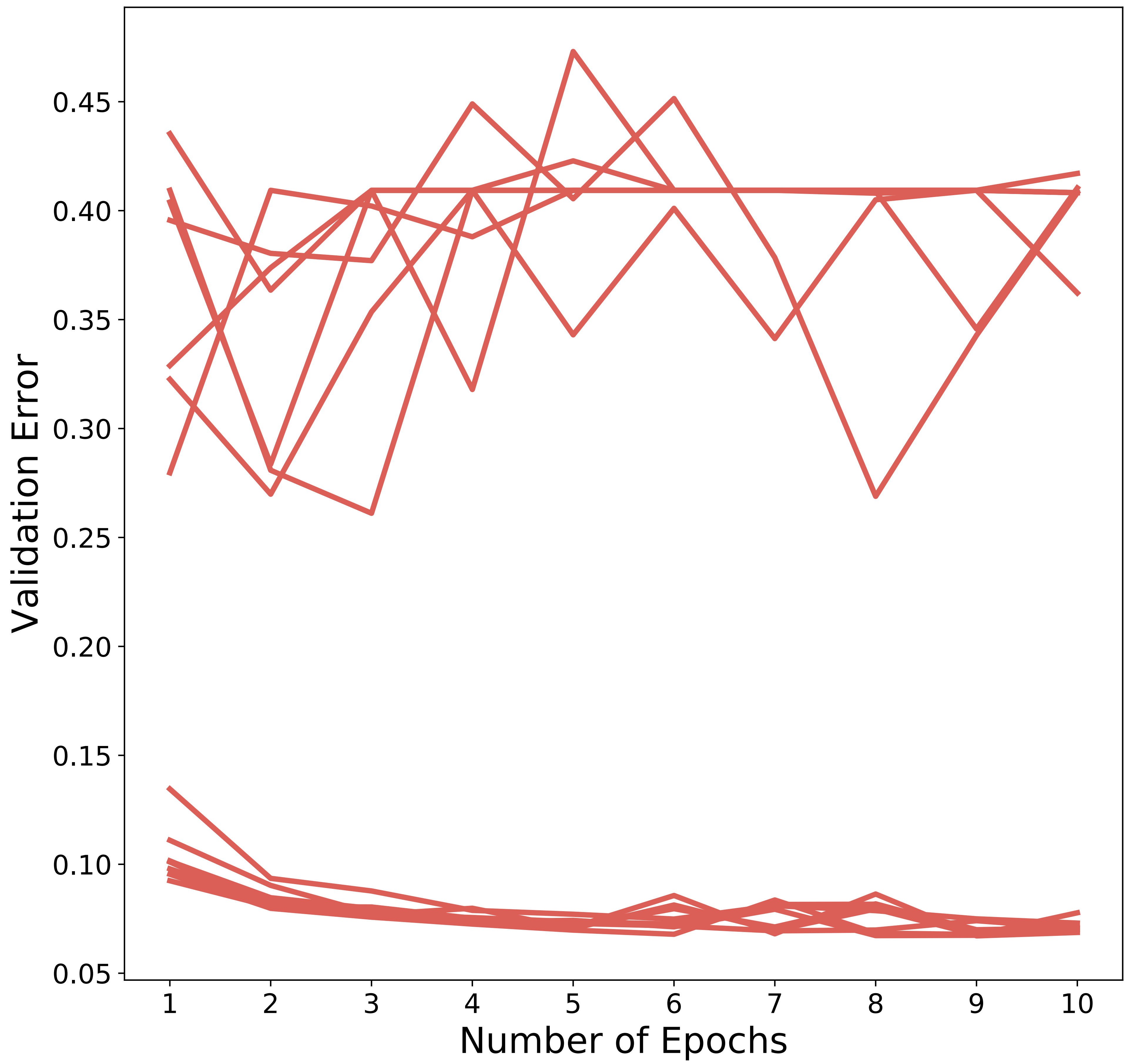
Model Selection on our Data

2 values each: time windows,
layers, learning rate, L2 regularizer

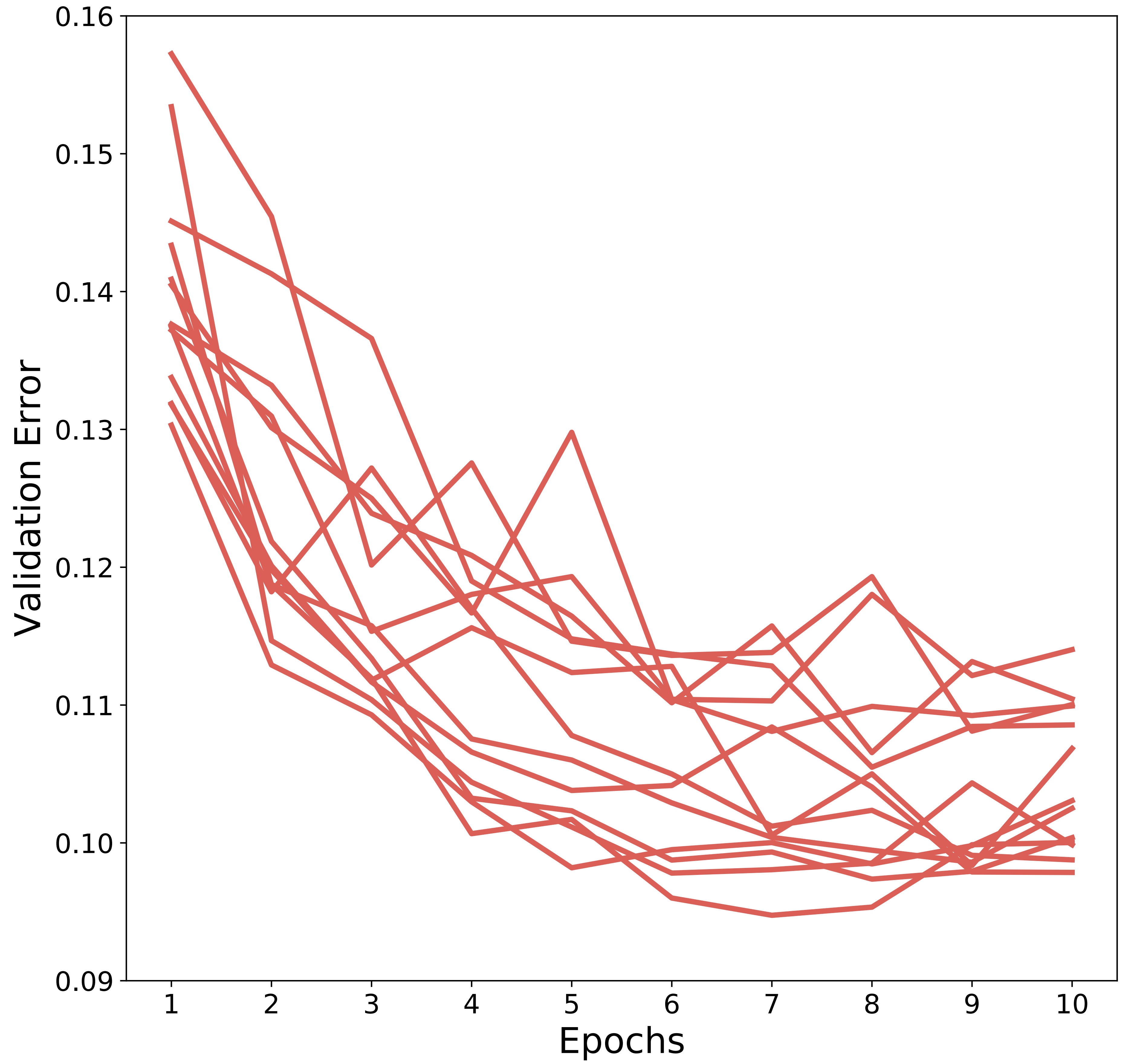


Model Selection on our Data

2 values each: time windows, # layers, learning rate, L2 regularizer



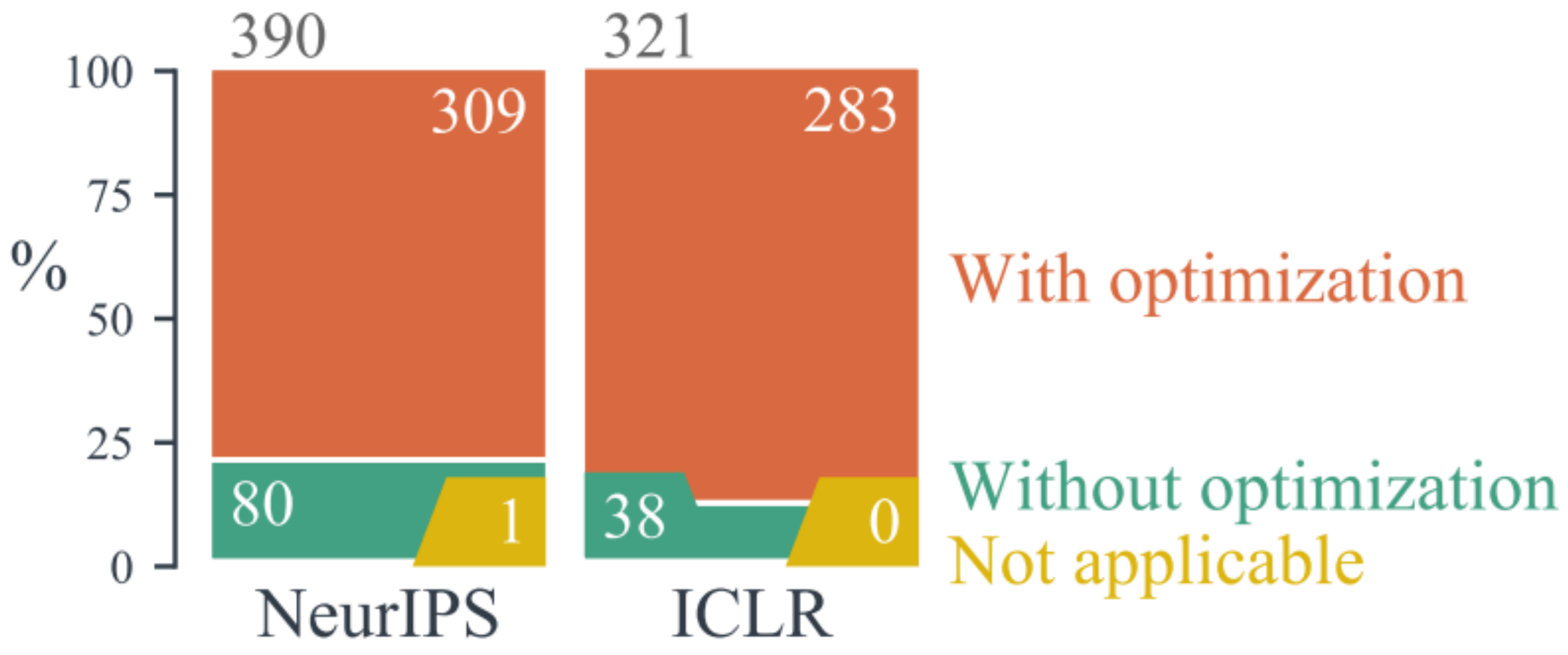
2x-4x network capacity; interpolation regime is hard to reach; much slower!



Abysmal State of Model Sel. IRL

Question 2)
Did you optimize your hyperparameters?

Results are for empirical papers only.

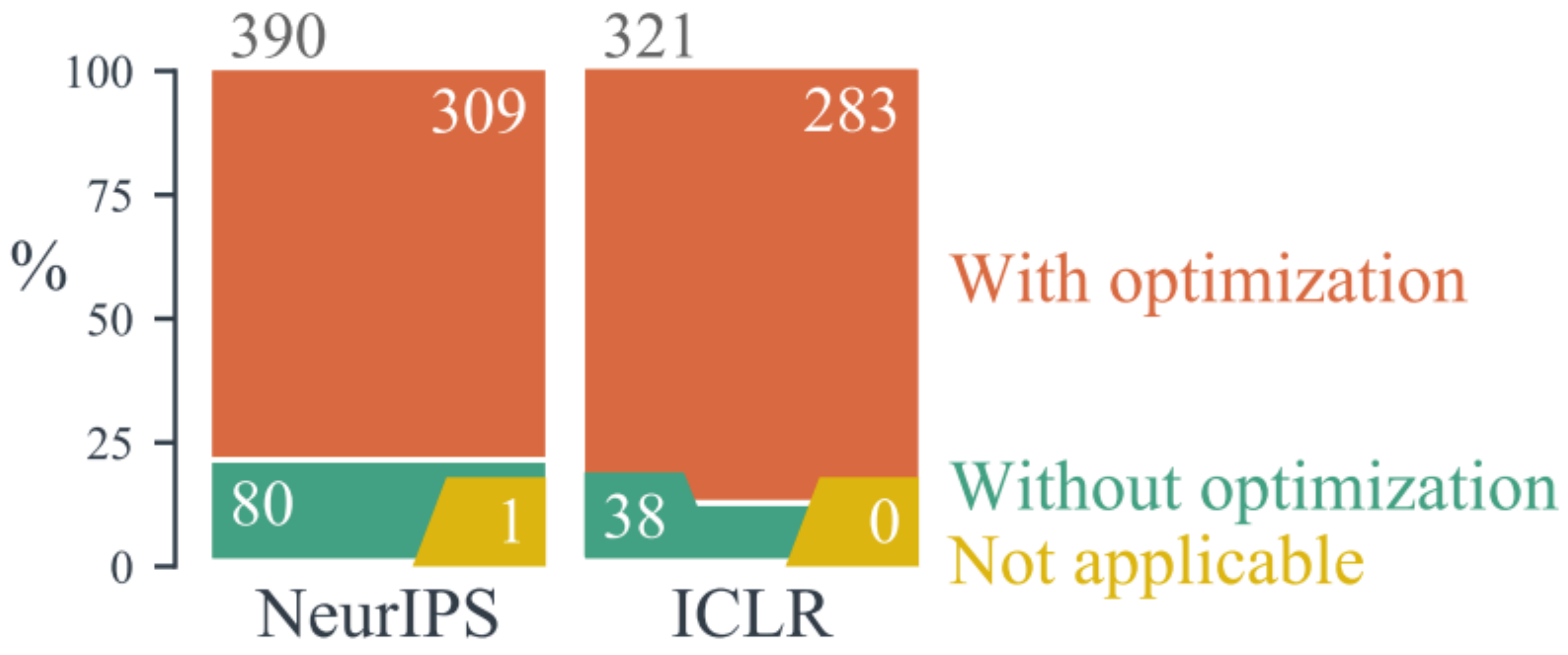


https://hal.archives-ouvertes.fr/hal-02447823/document

Abysmal State of Model Sel. IRL

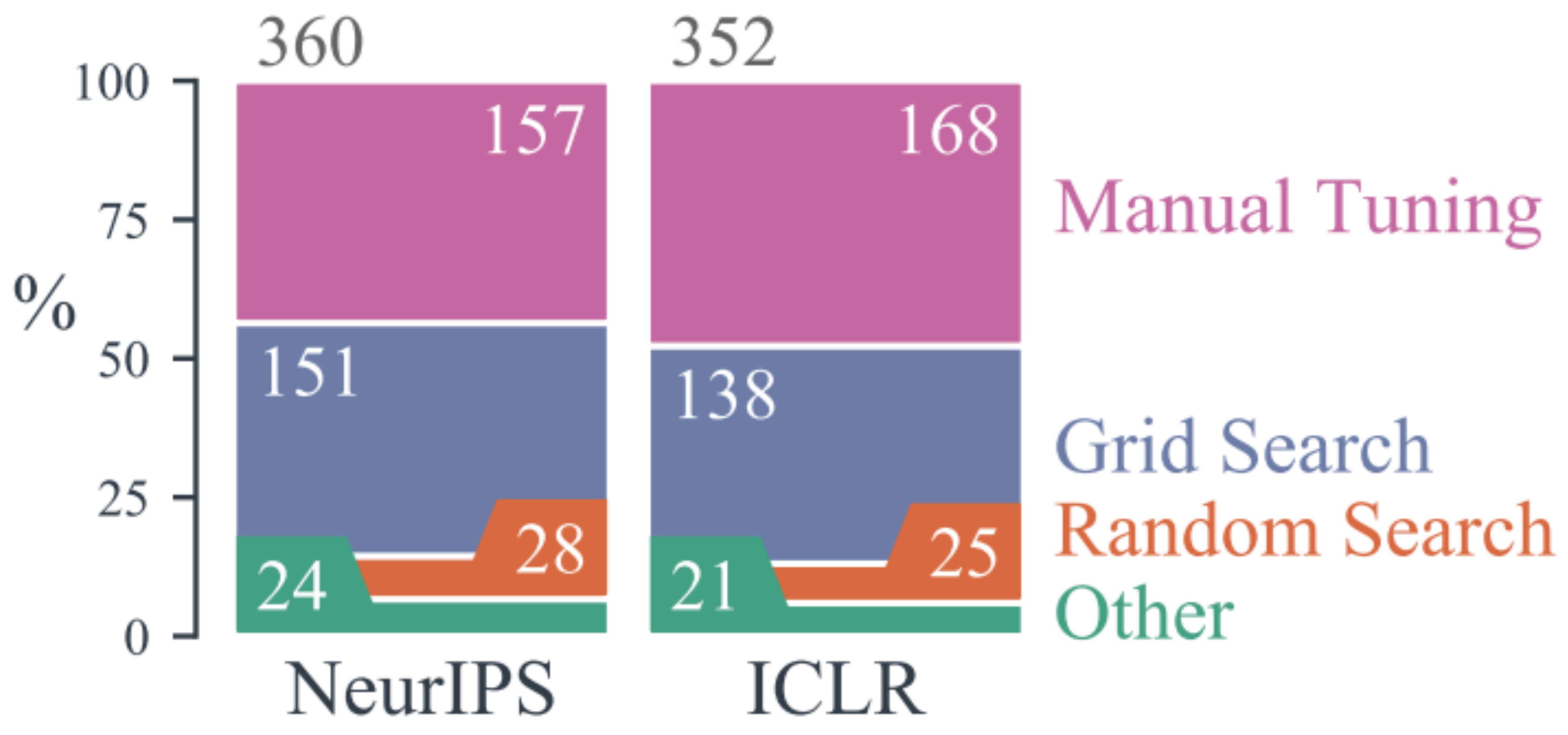
Question 2)
Did you optimize your hyperparameters?

Results are for empirical papers only.



Question 3)
If yes, how did you tune them?

Results are for empirical papers with optimization only. Papers may use more than one method, hence it sums to more than the number of empirical papers.

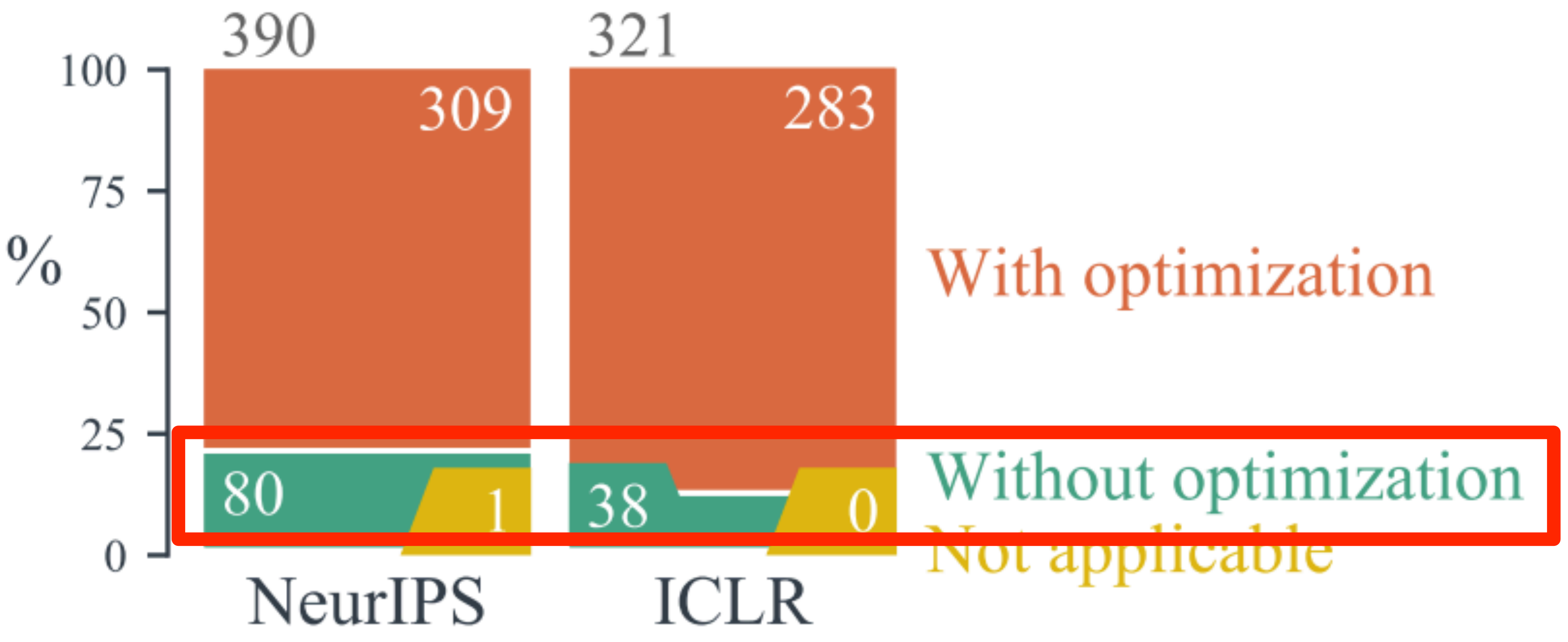


https://hal.archives-ouvertes.fr/hal-02447823/document

Abysmal State of Model Sel. IRL

Question 2)
Did you optimize your hyperparameters?

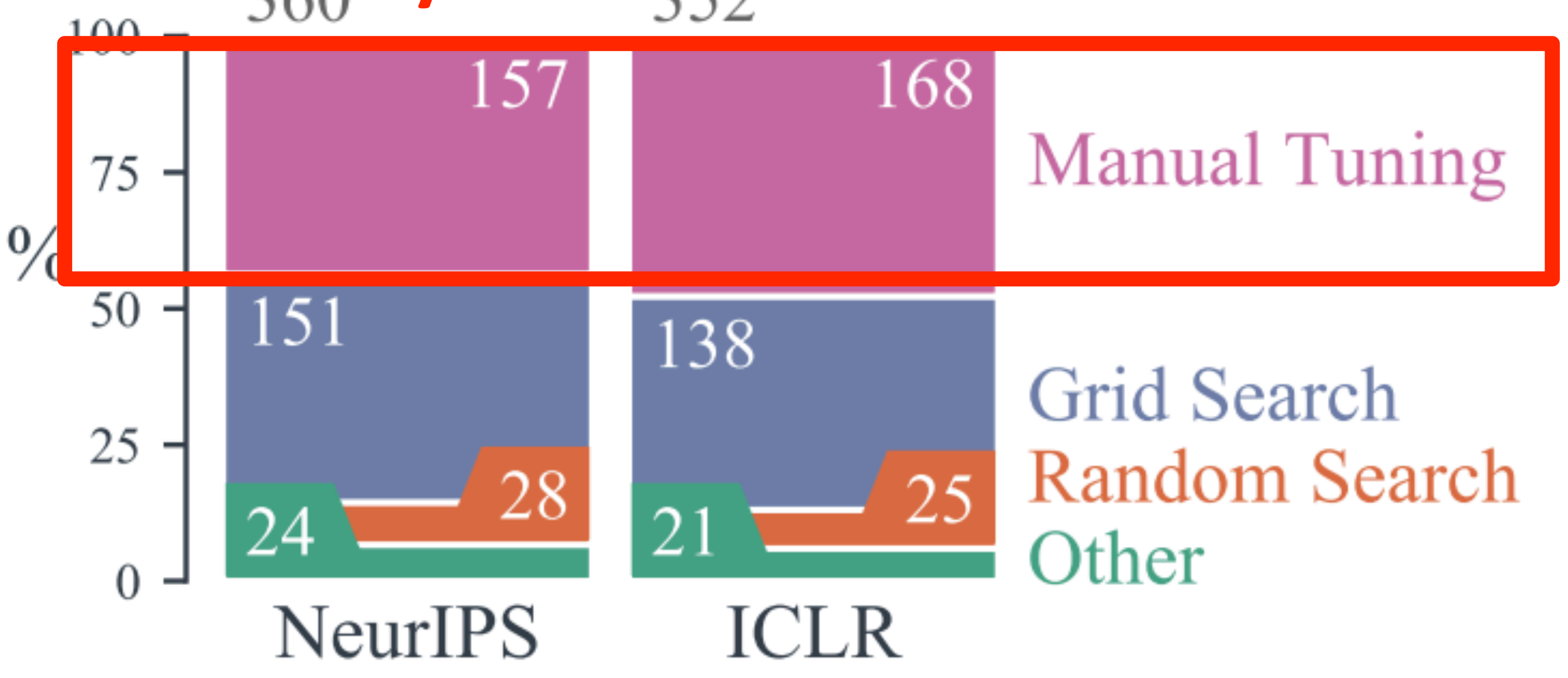
Results are for empirical papers only.



OMG!! Is ML/AI even a "science" anymore?!

Question 3)
If yes, how did you tune them?

Results are for empirical papers with optimization only. Papers may use more than one method, hence it sums to more than the number of empirical papers.



https://hal.archives-ouvertes.fr/hal-02447823/document

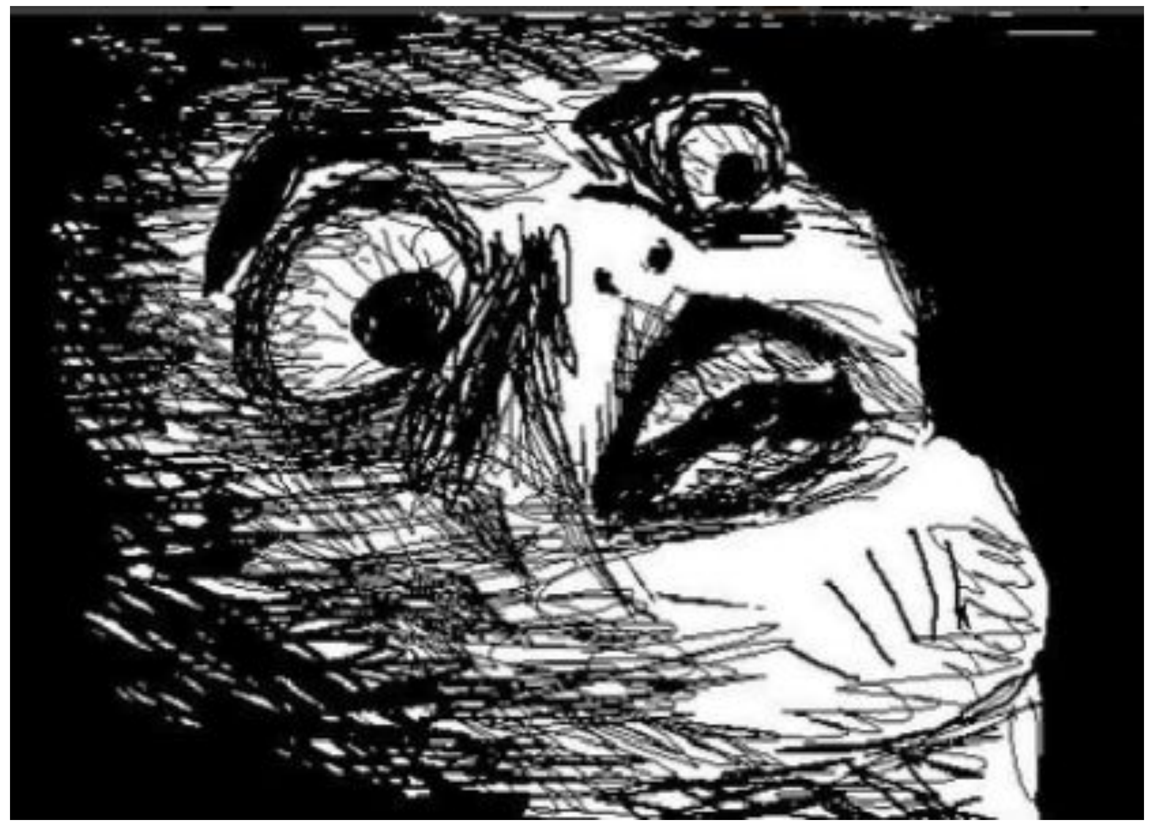
But it makes for… a lot of fun! :)

ML/AI types

"Yay, my fancy new model beats the baselines by a huge margin!"

"Properly tune all hyperparameters first"

Me



Baselines now match/beat new model!

https://datasystemsfun.tumblr.com/

Poor Model Sel. = Squander Labeled Data!

Poor Model Sel. = Squander Labeled Data!

Poor Model Sel. = Squander Labeled Data!

Poor Model Sel. = Squander Labeled Data!

Model selection is oft ignored by DL types.
In the worlds of ML, systems, DB, all stripes.
Are they deluded or just lazy?
Or just marketing like crazy?
Boy, for sure they are living stereotypes!

How to Avoid Modeling Delusion # 1:
Perform rigorous and repeatable model selection to optimize task-specific B-V-N tradeoffs

Is Automated ML the Savior?!

AutoML heuristics are indeed useful.

Is Automated ML the Savior?!

AutoML heuristics are indeed useful.

But they are very easy to *abuse*.

Is Automated ML the Savior?!

AutoML heuristics are indeed useful.

YEAH! AutoML tuned EVERYTHING for us!

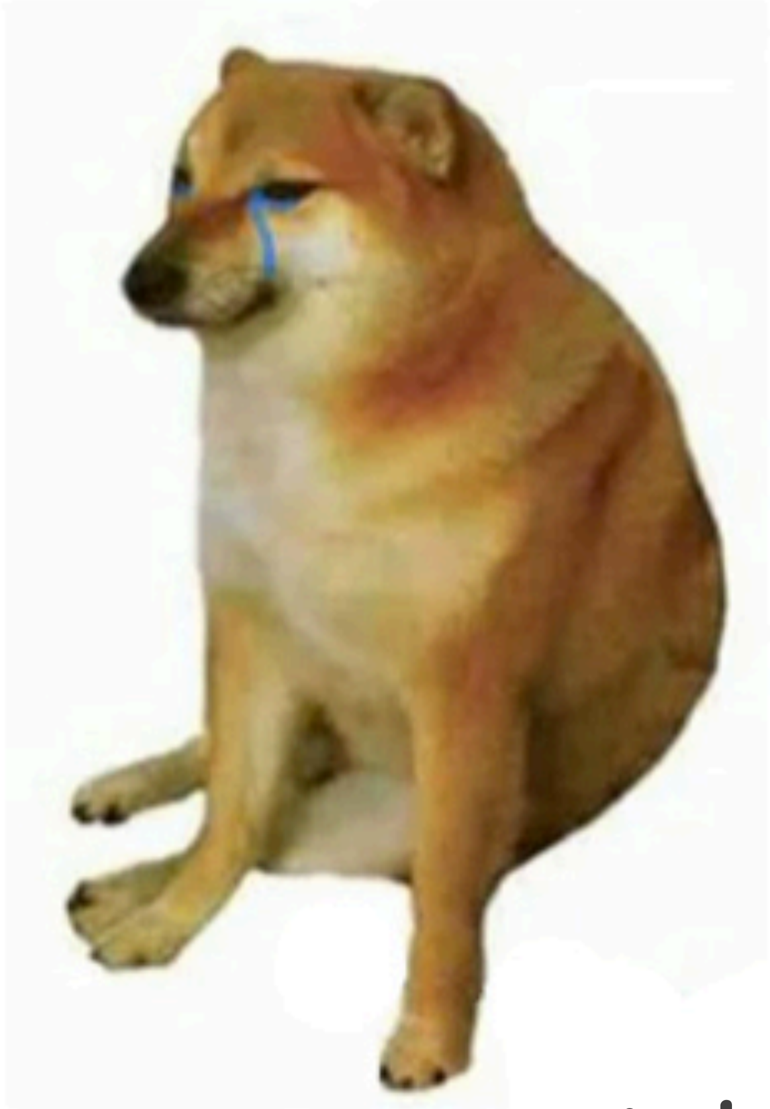
But they are very easy to *abuse*.

Is Automated ML the Savior?!

AutoML heuristics are indeed useful.

YEAH! AutoML tuned
EVERYTHING for us!

But they are very easy to *abuse*.



But we burned $1mil
for just 1 model!

How to Avoid Modeling Delusion # 2:
Hybrid human-in-the-loop + AutoML specification to rein in resource bloat

Is Transfer Learning the Savior?!

Is Transfer Learning the Savior?!

Is Transfer Learning the Savior?!

Model hubs, HuggingFace, etc. indeed help democratize SOTA DL

But the world is far, far bigger than just a few NLP or image tasks!

Is Transfer Learning the Savior?!

Model hubs, HuggingFace, etc. indeed help democratize SOTA DL
But the world is far, far bigger than just a few NLP or image tasks!

Q: *What does Transfer Learning have to do with model selection?*

Is Transfer Learning the Savior?!

Model hubs, HuggingFace, etc. indeed help democratize SOTA DL
But the world is far, far bigger than just a few NLP or image tasks!

Q: *What does Transfer Learning have to do with model selection?*

Er, literally everything!
Pre-trained models are *seeds* for featurization, fine-tuning, etc.
Raises Bias, reduces Variance; in overall mix test error drops

Is Transfer Learning the Savior?!

Model hubs, HuggingFace, etc. indeed help democratize SOTA DL
But the world is far, far bigger than just a few NLP or image tasks!

Q: *What does Transfer Learning have to do with model selection?*

Er, literally everything!
Pre-trained models are *seeds* for featurization, fine-tuning, etc.
Raises Bias, reduces Variance; in overall mix test error drops
Multimodal models have bespoke task-specific B-V-N tradeoffs

How to Avoid Modeling Delusion # 3:
Treat transfer learning rigorously as another part of model selection

But training so many models is painful!

Outline

- Why am I here to speak?
- Modeling-related DL Delusions
- Systems-related DL Delusions

So many DL Systems are so poor at scaling.
One wonders why there is so much failing.
Boring wasteful execution.
Is not really a scaling solution.
Against DL Systems I will now start railing.

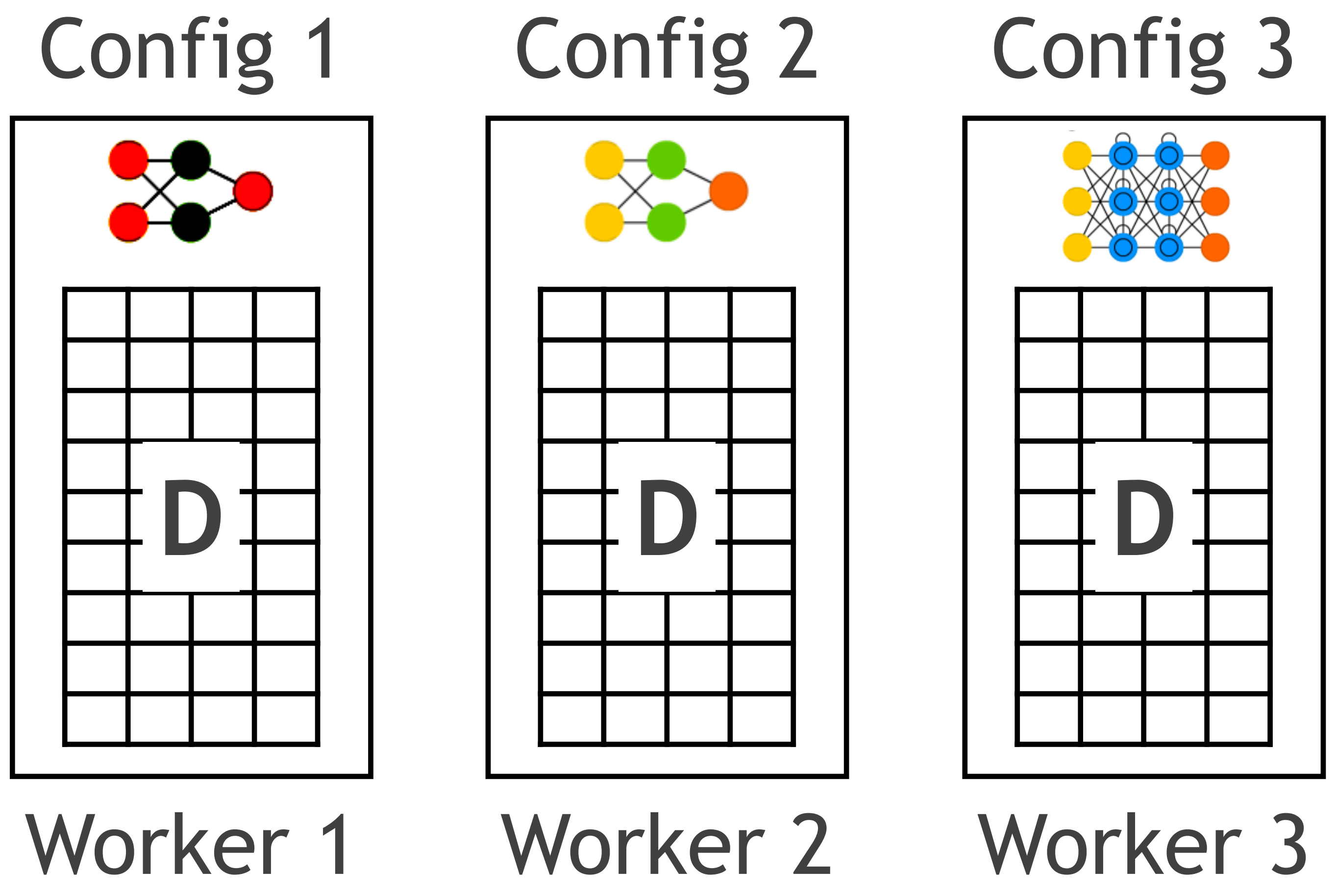
Boring Wasteful Execution at Scale

Q: *How do almost all DL Systems scale model selection today?*

Boring Wasteful Execution at Scale

Q: How do almost all DL Systems scale model selection today?

Task Parallelism:



1 worker per config

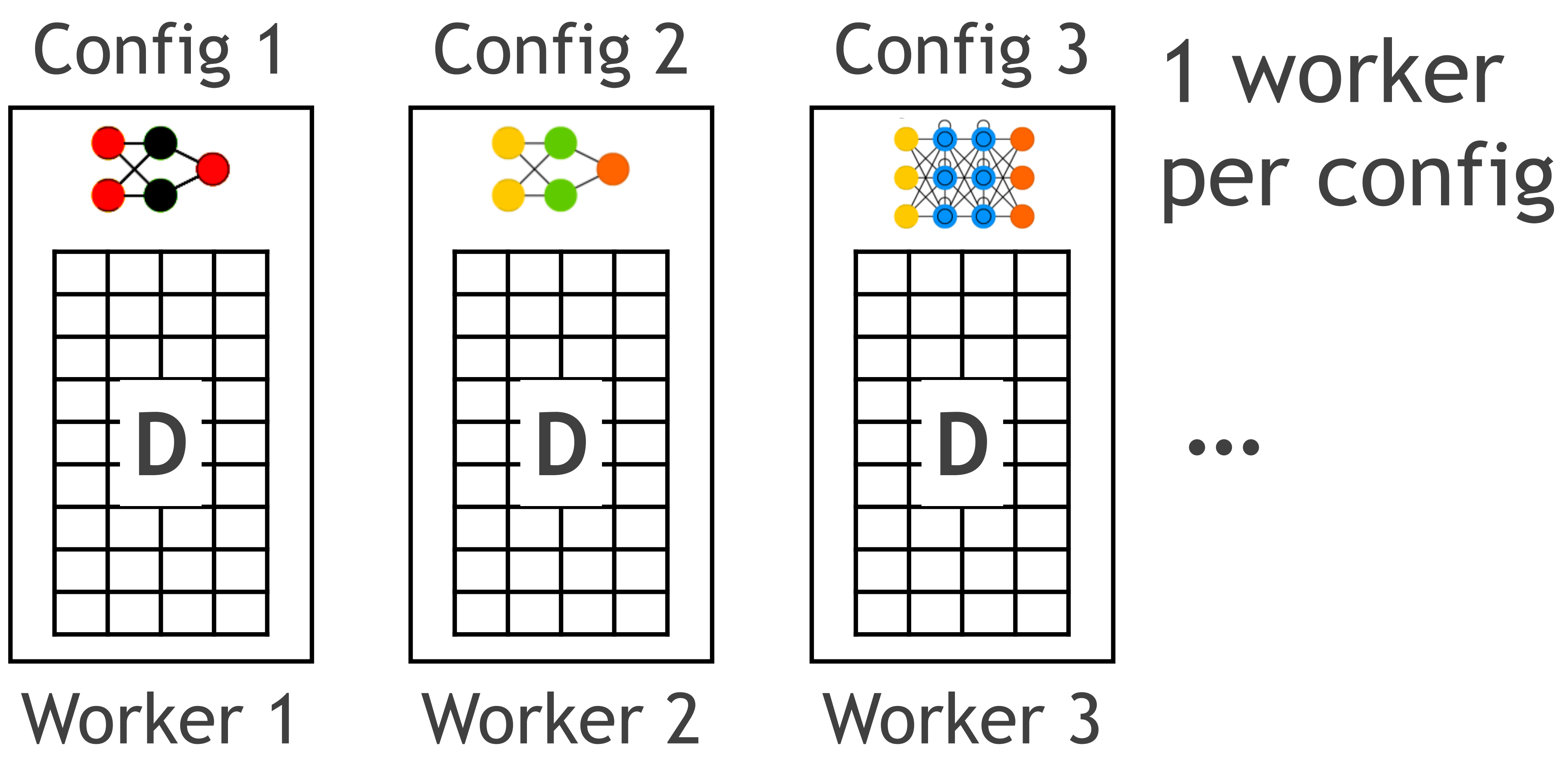
...

Ray, Google Vizier, Dask, Celery, ASHA, Determined

Boring Wasteful Execution at Scale

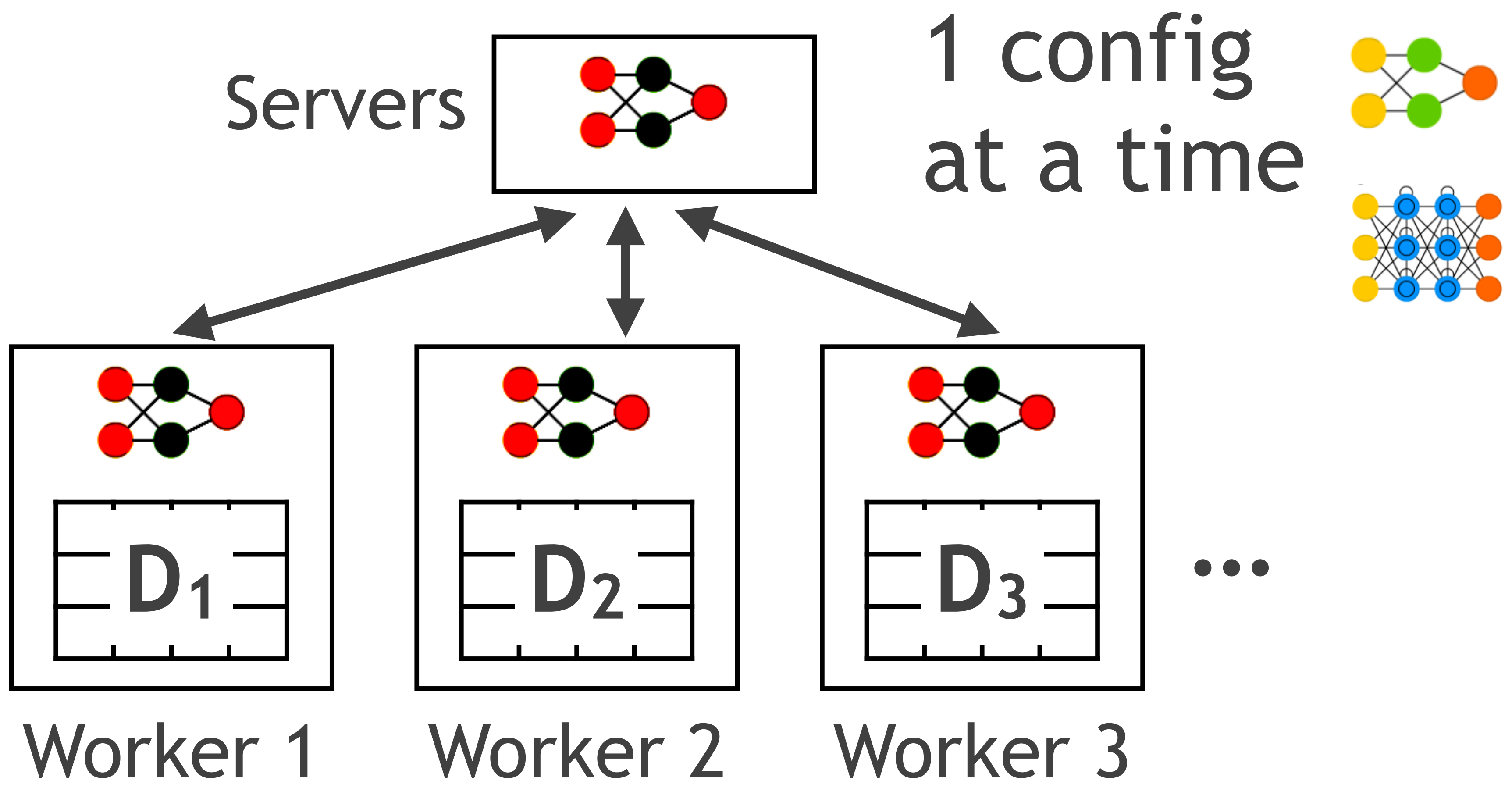
Q: How do almost all DL Systems scale model selection today?

Task Parallelism:



Ray, Google Vizier, Dask,
Celery, ASHA, Determined

Data Parallelism:

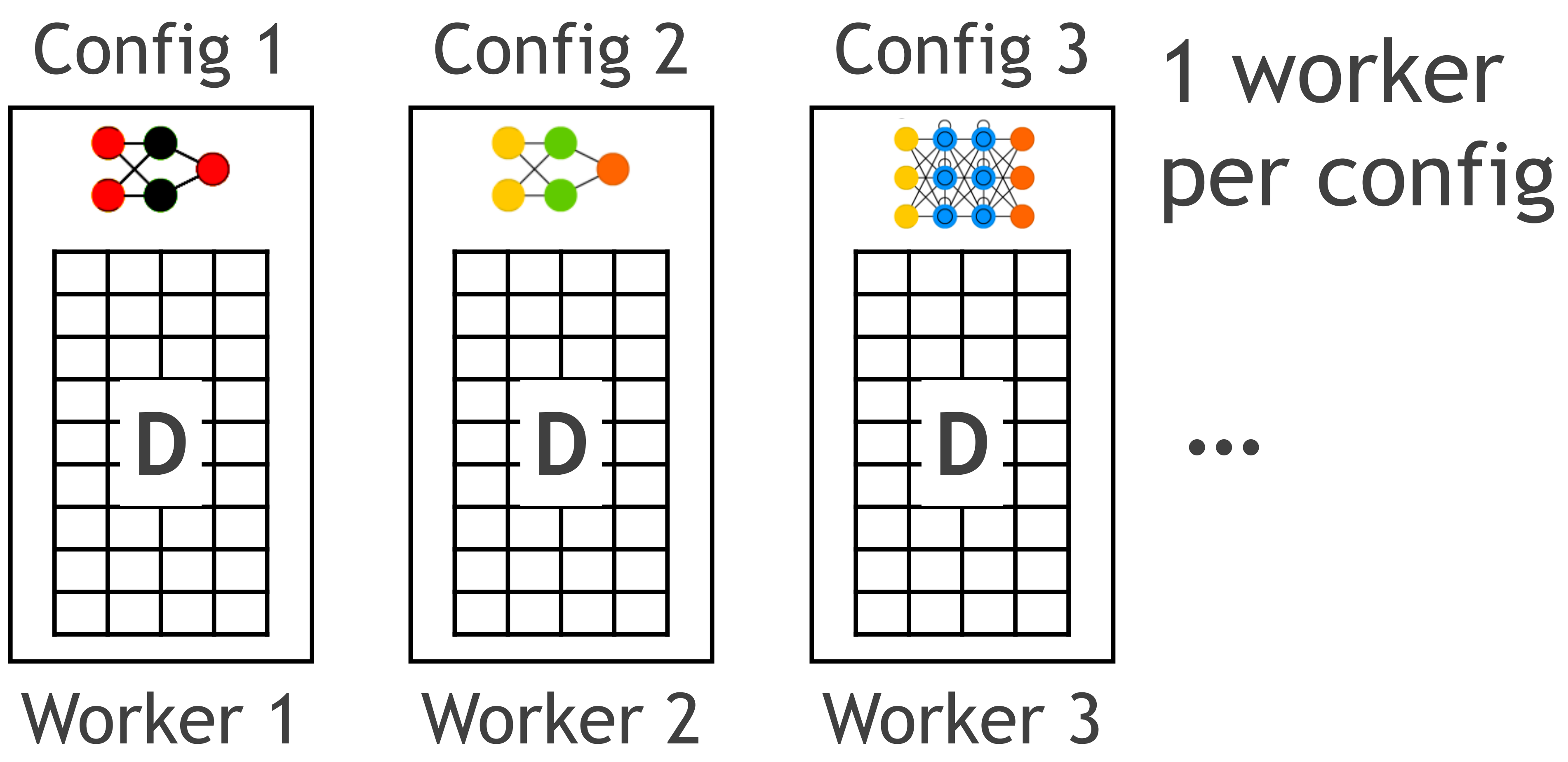


Horovod, Parameter Server,
Petuum AutoDist :)

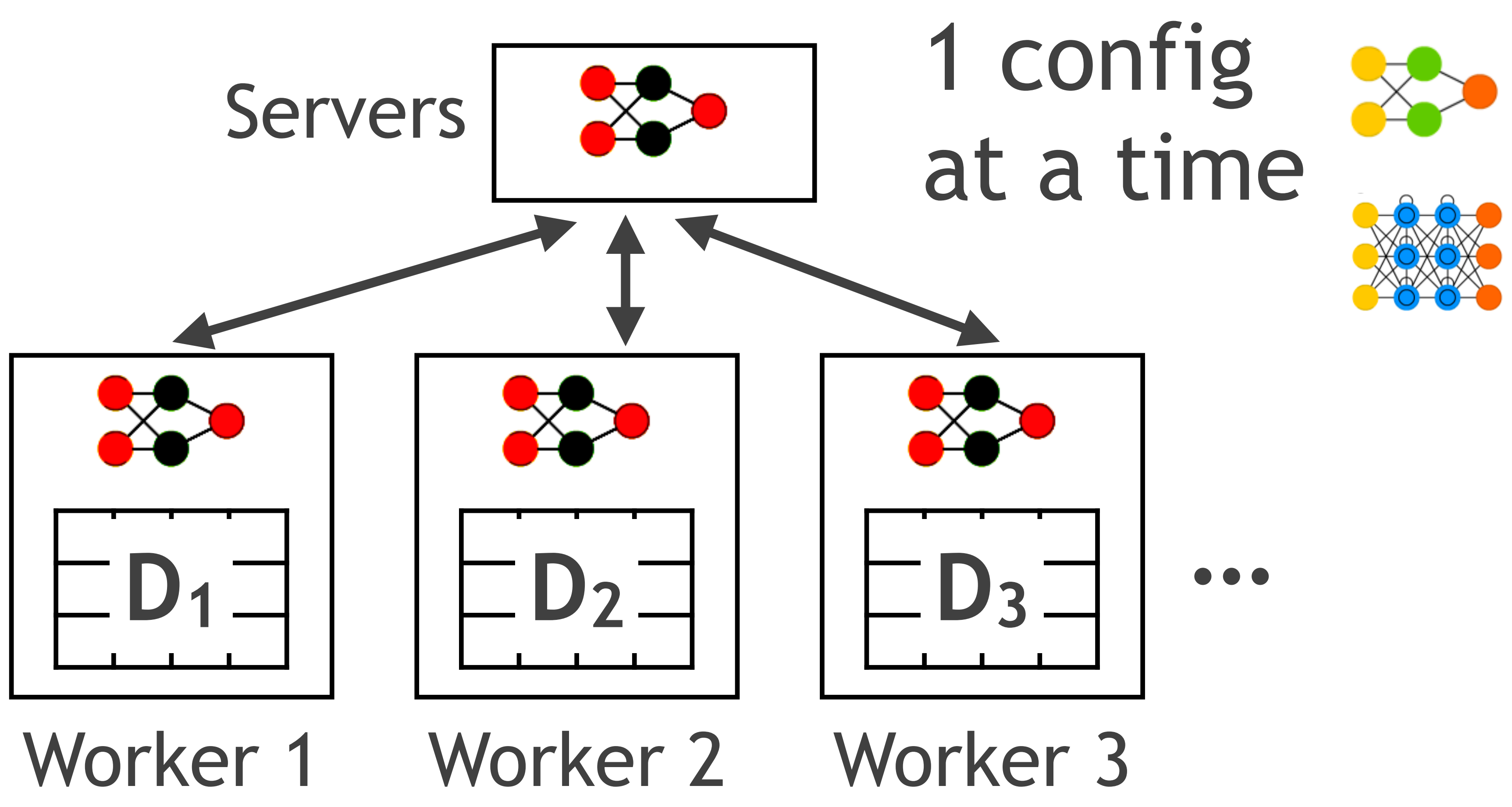
Boring Wasteful Execution at Scale

Q: How do almost all DL Systems scale model selection today?

Task Parallelism:



Data Parallelism:

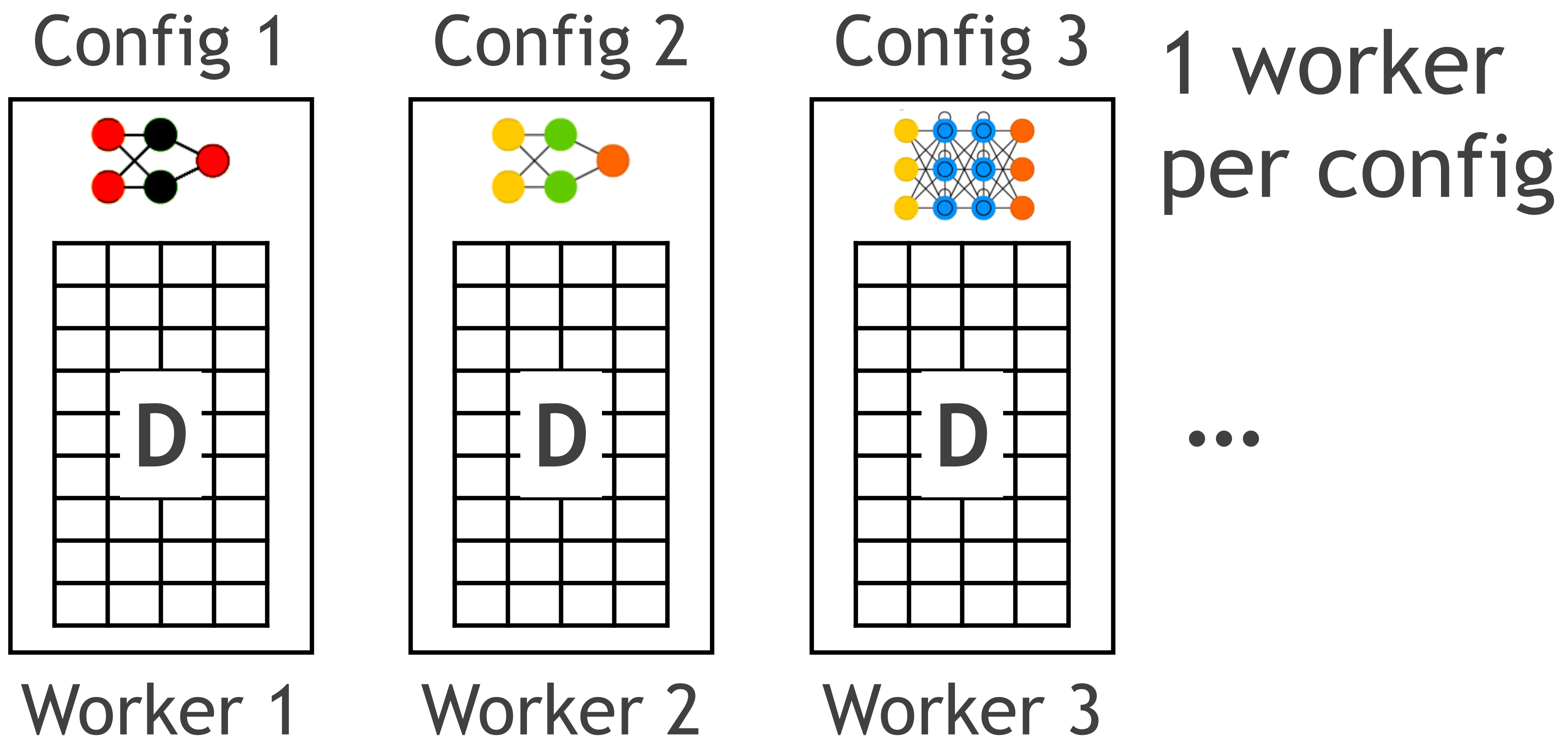


+ High throughput model selection
+ Best SGD accuracy
— Low data scalability; wastes memory/ storage (copy) or network (remote read)

Boring Wasteful Execution at Scale

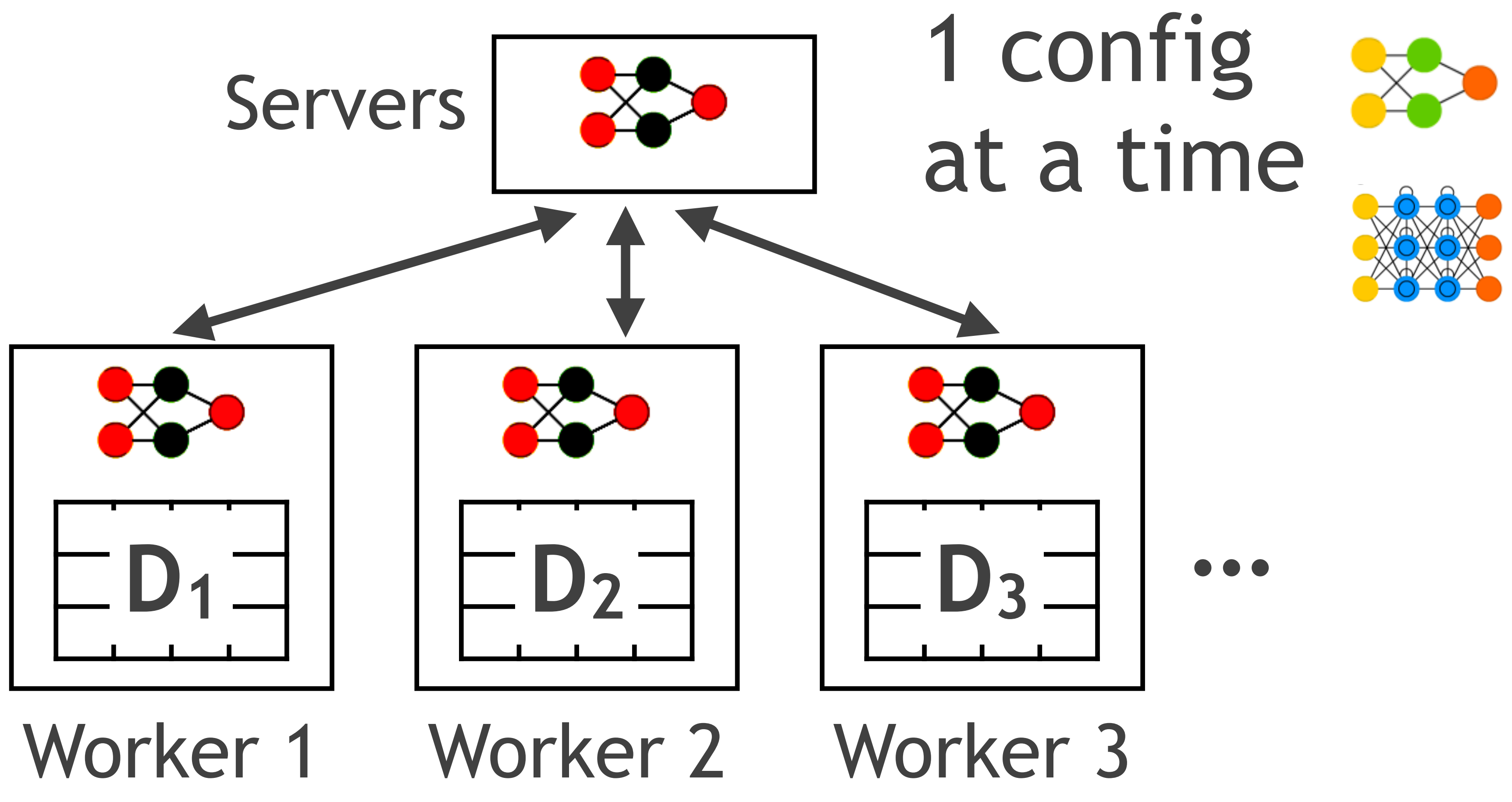
Q: How do almost all DL Systems scale model selection today?

Task Parallelism:



+ High throughput model selection
+ Best SGD accuracy
— Low data scalability; wastes memory/ storage (copy) or network (remote read)

Data Parallelism:

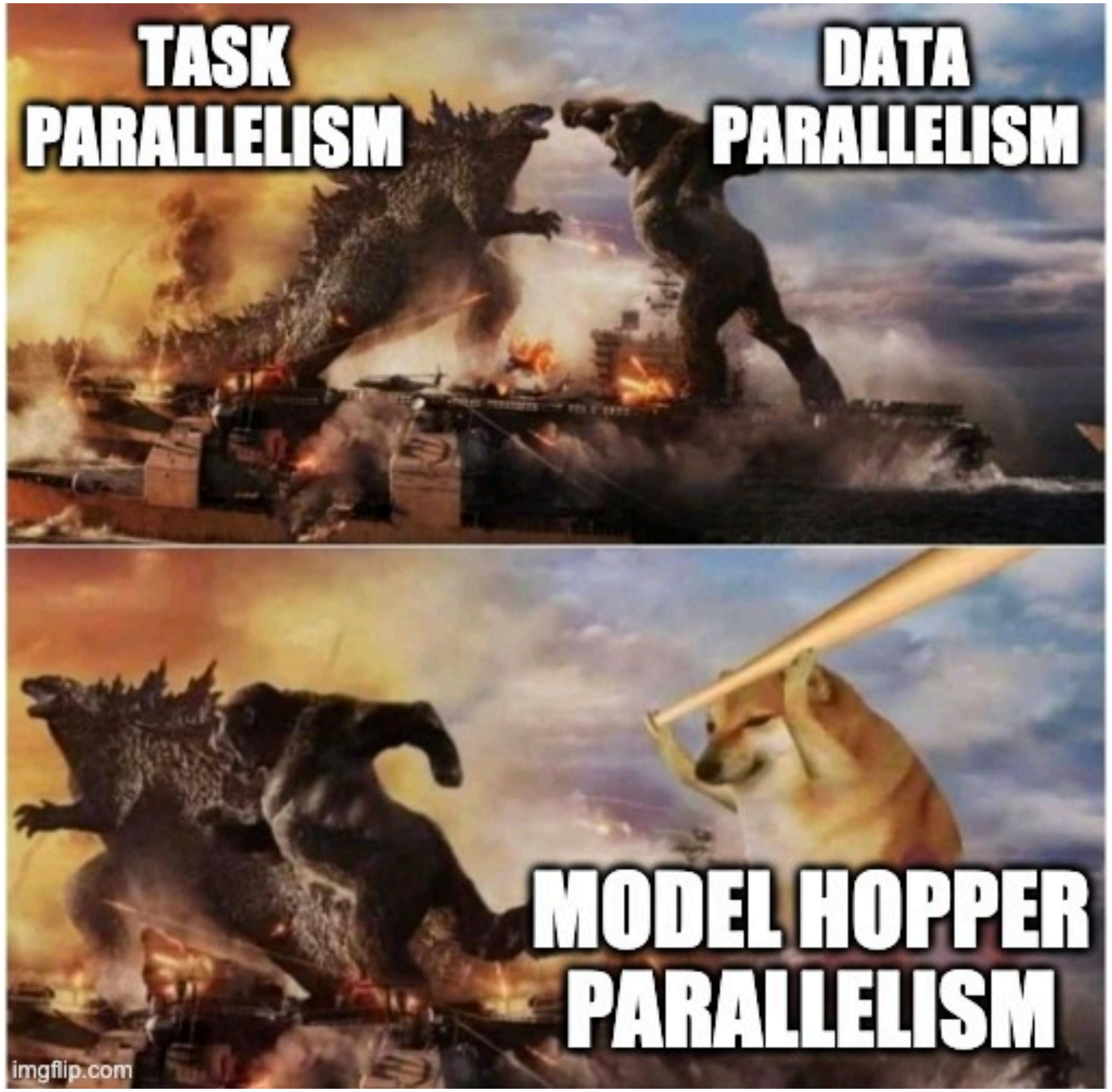


+ High data scalability
— Low throughput model selection
— Ultra-high communication costs

Enter Hybrid Parallelism for Scaling

https://adalabucsd.github.io/cerebro.html

Enter Hybrid Parallelism for Scaling



https://adalabucsd.github.io/cerebro.html

Cerebro: Model Hopper Parallelism

SGD is robust to *data ordering randomness*

https://adalabucsd.github.io/cerebro.html

Cerebro: Model Hopper Parallelism

SGD is robust to *data ordering randomness*

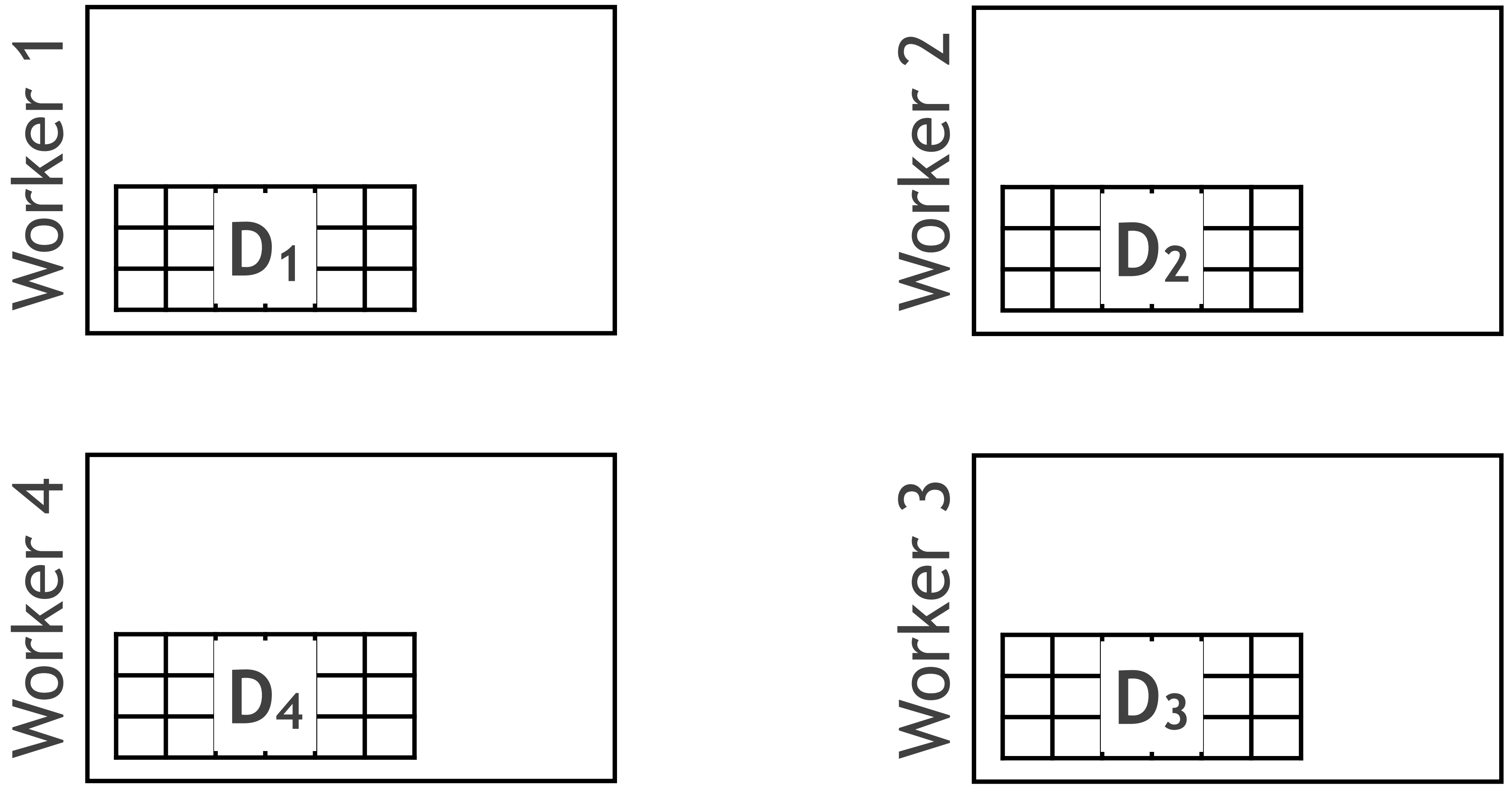
Shuffle and shard dataset
Run *n* DNNs on *n* workers

https://adalabucsd.github.io/cerebro.html

Cerebro: Model Hopper Parallelism

SGD is robust to *data ordering randomness*

Shuffle and shard dataset
Run *n* DNNs on *n* workers



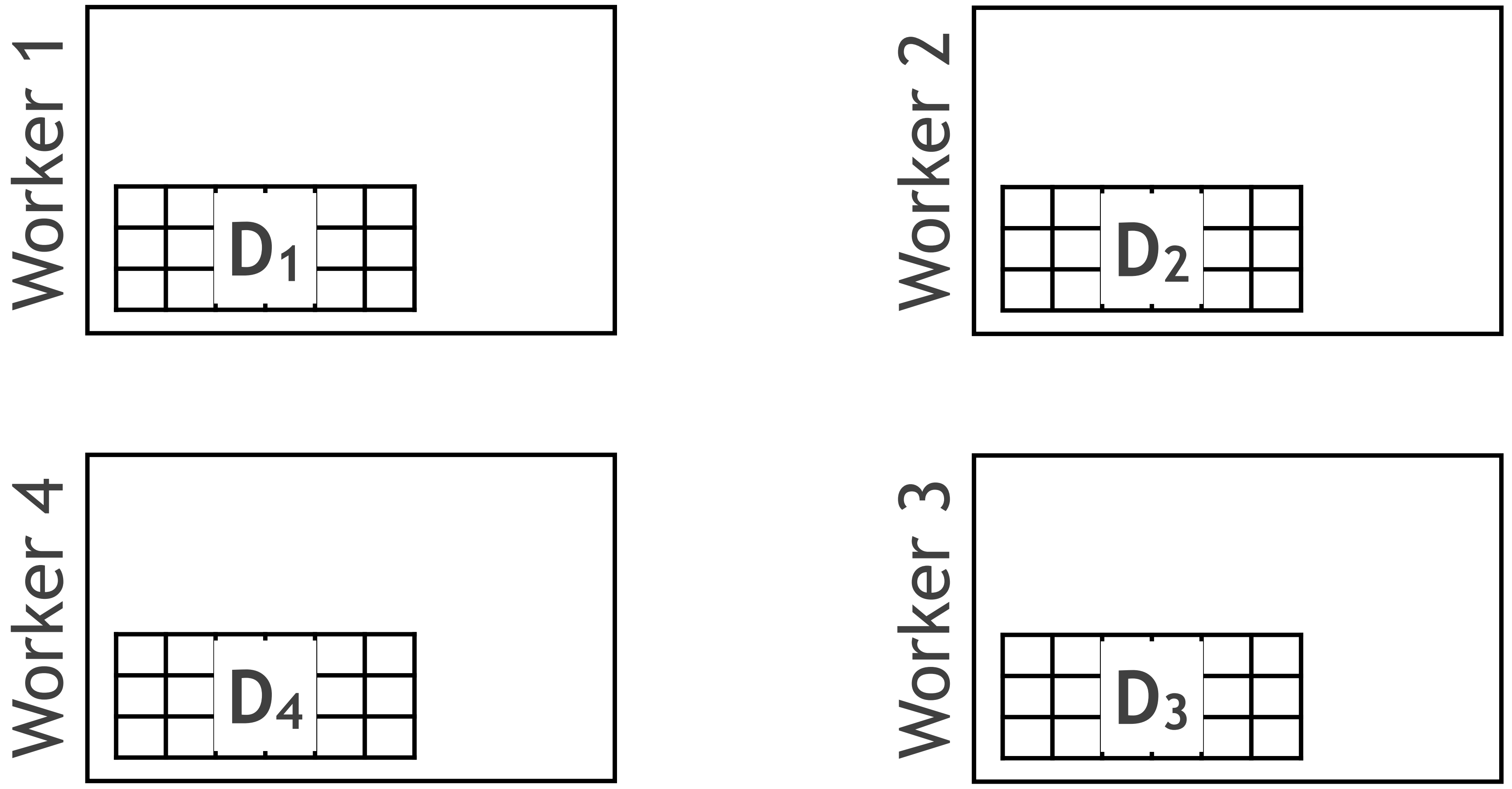
https://adalabucsd.github.io/cerebro.html

Cerebro: Model Hopper Parallelism

SGD is robust to *data ordering randomness*

Shuffle and shard dataset
Run *n* DNNs on *n* workers

Epoch 1.1 starts in parallel



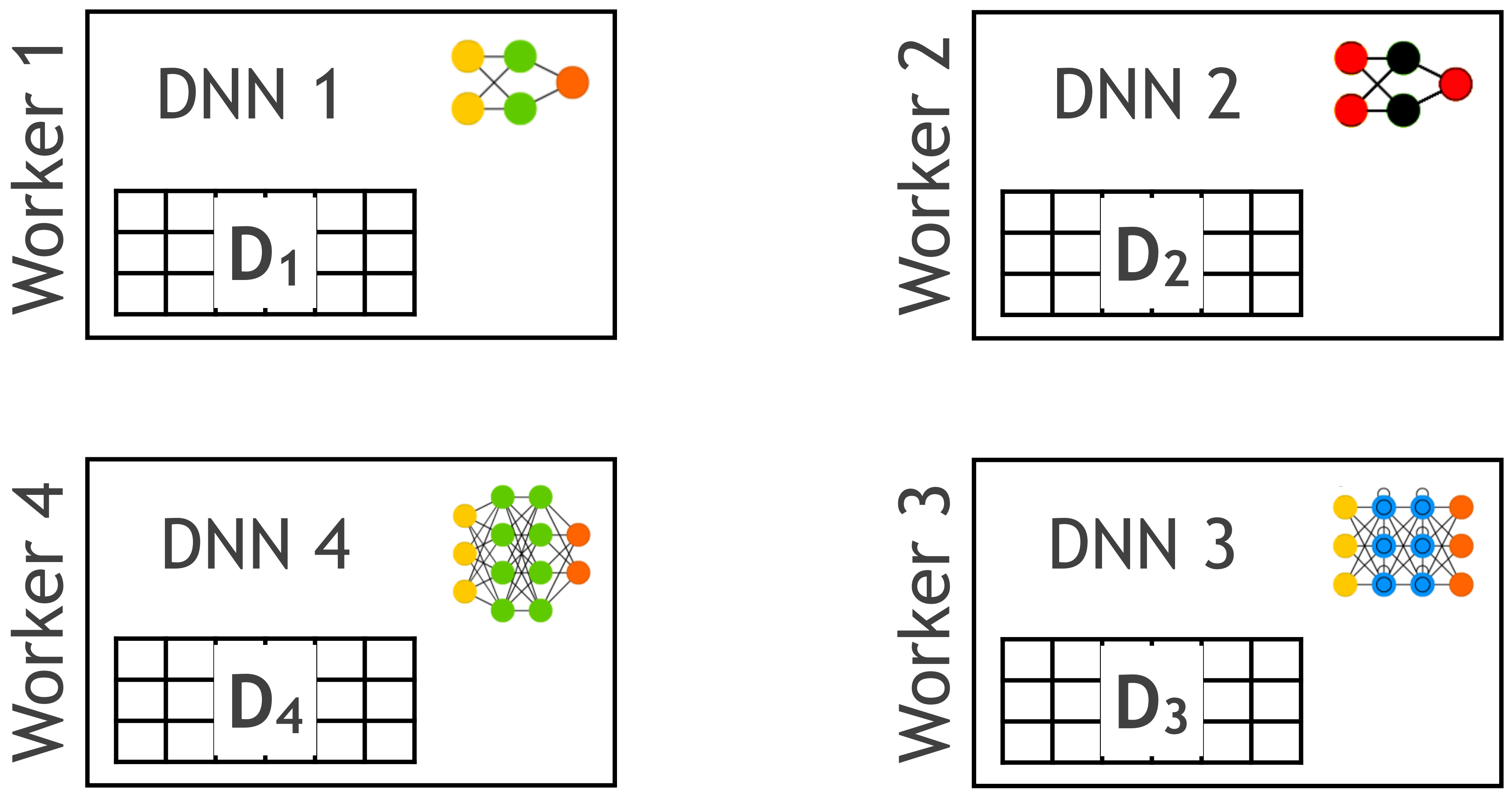
https://adalabucsd.github.io/cerebro.html

Cerebro: Model Hopper Parallelism

SGD is robust to *data ordering randomness*

Shuffle and shard dataset
Run *n* DNNs on *n* workers

Epoch 1.1 starts in parallel



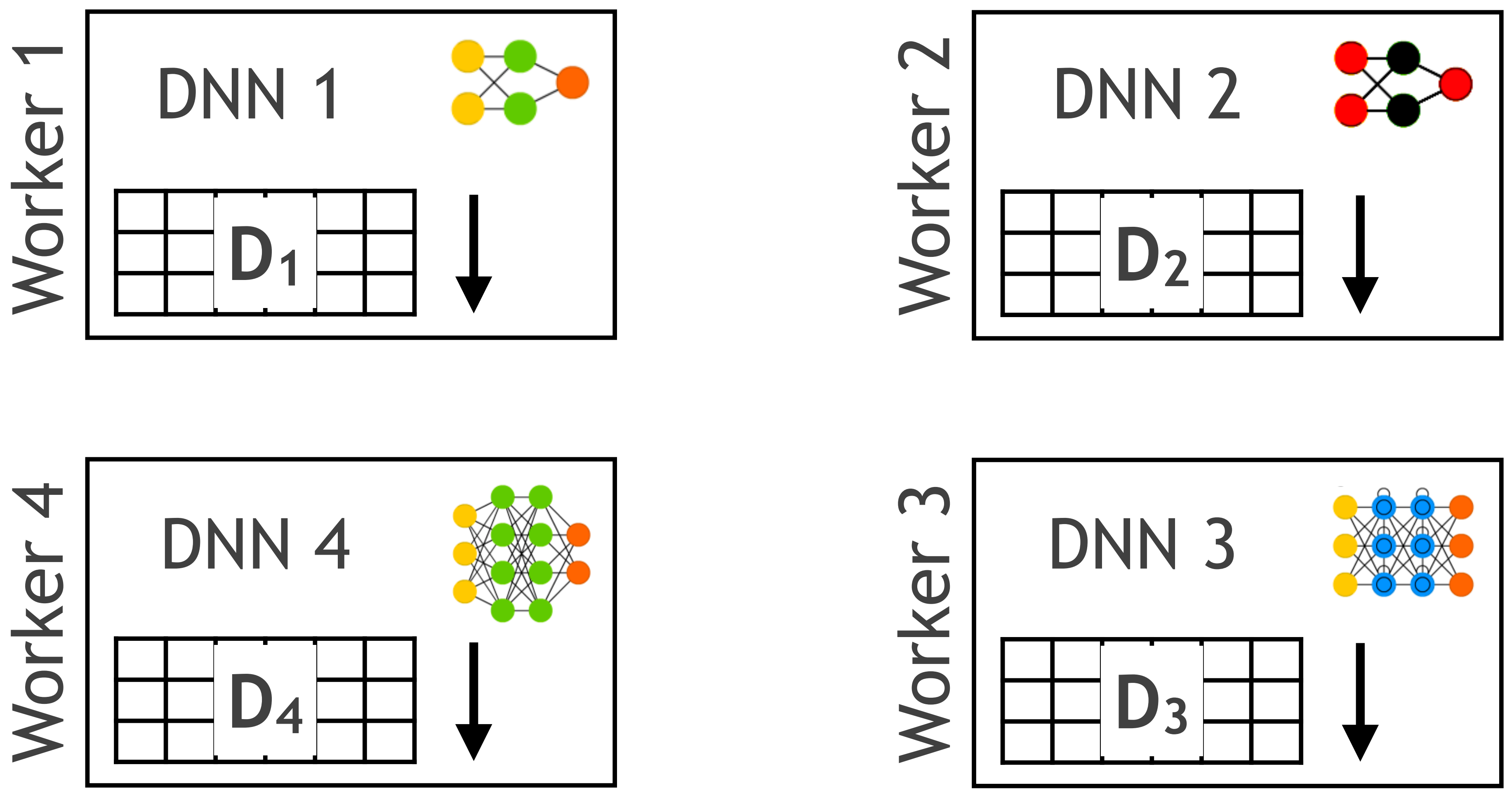
https://adalabucsd.github.io/cerebro.html

Cerebro: Model Hopper Parallelism

SGD is robust to *data ordering randomness*

Shuffle and shard dataset
Run *n* DNNs on *n* workers

Epoch 1.1 starts in parallel



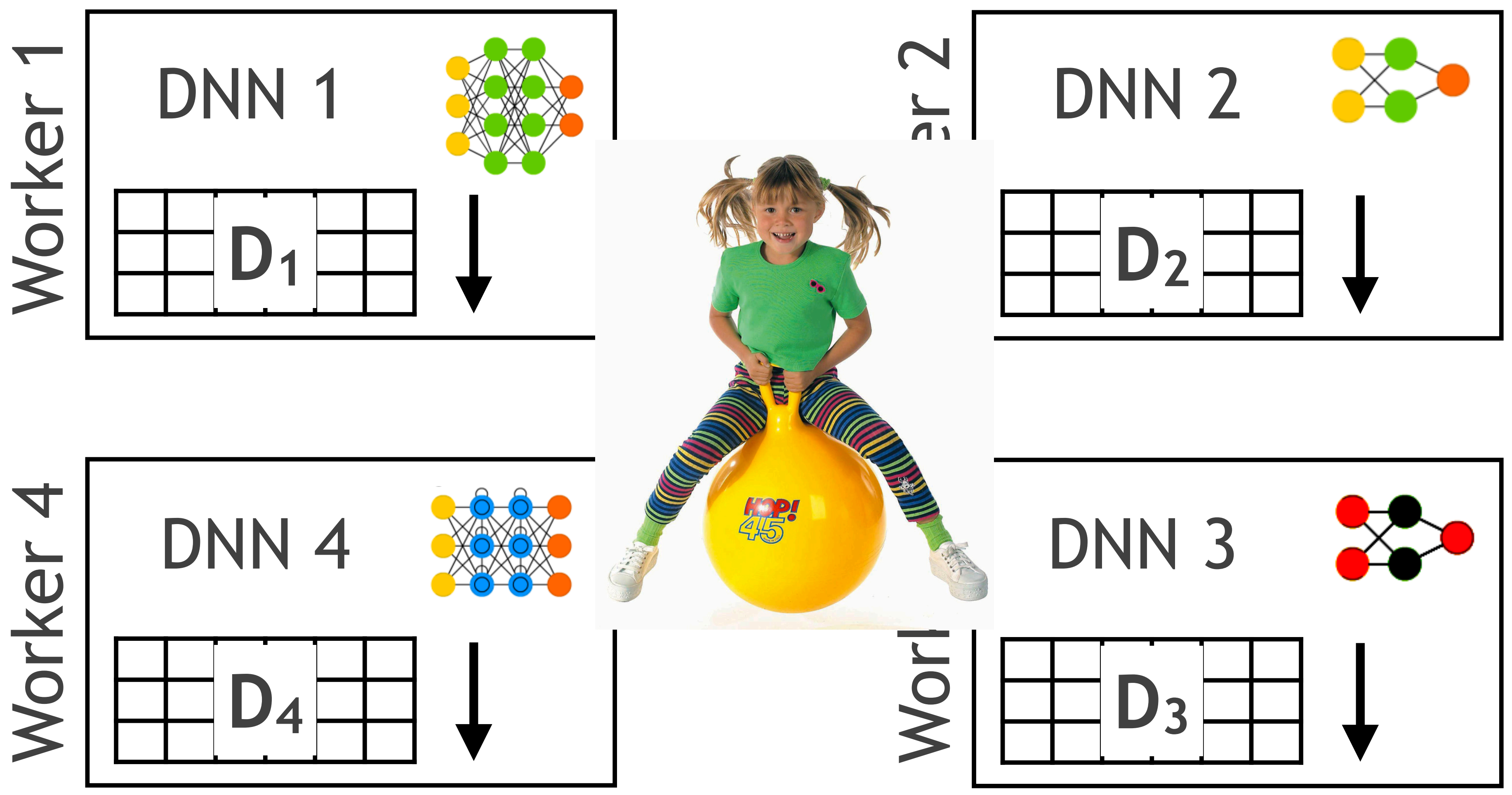
https://adalabucsd.github.io/cerebro.html

Cerebro: Model Hopper Parallelism

SGD is robust to *data ordering randomness*

Shuffle and shard dataset
Run *n* DNNs on *n* workers

Epoch 1.1 starts in parallel



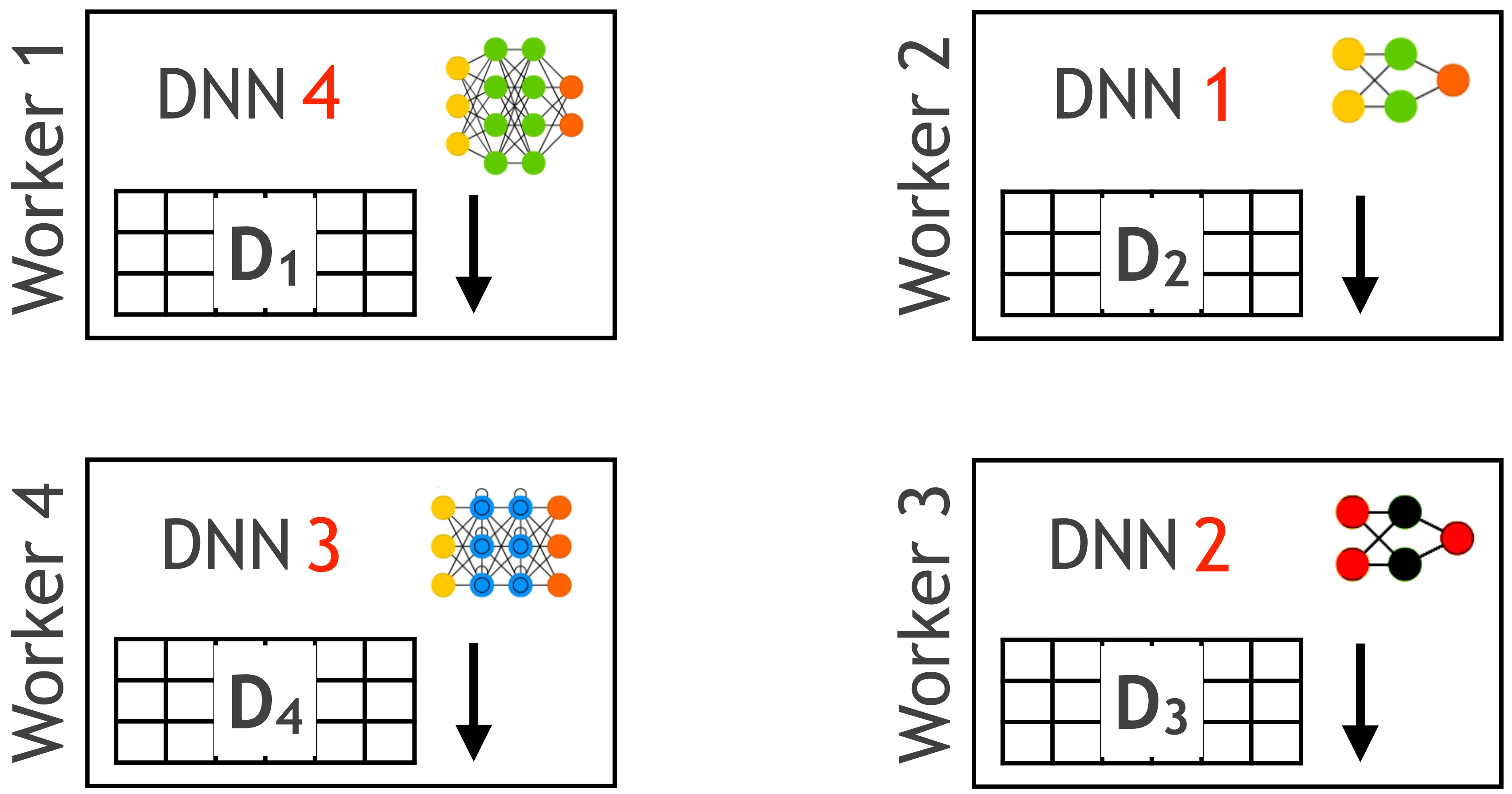
https://adalabucsd.github.io/cerebro.html

Cerebro: Model Hopper Parallelism

SGD is robust to *data ordering randomness*

Shuffle and shard dataset
Run *n* DNNs on *n* workers

Epoch 1.1 starts in parallel



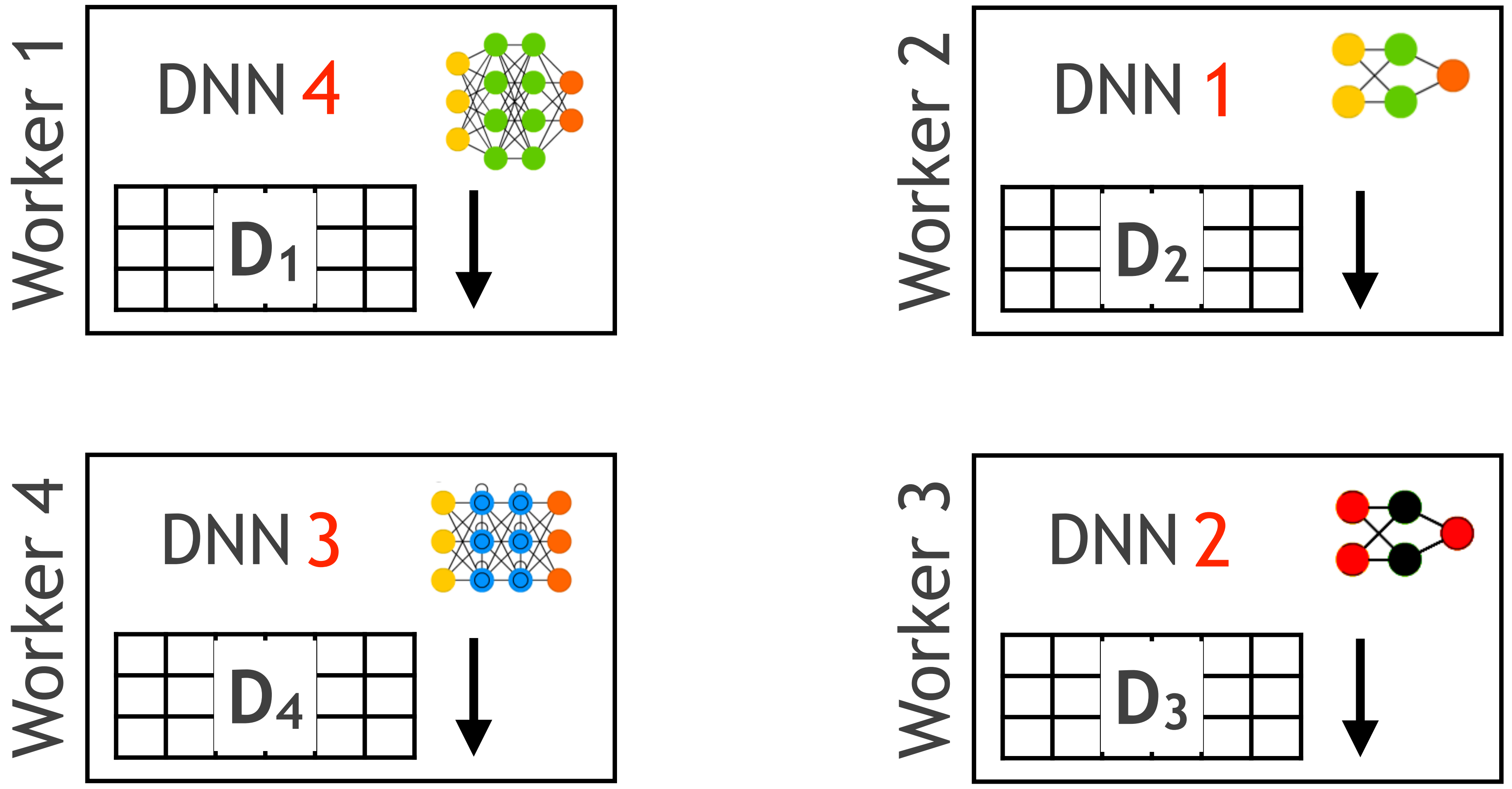
https://adalabucsd.github.io/cerebro.html

Cerebro: Model Hopper Parallelism

SGD is robust to *data ordering randomness*

Shuffle and shard dataset
Run *n* DNNs on *n* workers

Epoch 1.2 starts in parallel



https://adalabucsd.github.io/cerebro.html

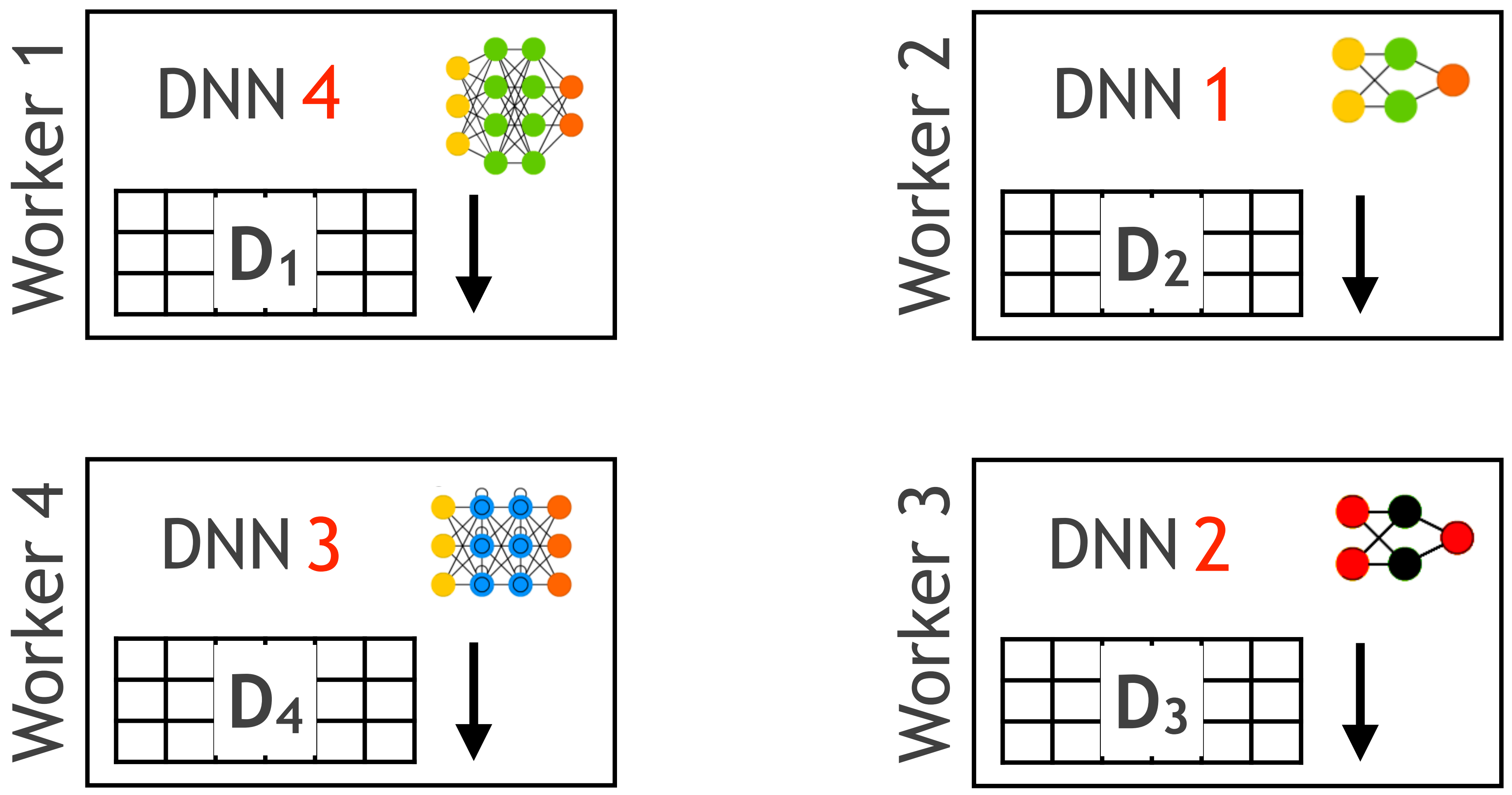
Cerebro: Model Hopper Parallelism

SGD is robust to *data ordering randomness*

Shuffle and shard dataset
Run *n* DNNs on *n* workers

Each model keeps "hopping" across shards until it sees all of D

Epoch 1.2 starts in parallel



https://adalabucsd.github.io/cerebro.html

Cerebro: Model Hopper Parallelism

SGD is robust to *data ordering randomness*

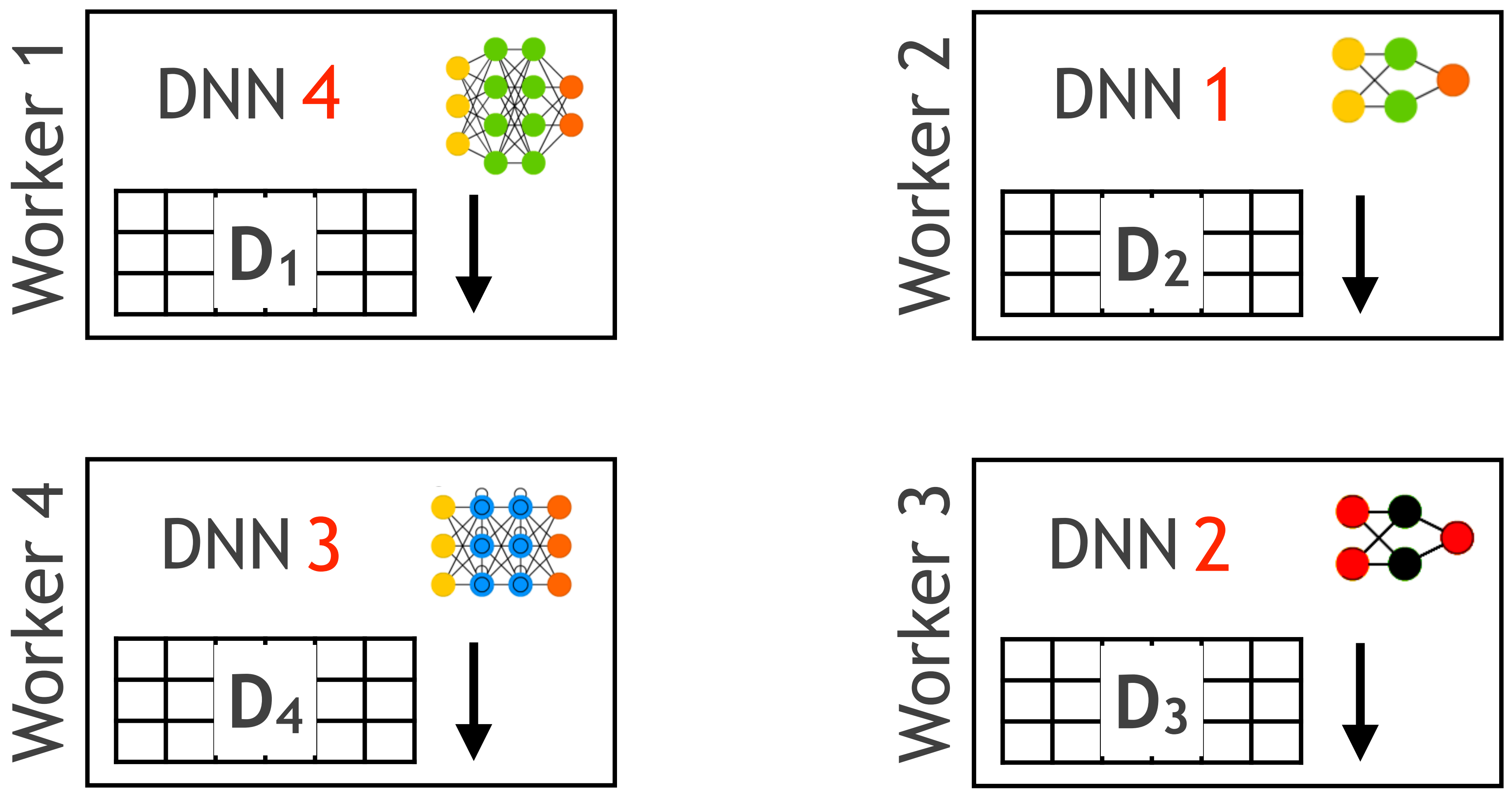
Shuffle and shard dataset
Run *n* DNNs on *n* workers

Each model keeps "hopping" across shards until it sees all of D

Strong theoretical guarantees:

1. *Equivalent* to sequential SGD
2. Hits lower bound on comm. cost

Epoch 1.2 starts in parallel



https://adalabucsd.github.io/cerebro.html

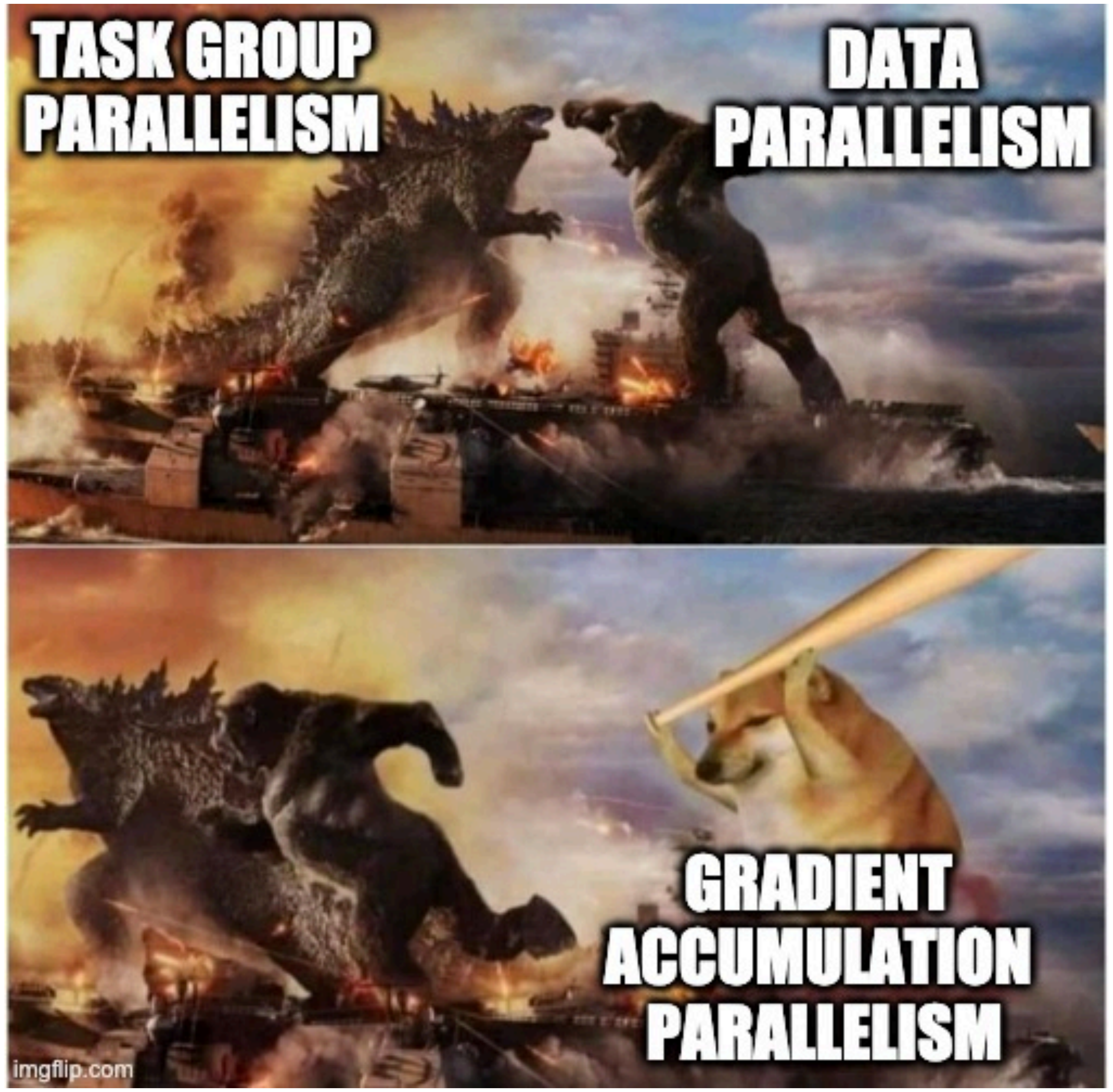
Lots More in Cerebro!

Suite of new hybrid parallelism schemes for *genuine scalability* on all possible axes: data sizes, tasks, groups, model sizes, etc.

https://adalabucsd.github.io/cerebro.html

Lots More in Cerebro!

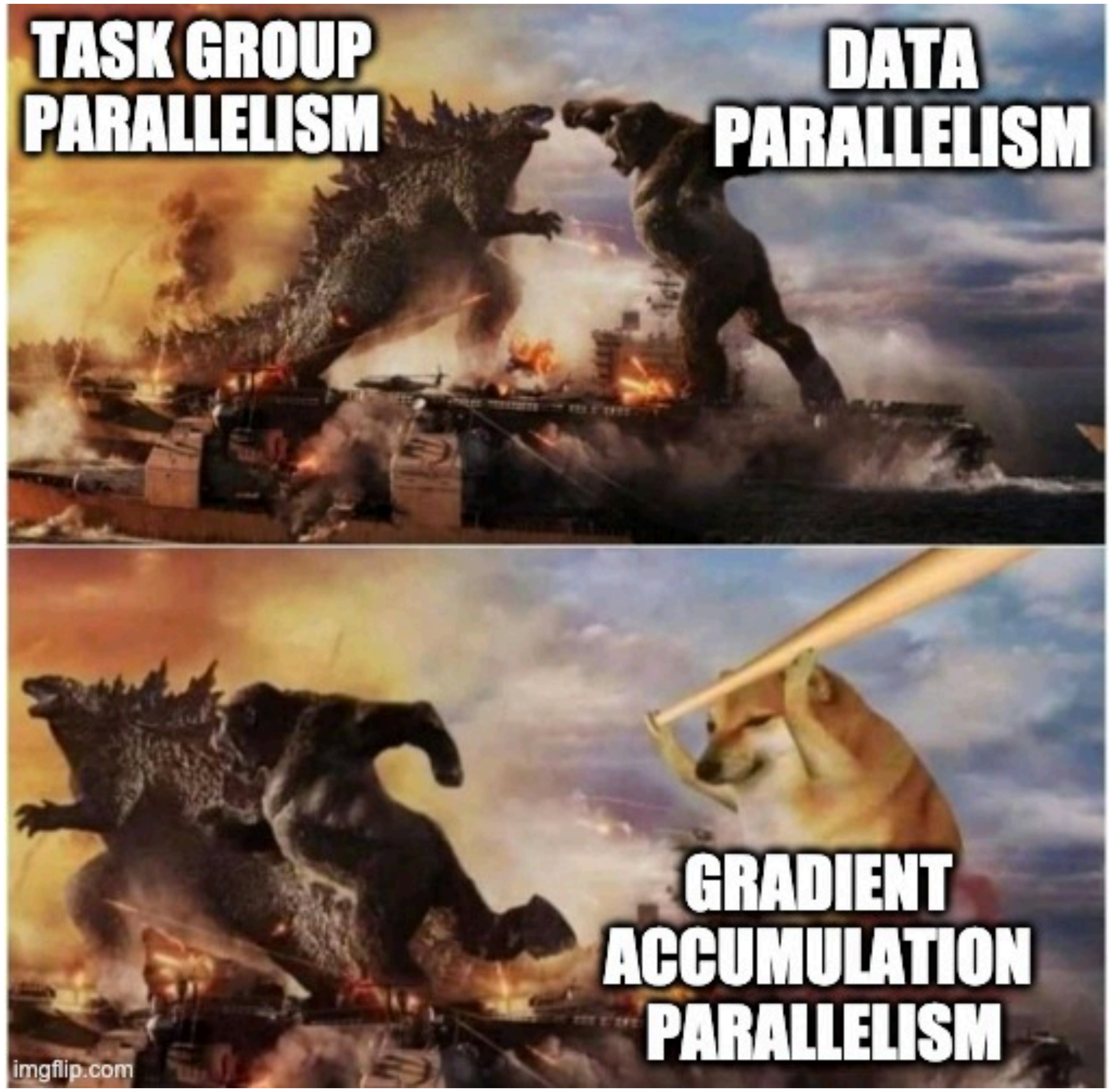
Suite of new hybrid parallelism schemes for *genuine scalability* on all possible axes: data sizes, tasks, groups, model sizes, etc.

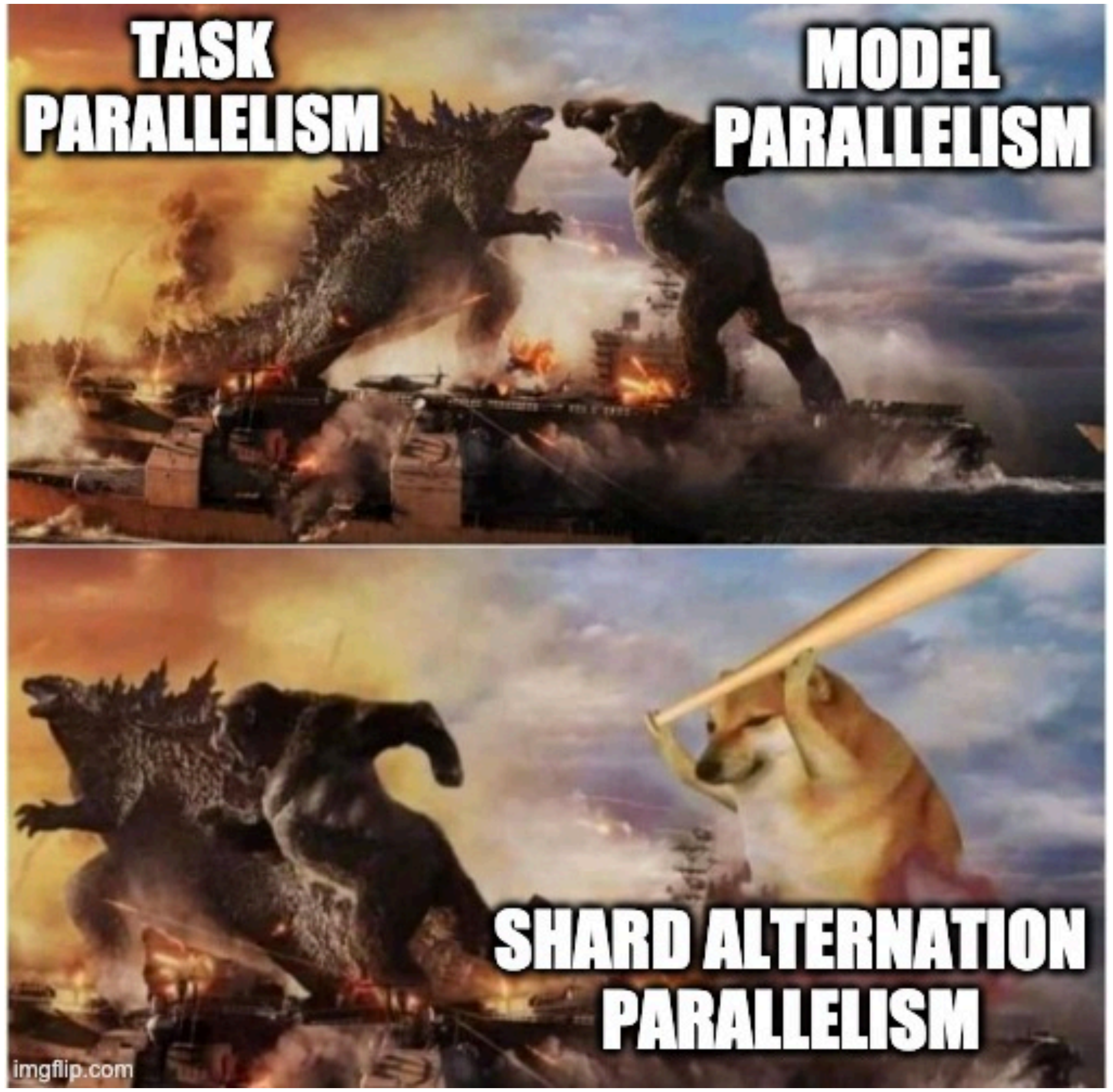


https://adalabucsd.github.io/cerebro.html

Lots More in Cerebro!

Suite of new hybrid parallelism schemes for *genuine scalability* on all possible axes: data sizes, tasks, groups, model sizes, etc.





https://adalabucsd.github.io/cerebro.html

Hybrid parallelism scales so much better.
It may even become a new trend setter.
Data, tasks, models—all are on.
Boring scaling now be gone.
Free DL systems from every scaling fetter!

How to Avoid Systems Delusion # 1:
DL Systems need hybrid parallelism to scale well

Big Tech’s Gospel of Gluttony

I can haz more GPUs, more memuhry, more masheens, more, more, more… plz!

Beware Cloud Whales' CoI!

https://adalabucsd.github.io/cerebro.html

Beware Cloud Whales' CoI!

Cloud computing indeed democratizes access to resources

https://adalabucsd.github.io/cerebro.html

Beware Cloud Whales' CoI!

Cloud computing indeed democratizes access to resources

But *pay-as-you-go* is a double-edged sword!

Cloud Whales feast on money of enterprises, small Web firms, etc.

https://adalabucsd.github.io/cerebro.html

Beware Cloud Whales' CoI!

Cloud computing indeed democratizes access to resources

But *pay-as-you-go* is a double-edged sword!

Cloud Whales feast on money of enterprises, small Web firms, etc.

Q: *How to ensure DL systems design optimizes resources holistically?*

https://adalabucsd.github.io/cerebro.html

Beware Cloud Whales' CoI!

Cloud computing indeed democratizes access to resources
But *pay-as-you-go* is a double-edged sword!
Cloud Whales feast on money of enterprises, small Web firms, etc.

Q: *How to ensure DL systems design optimizes resources holistically?*

In the RDBMS world, **query optimization** is at the heart of holistic resource efficiency that helps reduce costs

We are bringing the analog of that to scalable DL Systems in Cerebro!

https://adalabucsd.github.io/cerebro.html

Just throw more machines, says a greedy sneer.
Money or energy concerns, who cares dear?
Cloud Whales hunger for ka-ching.
Optimizing systems ain’t a thing.
But we see through their folly—and jeer!

How to Avoid Systems Delusion # 2:
DL Systems need query optimization to raise overall resource efficiency and reduce costs

My Terrific Advisees Driving Cerebro

Supun Nakandala
PhD

Yuhao Zhang
PhD & MS

Kabir Nagrecha
BS -> PhD

https://ADALabUCSD.github.io

https://ADALabUCSD.github.io

arunkk@eng.ucsd.edu

github.com/ADALabUCSD

@TweetAtAKK

ACKS:

Wake up and smell the coffee!

How to Avoid Modeling Delusion # 1:
Perform rigorous and repeatable model selection to tune task-specific B-V-N tradeoffs

How to Avoid Modeling Delusion # 2:
Hybrid human-in-the-loop + AutoML specification to rein in resource bloat

How to Avoid Modeling Delusion # 3:
Treat transfer learning rigorously as another part of model selection

How to Avoid Systems Delusion # 1:
DL Systems need hybrid parallelism to scale well

How to Avoid Systems Delusion # 2:
DL Systems need query optimization to raise overall resource efficiency and reduce costs