ReFoRM Reading Group

Rethinking Foundations for Real-world ML

Amin Saberi & Andrew Ilyas

Welcome to ReFoRM!

Welcome to ReFoRM!

What this is: an experimental reading group on foundations of “real-world” ML

Welcome to ReFoRM!

What this is: an experimental reading group on foundations of “real-world” ML

What does this mean?

Welcome to ReFoRM!

Not a slight to other ML!
Just starts with R 😅

What this is: an experimental reading group on foundations of "real-world" ML

What does this mean?

Welcome to ReFoRM!

Not a slight to other ML!
Just starts with R 😅

What this is: an experimental reading group on foundations of "real-world" ML

What does this mean?

Idealized picture of ML: something like $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{z_i \sim D}[\ell(z_i; \theta)]$

Welcome to ReFoRM!

Not a slight to other ML!
Just starts with R 😅

What this is: an experimental reading group on foundations of "real-world" ML

What does this mean?

Idealized picture of ML: something like $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{z_i \sim D}[\ell(z_i; \theta)]$

Design decisions: choosing Θ to avoid overfitting, choosing a good (convex) loss function ℓ, what optimizer to use for efficiency...

Welcome to ReFoRM!

Not a slight to other ML!
Just starts with R 😅

What this is: an experimental reading group on foundations of "real-world" ML

What does this mean?

Idealized picture of ML: something like $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{z_i \sim D}[\ell(z_i; \theta)]$

Design decisions: choosing Θ to avoid overfitting, choosing a good (convex) loss function ℓ, what optimizer to use for efficiency…

Guarantees: Convergence rates, generalization bounds, out-of-distribution error control, uncertainty quantification (e.g., via confidence intervals)

Welcome to ReFoRM!

Not a slight to other ML!
Just starts with R 😅

What this is: an experimental reading group on foundations of "real-world" ML

What does this mean?

Idealized picture of ML: something like $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{z_i \sim D}[\ell(z_i; \theta)]$

ML powering systems like Claude, DALL-E, Google Photos:

Welcome to ReFoRM!

Not a slight to other ML!
Just starts with R 😅

What this is: an experimental reading group on foundations of "real-world" ML

What does this mean?

Idealized picture of ML: something like $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{z_i \sim D}[\ell(z_i; \theta)]$

ML powering systems like Claude, DALL-E, Google Photos:

Huge, messy dataset D_{full}

Welcome to ReFoRM!

Not a slight to other ML!
Just starts with R 😅

What this is: an experimental reading group on foundations of "real-world" ML

What does this mean?

Idealized picture of ML: something like $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{z_i \sim D}[\ell(z_i; \theta)]$

ML powering systems like Claude, DALL-E, Google Photos:

Huge, messy dataset D_{full} $\xrightarrow{\text{Data selection, curation, synthesis}}$

Welcome to ReFoRM!

Not a slight to other ML!
Just starts with R 😅

What this is: an experimental reading group on foundations of "real-world" ML

What does this mean?

Idealized picture of ML: something like $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{z_i \sim D}[\ell(z_i; \theta)]$

ML powering systems like Claude, DALL-E, Google Photos:

$$\text{Huge, messy dataset } D_{full} \xrightarrow{\text{Data selection, curation, synthesis}} \text{Training set } D$$

Welcome to ReFoRM!

Not a slight to other ML!
Just starts with R 😅

What this is: an experimental reading group on foundations of "real-world" ML

What does this mean?

Idealized picture of ML: something like $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{z_i \sim D}[\ell(z_i; \theta)]$

ML powering systems like Claude, DALL-E, Google Photos:

Huge, messy dataset D_{full} $\xrightarrow{\text{Data selection, curation, synthesis}}$ Training set D

Extremely expressive hypothesis class Θ

Welcome to ReFoRM!

Not a slight to other ML!
Just starts with R 😅

What this is: an experimental reading group on foundations of "real-world" ML

What does this mean?

Idealized picture of ML: something like $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{z_i \sim D}[\ell(z_i; \theta)]$

ML powering systems like Claude, DALL-E, Google Photos:

Huge, messy dataset D_{full} $\xrightarrow{\text{Data selection, curation, synthesis}}$ Training set D

Extremely expressive hypothesis class Θ + Bespoke regularization

Welcome to ReFoRM!

Not a slight to other ML!
Just starts with R 😅

What this is: an experimental reading group on foundations of "real-world" ML

What does this mean?

Idealized picture of ML: something like $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{z_i \sim D}[\ell(z_i; \theta)]$

ML powering systems like Claude, DALL-E, Google Photos:

Huge, messy dataset D_{full} $\xrightarrow{\text{Data selection, curation, synthesis}}$ Training set D

Extremely expressive hypothesis class Θ + Bespoke regularization + Optimizer

Welcome to ReFoRM!

Not a slight to other ML!
Just starts with R 😅

What this is: an experimental reading group on foundations of “real-world” ML

What does this mean?

Idealized picture of ML: something like $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{z_i \sim D}[\ell(z_i; \theta)]$

ML powering systems like Claude, DALL-E, Google Photos:

Huge, messy dataset D_{full} $\xrightarrow{\text{Data selection, curation, synthesis}}$ Training set D

Learning algorithm

=

Extremely expressive hypothesis class Θ + Bespoke regularization + Optimizer

Welcome to ReFoRM!

Not a slight to other ML!
Just starts with R 😅

What this is: an experimental reading group on foundations of "real-world" ML

What does this mean?

Idealized picture of ML: something like $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{z_i \sim D}[\ell(z_i; \theta)]$

ML powering systems like Claude, DALL-E, Google Photos:

Huge, messy dataset D_{full} $\xrightarrow{\text{Data selection, curation, synthesis}}$ Training set D + Learning algorithm = Learned θ

Learning algorithm = **Extremely** expressive hypothesis class Θ + Bespoke regularization + Optimizer

Welcome to ReFoRM!

Not a slight to other ML!
Just starts with R 😅

What this is: an experimental reading group on foundations of "real-world" ML

What does this mean?

Idealized picture of ML: something like $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{z_i \sim D}[\ell(z_i; \theta)]$

ML powering systems like Claude, DALL-E, Google Photos:

Huge, messy dataset D_{full} $\xrightarrow{\text{Data selection, curation, synthesis}}$ Training set D + Learning algorithm = Learned θ $\xrightarrow{\text{"Post-training"}}$

Learning algorithm = **Extremely** expressive hypothesis class Θ + Bespoke regularization + Optimizer

Welcome to ReFoRM!

Not a slight to other ML!
Just starts with R 😅

What this is: an experimental reading group on foundations of "real-world" ML

What does this mean?

Idealized picture of ML: something like $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{z_i \sim D}[\ell(z_i; \theta)]$

ML powering systems like Claude, DALL-E, Google Photos:

$$\text{Huge, messy dataset } D_{full} \xrightarrow{\text{Data selection, curation, synthesis}} \text{Training set } D + \text{Learning algorithm} = \text{Learned } \theta \xrightarrow{\text{"Post-training"}} \text{Deployed model}$$

Learning algorithm = **Extremely** expressive hypothesis class Θ + Bespoke regularization + Optimizer

Welcome to ReFoRM!

Not a slight to other ML!
Just starts with R 😅

What this is: an experimental reading group on foundations of "real-world" ML

What does this mean?

Idealized picture of ML: something like $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{z_i \sim D}[\ell(z_i; \theta)]$

ML powering systems like Claude, DALL-E, Google Photos:

Huge, messy dataset D_{full} $\xrightarrow{\text{Data selection, curation, synthesis}}$ Training set D + Learning algorithm = Learned θ $\xrightarrow{\text{“Post-training"}}$ Deployed model

Evaluation and auditing

Learning algorithm = **Extremely** expressive hypothesis class Θ + Bespoke regularization + Optimizer

Welcome to ReFoRM!

Not a slight to other ML!
Just starts with R 😅

What this is: an experimental reading group on foundations of "real-world" ML

What does this mean?

Idealized picture of ML: something like $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{z_i \sim D}[\ell(z_i; \theta)]$

ML powering systems like Claude, DALL-E, Google Photos:

Huge, messy dataset D_{full} $\xrightarrow{\text{Data selection, curation, synthesis}}$ Training set D + Learning algorithm = Learned θ $\xrightarrow{\text{"Post-training"}}$ Deployed model

Learning algorithm = **Extremely** expressive hypothesis class Θ + Bespoke regularization + Optimizer

Deployed model $\xrightarrow{\text{Evaluation and auditing}}$ Actual target metric

Welcome to ReFoRM!

Not a slight to other ML!
Just starts with R 😅

What this is: an experimental reading group on foundations of "real-world" ML

What does this mean?

Idealized picture of ML: something like $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{z_i \sim D}[\ell(z_i; \theta)]$

ML powering systems like Claude, DALL-E, Google Photos:

Huge, messy dataset D_{full} $\xrightarrow{\text{Data selection, curation, synthesis}}$ Training set D + Learning algorithm = Learned θ $\xrightarrow{\text{"Post-training"}}$ Deployed model

Learning algorithm = **Extremely** expressive hypothesis class Θ + Bespoke regularization + Optimizer

Bigger is better! But more expensive...

Deployed model $\xrightarrow{\text{Evaluation and auditing}}$ Actual target metric

Welcome to ReFoRM!

Not a slight to other ML!
Just starts with R 😅

What this is: an experimental reading group on foundations of "real-world" ML

What does this mean?

Idealized picture of ML: something like $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{z_i \sim D}[\ell(z_i; \theta)]$

ML powering systems like Claude, DALL-E, Google Photos:

Huge, messy dataset D_{full} $\xrightarrow{\text{Data selection, curation, synthesis}}$ Training set D + Learning algorithm = Learned θ $\xrightarrow{\text{"Post-training"}}$ Deployed model

Learning algorithm = **Extremely** expressive hypothesis class Θ + Bespoke regularization + Optimizer

Bigger is better! But more expensive...

...and regularization still matters (a lot!)

Deployed model $\xrightarrow{\text{Evaluation and auditing}}$ Actual target metric

Welcome to ReFoRM!

Not a slight to other ML!
Just starts with R 😅

What this is: an experimental reading group on foundations of "real-world" ML

What does this mean?

Idealized picture of ML: something like $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{z_i \sim D}[\ell(z_i; \theta)]$

ML powering systems like Claude, DALL-E, Google Photos:

Huge, messy dataset D_{full} $\xrightarrow{\text{Data selection, curation, synthesis}}$ Training set D + Learning algorithm = Learned θ $\xrightarrow{\text{"Post-training"}}$ Deployed model

Deployed model $\xrightarrow{\text{Evaluation and auditing}}$ Actual target metric

Learning algorithm

||

Extremely expressive hypothesis class Θ + Bespoke regularization + Optimizer

Bigger is better! But more expensive...

...and regularization still matters (a lot!)

Optimizer is now part efficiency part regularization

Welcome to ReFoRM!

ML powering systems like Claude, DALL-E, Google Photos:

Huge, messy dataset D_{full} $\xrightarrow{\text{Data selection, curation, synthesis}}$ Training set D + Learning algorithm = Learned θ $\xrightarrow{\text{"Post-training"}}$ Deployed model $\xrightarrow{\text{Evaluation and auditing}}$ Actual target metric

Learning algorithm = **Extremely** expressive hypothesis class Θ + Bespoke regularization + Optimizer

Bigger is better! But more expensive…

…and regularization still matters (a lot!)

Optimizer is now part efficiency part regularization

Implications: unpredictability, new considerations, invalidated assumptions

Goal of this group

Goal of this group

What do rigorous foundations for this new age of ML look like?

Goal of this group

What do rigorous foundations for this new age of ML look like?

How can tools from statistics, CS theory, and operations inform a better understanding of machine learning algorithms and systems?

Goal of this group

What do rigorous foundations for this new age of ML look like?

How can tools from statistics, CS theory, and operations inform a better understanding of machine learning algorithms and systems?

What are the right questions to ask, and phenomena to explain—at what level of abstraction should we be aiming to explain them?

Goal of this group

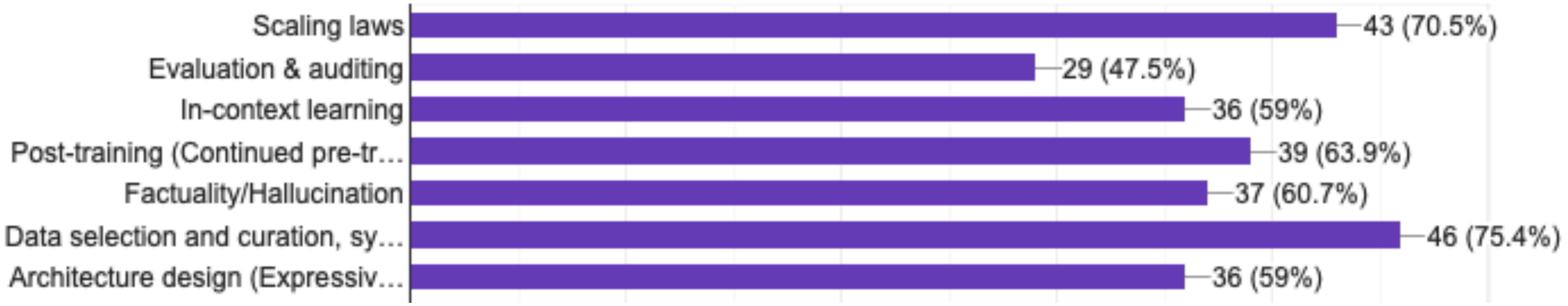
What do rigorous foundations for this new age of ML look like?

How can tools from statistics, CS theory, and operations inform a better understanding of machine learning algorithms and systems?

What are the right questions to ask, and phenomena to explain—at what level of abstraction should we be aiming to explain them?

What theoretical models not only explain unexpected phenomena, but also predict new phenomena that we can verify experimentally?

List of topics



List of topics

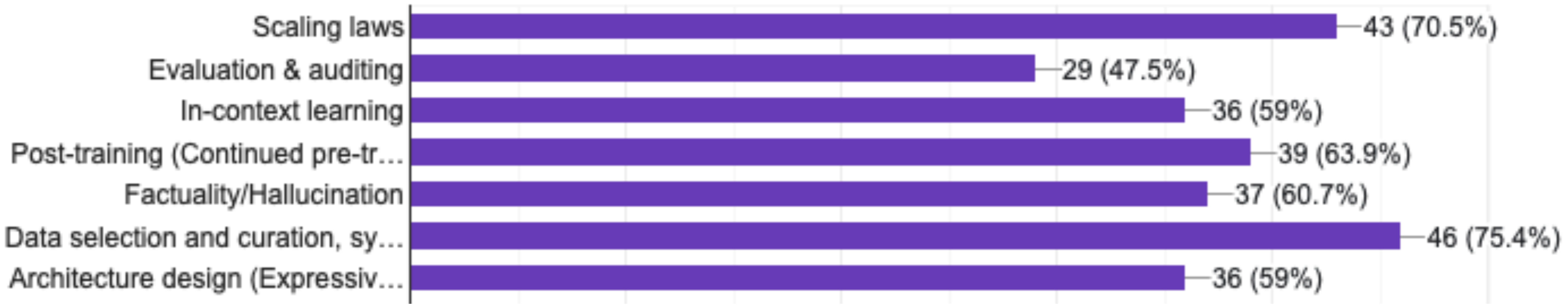
Topics by weighted combination of {interest, coverage}:

Data selection, curation, and synthesis

Scaling laws & prediction

Expressivity & architectures/Evaluation & Auditing/Factuality

Post-training (continued pre-training, preference tuning, …)



List of topics

Topics by weighted combination of {interest, coverage}:

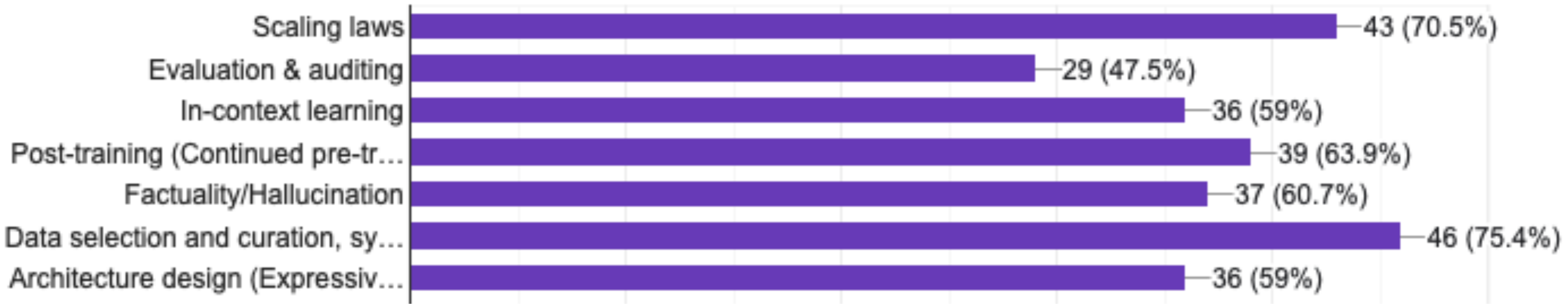
Data selection, curation, and synthesis

Scaling laws & prediction

2025

Expressivity & architectures/Evaluation & Auditing/Factuality

Post-training (continued pre-training, preference tuning, ...)



List of topics

Topics by weighted combination of {interest, coverage}:

Data selection, curation, and synthesis

Scaling laws & prediction

2025

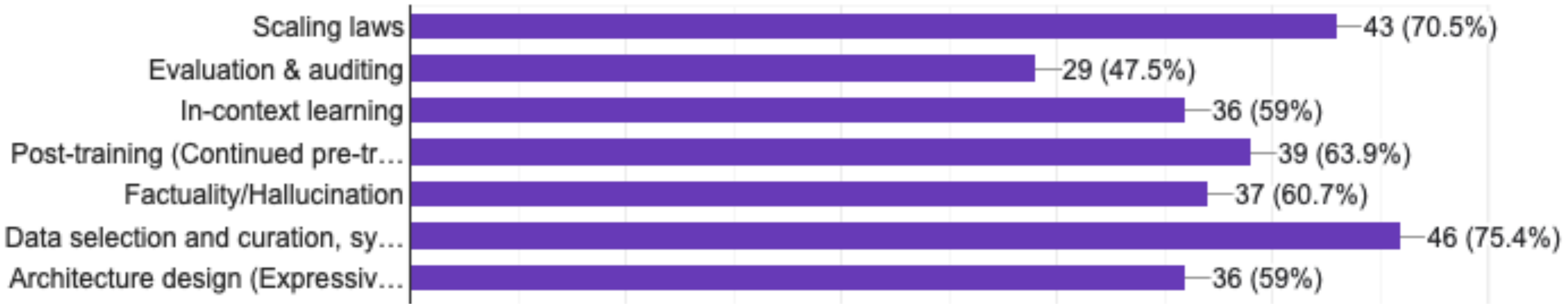
Expressivity & architectures/Evaluation & Auditing/Factuality

Post-training (continued pre-training, preference tuning, ...)

Simple descriptive and predictive models

Where does theory agree/ disagree with practice?

Where can we draw from known techniques?



Intended format (please sign up!)

Intended format (please sign up!)

Goal: Build intuition, leverage diversity in this group, start collaborations (bringing new perspectives from everyone's field)

Intended format (please sign up!)

Goal: Build intuition, leverage diversity in this group, start collaborations (bringing new perspectives from everyone's field)

Sign up to be a discussant at https://tinyurl.com/reform-ml-signup

Intended format (please sign up!)

Goal: Build intuition, leverage diversity in this group, start collaborations (bringing new perspectives from everyone's field)

Sign up to be a discussant at https://tinyurl.com/reform-ml-signup

Goal(s) of the discussant:

Intended format (please sign up!)

Goal: Build intuition, leverage diversity in this group, start collaborations (bringing new perspectives from everyone's field)

Sign up to be a discussant at https://tinyurl.com/reform-ml-signup

Goal(s) of the discussant:

1. A single "deep dive" per week about one subject (can be multiple papers) by 1-2 discussants

Intended format (please sign up!)

Goal: Build intuition, leverage diversity in this group, start collaborations (bringing new perspectives from everyone's field)

Sign up to be a discussant at https://tinyurl.com/reform-ml-signup

Goal(s) of the discussant:

1. A single "deep dive" per week about one subject (can be multiple papers) by 1-2 discussants
2. We have suggested several papers for each week, more than one can cover thoroughly in a week. **Pick a small, focused set of papers and read them thoroughly**

Intended format (please sign up!)

Goal: Build intuition, leverage diversity in this group, start collaborations (bringing new perspectives from everyone's field)

Sign up to be a discussant at https://tinyurl.com/reform-ml-signup

Goal(s) of the discussant:

1. A single "deep dive" per week about one subject (can be multiple papers) by 1-2 discussants
2. We have suggested several papers for each week, more than one can cover thoroughly in a week. **Pick a small, focused set of papers and read them thoroughly**
3. Prepare a 20-30 minute presentation, accessible to a second year PhD student, focusing on (a) seeding discussion and (b) identifying gaps and connections, and (c) formulating open problems

Intended format (please sign up!)

Goal: Build intuition, leverage diversity in this group, start collaborations (bringing new perspectives from everyone's field)

Sign up to be a discussant at https://tinyurl.com/reform-ml-signup

Goal(s) of the discussant:

1. A single "deep dive" per week about one subject (can be multiple papers) by 1-2 discussants
2. We have suggested several papers for each week, more than one can cover thoroughly in a week. **Pick a small, focused set of papers and read them thoroughly**
3. Prepare a 20-30 minute presentation, accessible to a second year PhD student, focusing on (a) seeding discussion and (b) identifying gaps and connections, and (c) formulating open problems

Everyone else: Read the paper/watch a podcast/something! Try to come with some familiarity

Today's meeting

Introduce topics & papers for this year (scaling laws & data selection)

For each topic:

- Problem setup/definition
- Motivation
- Methodology
- Extensions

Scaling laws

Scaling laws

Overarching question: How does “scaling up” a given training setup change the resulting machine learning model behavior?

Scaling laws

Overarching question: How does “scaling up” a given training setup change the resulting machine learning model behavior?

Problem setup: Scaling in deep learning is typically along one of three axes: model complexity (proxied by number of parameters), dataset size, amount of compute (# of training steps over the data)

Scaling laws

Overarching question: How does “scaling up” a given training setup change the resulting machine learning model behavior?

Problem setup: Scaling in deep learning is typically along one of three axes: model complexity (proxied by number of parameters), dataset size, amount of compute (# of training steps over the data)

Goal: Make a predictor for the average test loss as a function of these scaling axes

Scaling laws

Overarching question: How does "scaling up" a given training setup change the resulting machine learning model behavior?

Problem setup: Scaling in deep learning is typically along one of three axes: model complexity (proxied by number of parameters), dataset size, amount of compute (# of training steps over the data)

Goal: Make a predictor for the average test loss as a function of these scaling axes

$$\ell \propto f_\beta(N, D, C)$$

Scaling laws

Scaling laws

Classical analog: Estimation error bounds

e.g., fixed-design linear regression:

Scaling laws

Classical analog: Estimation error bounds

e.g., fixed-design linear regression:

$$\|\theta - \theta^*\|_{\Sigma}^2 \leq O\left(\frac{d}{n}\right)$$

Scaling laws

Classical analog: Estimation error bounds

e.g., fixed-design linear regression:

$$\|\theta - \theta^*\|_\Sigma^2 \leq O\left(\frac{d}{n}\right)$$

In deep learning, we can't prove bounds, so we empirically fit trends to actual data

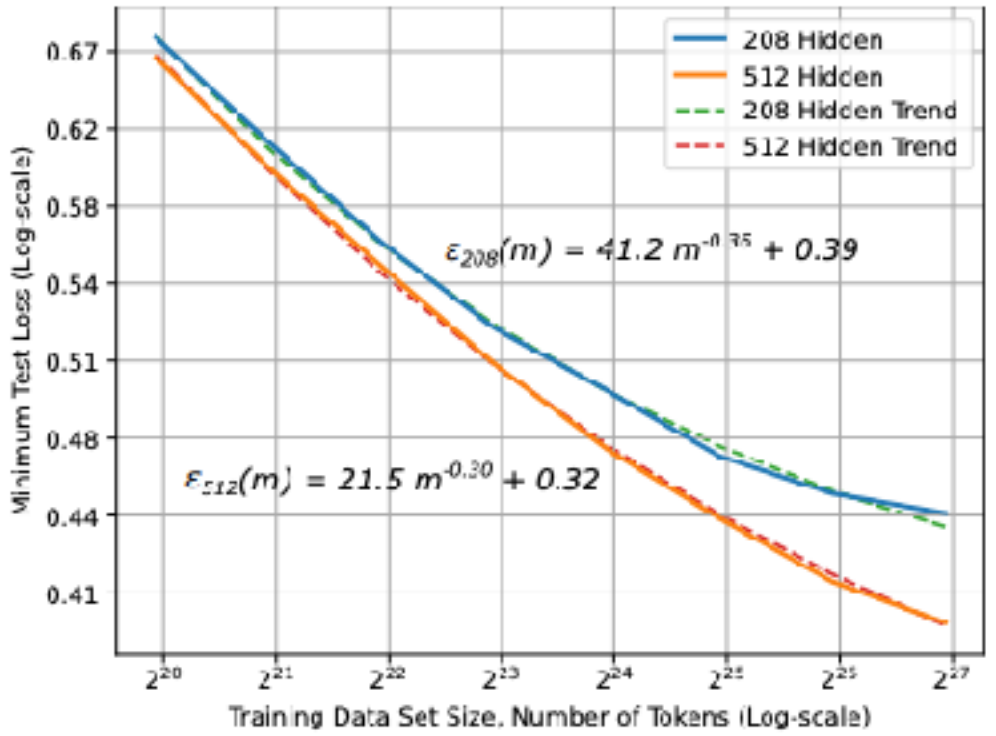
Scaling laws

Classical analog: Estimation error bounds

e.g., fixed-design linear regression:

$$||\theta - \theta^*||_{\Sigma}^2 \leq O\left(\frac{d}{n}\right)$$

In deep learning, we can't prove bounds, so we empirically fit trends to actual data



An early example of a neural "scaling law:" [Hestness et al. 2017] relate # data to minimum test loss for machine translation.

Scaling laws: Motivation

Scaling laws: Motivation

We can use scaling laws to:

Scaling laws: Motivation

We can use scaling laws to:

- Predict the behavior of larger models without training them
  - How big of a model & how much data do I need to achieve a target loss?

Scaling laws: Motivation

We can use scaling laws to:

Predict the behavior of larger models without training them

How big of a model & how much data do I need to achieve a target loss?

Design “compute-optimal” training by balancing terms **[Hoffman et al 2022]**

Given the amount of training data I have, what is the right model size?

Scaling laws: Motivation

We can use scaling laws to:

- Predict the behavior of larger models without training them
  - How big of a model & how much data do I need to achieve a target loss?
- Design “compute-optimal” training by balancing terms **[Hoffman et al 2022]**
  - Given the amount of training data I have, what is the right model size?
- Understand & diagnose bottlenecks to model performance
  - Are we “running out” of data?

Scaling laws: Motivation

We can use scaling laws to:

Predict the behavior of larger models without training them

How big of a model & how much data do I need to achieve a target loss?

Design “compute-optimal” training by balancing terms **[Hoffman et al 2022]**

Given the amount of training data I have, what is the right model size?

Understand & diagnose bottlenecks to model performance

Are we “running out” of data?

Make model selection decisions based on **predicted** behavior

Scaling laws: Methodology

Scaling laws: Methodology

General recipe (most basic version of a scaling law):

Scaling laws: Methodology

General recipe (most basic version of a scaling law):

Pick many (relatively small) values of N, D, C

Scaling laws: Methodology

General recipe (most basic version of a scaling law):

Pick many (relatively small) values of N, D, C

Note: C only comes into play when we do multiple passes over the data

Scaling laws: Methodology

General recipe (most basic version of a scaling law):

Pick many (relatively small) values of N, D, C

Note: C only comes into play when we do multiple passes over the data

Postulate a functional form for ℓ, e.g., for fixed D, suppose $\ell \propto \ell_0 + \left(\frac{D_0}{D}\right)^{\alpha}$

Scaling laws: Methodology

General recipe (most basic version of a scaling law):

Pick many (relatively small) values of N, D, C

Note: C only comes into play when we do multiple passes over the data

Postulate a functional form for ℓ, e.g., for fixed D, suppose $\ell \propto \ell_0 + \left(\frac{D_0}{D}\right)^\alpha$

Fit a power law using basic regression

Scaling laws: Methodology

General recipe (most basic version of a scaling law):

Pick many (relatively small) values of N, D, C

Note: C only comes into play when we do multiple passes over the data

Postulate a functional form for ℓ, e.g., for fixed D, suppose $\ell \propto \ell_0 + \left(\frac{D_0}{D}\right)^{\alpha}$

Fit a power law using basic regression

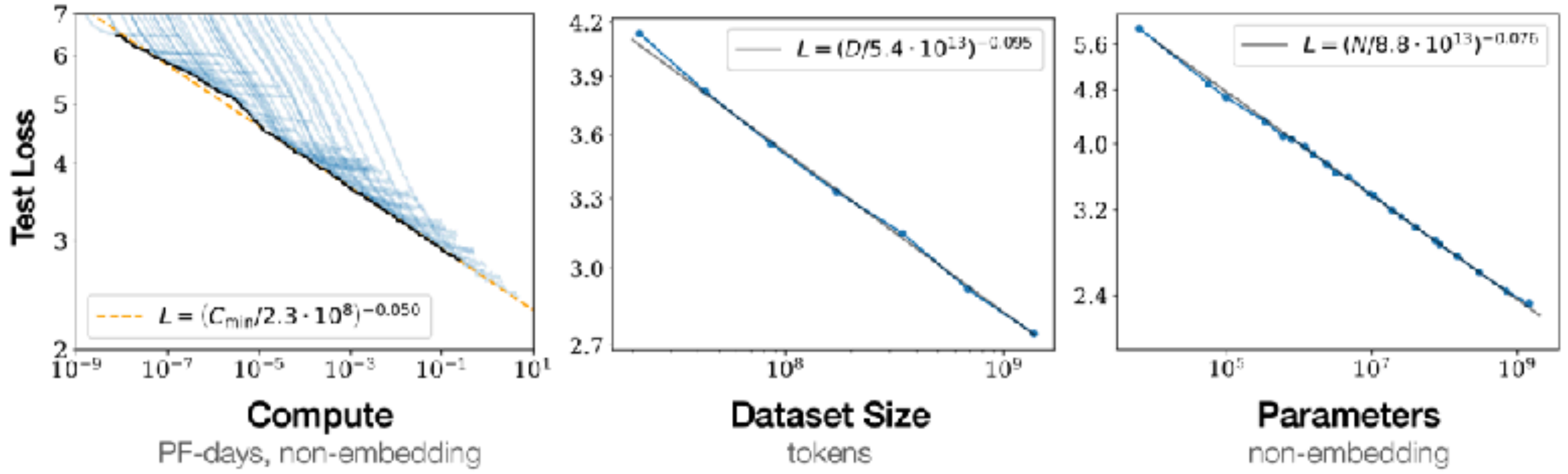


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not

Scaling laws: Methodology II

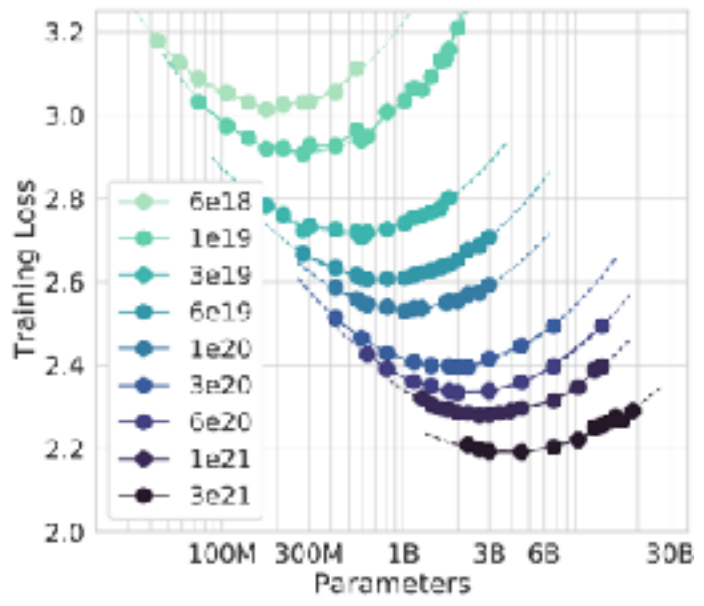
Several immediate limitations (& fixes):

Scaling laws: Methodology II

Several immediate limitations (& fixes):

Q: Can we vary > 1 scaling axes at once?

$$\ell = \left(\frac{N_0}{N}\right)^{\alpha} + \left(\frac{D_0}{D}\right)^{\beta} + \ell_0$$



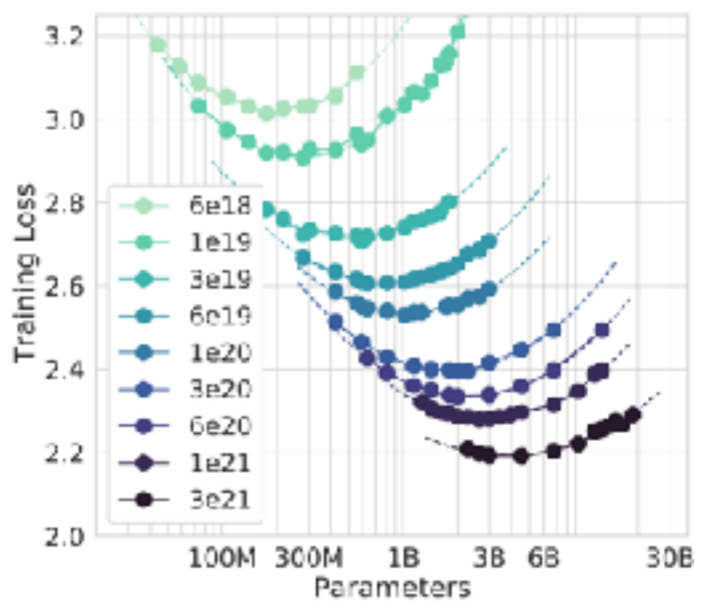
[Hoffman et al 2022]

Scaling laws: Methodology II

Several immediate limitations (& fixes):

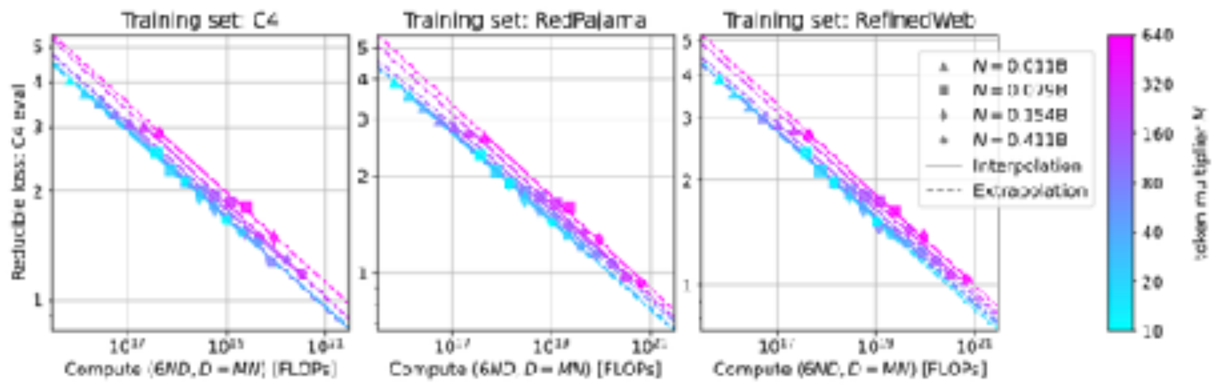
Q: Can we vary > 1 scaling axes at once?

$$\ell = \left(\frac{N_0}{N}\right)^{\alpha} + \left(\frac{D_0}{D}\right)^{\beta} + \ell_0$$



[Hoffman et al 2022]

Q: What if we see each datapoint > once?



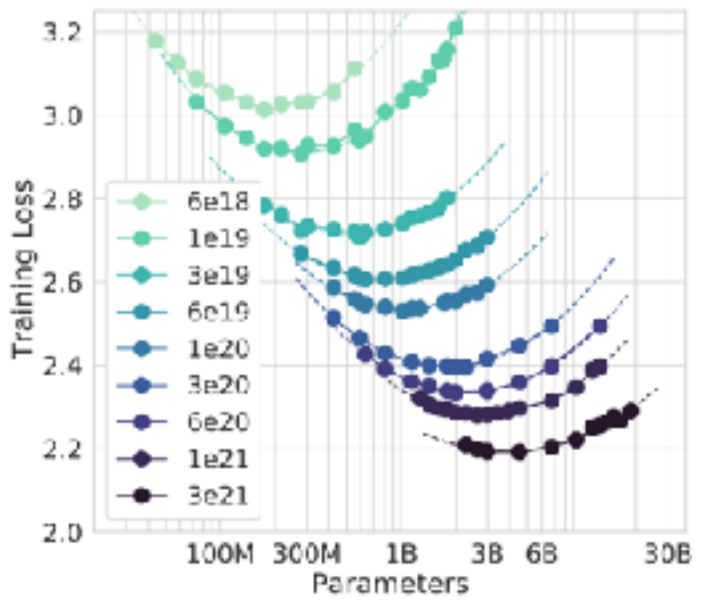
[Muennighoff et al 2021; Gadre et al 2024]

Scaling laws: Methodology II

Several immediate limitations (& fixes):

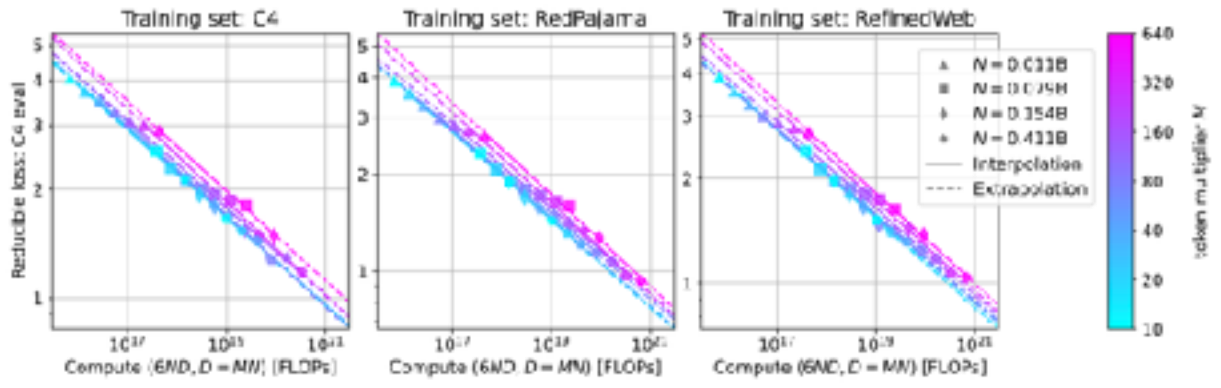
Q: Can we vary > 1 scaling axes at once?

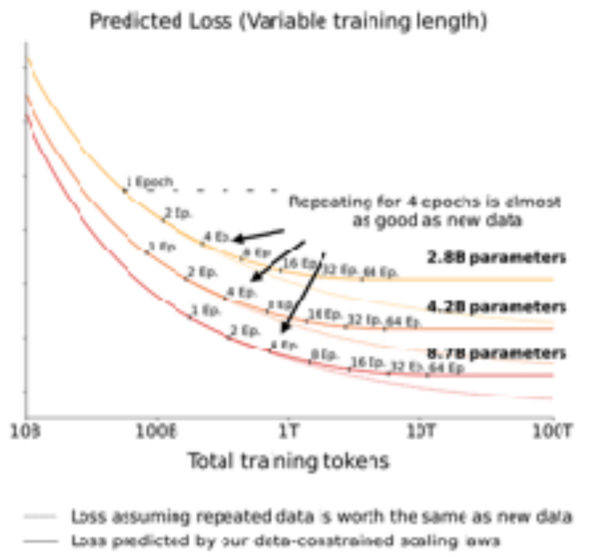
$$\ell = \left(\frac{N_0}{N}\right)^{\alpha} + \left(\frac{D_0}{D}\right)^{\beta} + \ell_0$$



[Hoffman et al 2022]

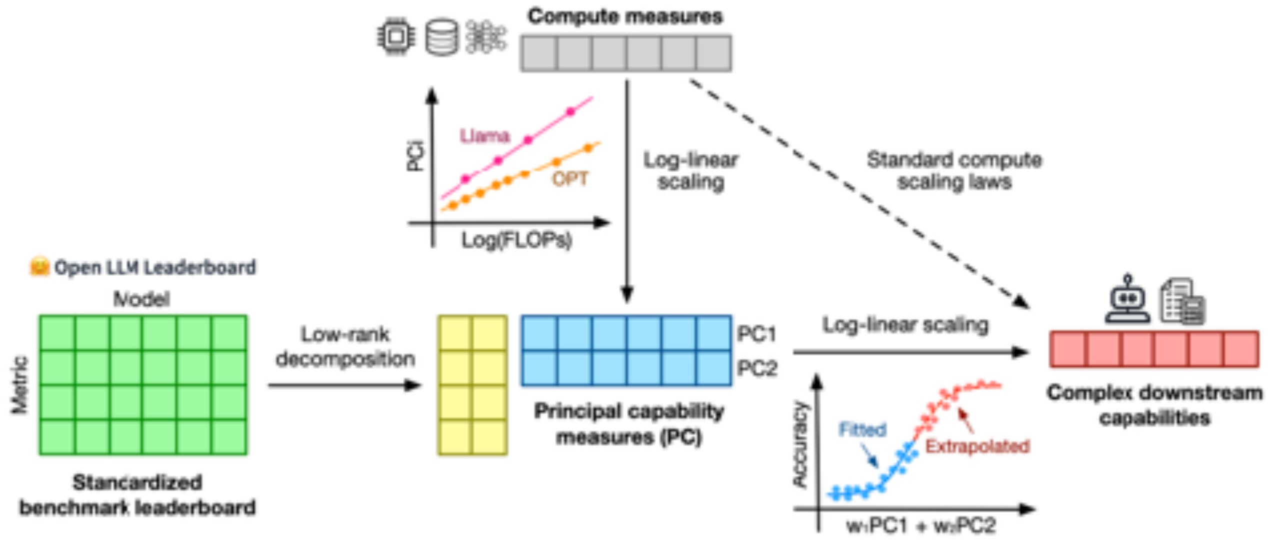
Q: What if we see each datapoint > once?





[Muennighoff et al 2021; Gadre et al 2024]

Q: Do we need to train hundreds of models?



[Ruan Maddison Hashimoto 2024]

Scaling laws: Explanations

Scaling laws: Explanations

Most theoretical explanations given using **random feature** models

Scaling laws: Explanations

Most theoretical explanations given using **random feature** models

Example **[Bahri Dyer Kaplan Lee Sharma 2021]**:

Study two-parameter scaling laws $\ell(N, D)$

Take one of $N, D \to \infty$, study the scaling behavior of the other

In these infinite limits, find similar phenomena to practice—training is sometimes "data-bottlenecked" and sometimes "compute-bottlenecked"

Scaling laws: Explanations

Most theoretical explanations given using **random feature** models

Example **[Bahri Dyer Kaplan Lee Sharma 2021]**:

Study two-parameter scaling laws $\ell(N, D)$

Take one of $N, D \to \infty$, study the scaling behavior of the other

In these infinite limits, find similar phenomena to practice—training is sometimes "data-bottlenecked" and sometimes "compute-bottlenecked"

Many refinements **[Bordelon Atanasov Pehlevan 2024]** and empirical caveats **[Vyas Bansal Nakkiran 2022]**

Data selection/curation/synthesis

Overarching question: how does the *composition* of the data we train on affect the ML models we get, and what interventions can we perform?

Problem setup: Learning algorithm A (mapping dataset → ML model), pool of messy/scraped data S, and a target metric f (mapping ML model → number)

Goal: Dataset D such that $A(D)$ maximizes the target metric f

$$D^* = \max_{D \in \mathcal{Z}^*} f(A(D))$$

General data design

$$D^* = \max_{D \subset S} f(A(D))$$

Data selection/curation

Data selection/curation

Data selection/curation

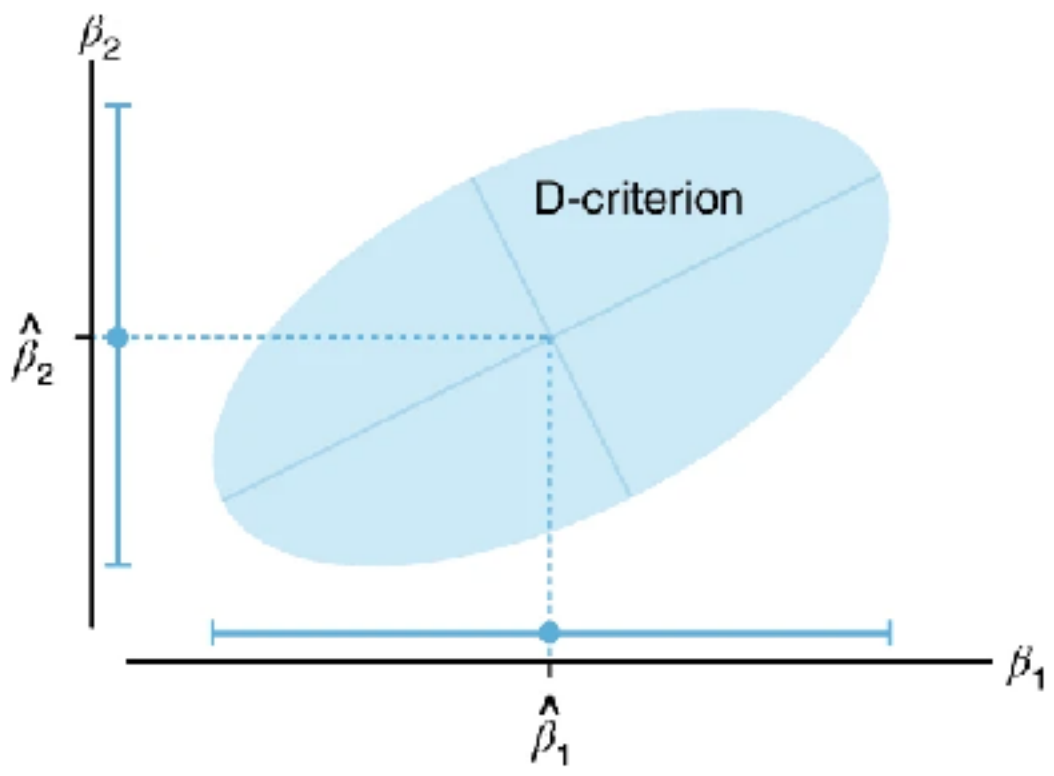
Classical analogs: Optimal experiment design, active learning, sample reweighing (e.g., in causal inference)

Data selection/curation

Classical analogs: Optimal experiment design, active learning, sample reweighing (e.g., in causal inference)

Which data minimizes the size of the resulting CI?

$$\max_X \det(X^\top X)$$



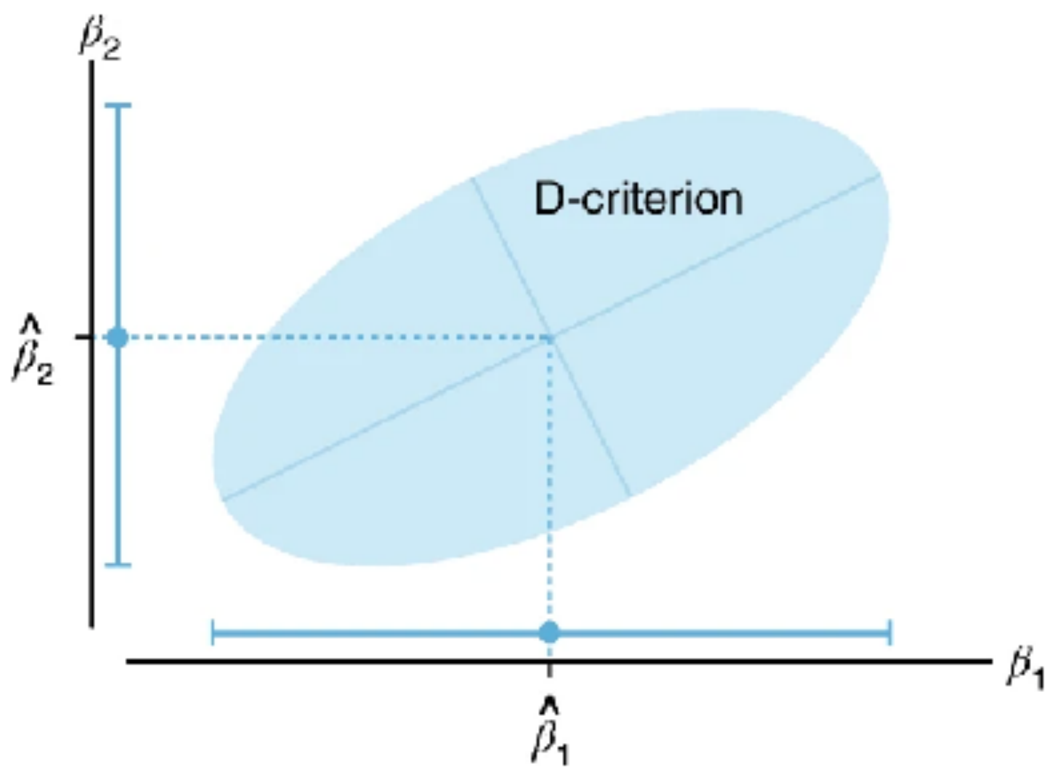
Data selection/curation

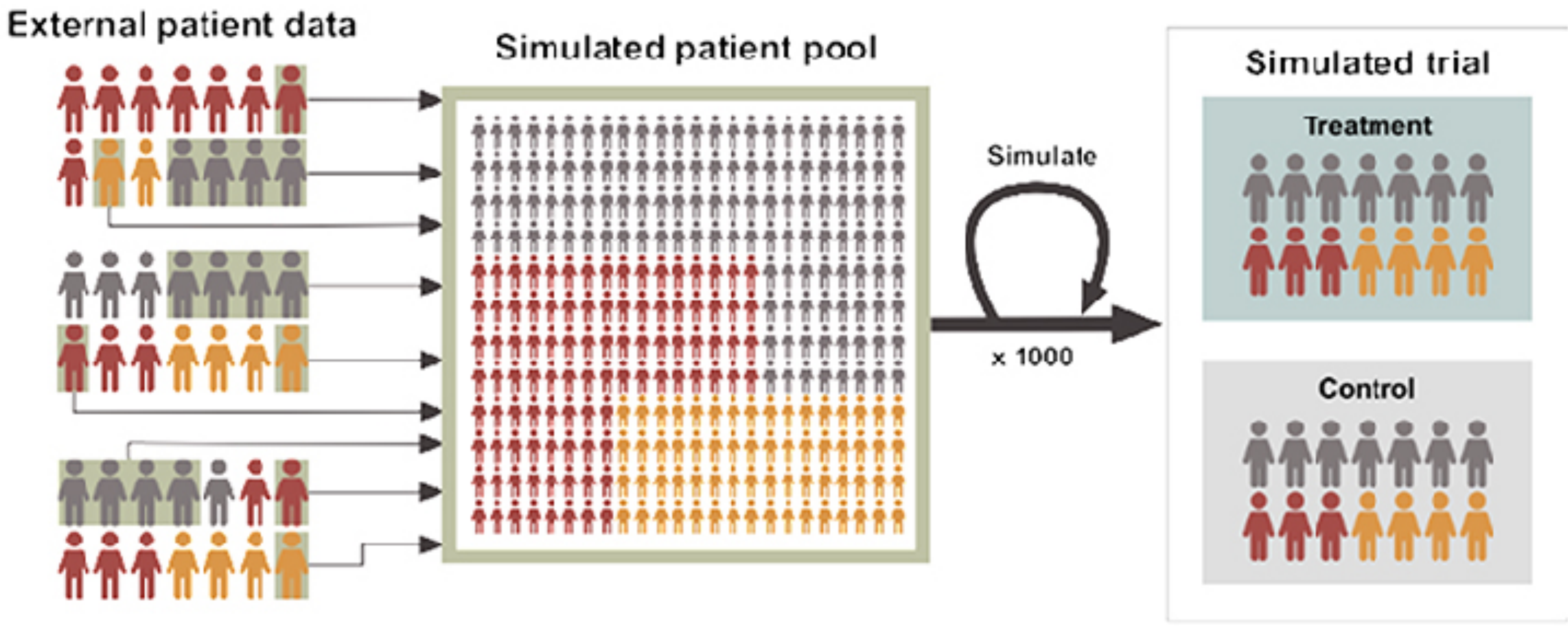
Classical analogs: Optimal experiment design, active learning, sample reweighing (e.g., in causal inference)

How do I combine data to make a valid inference?

Which data minimizes the size of the resulting CI?

$$\max_X \det(X^\top X)$$



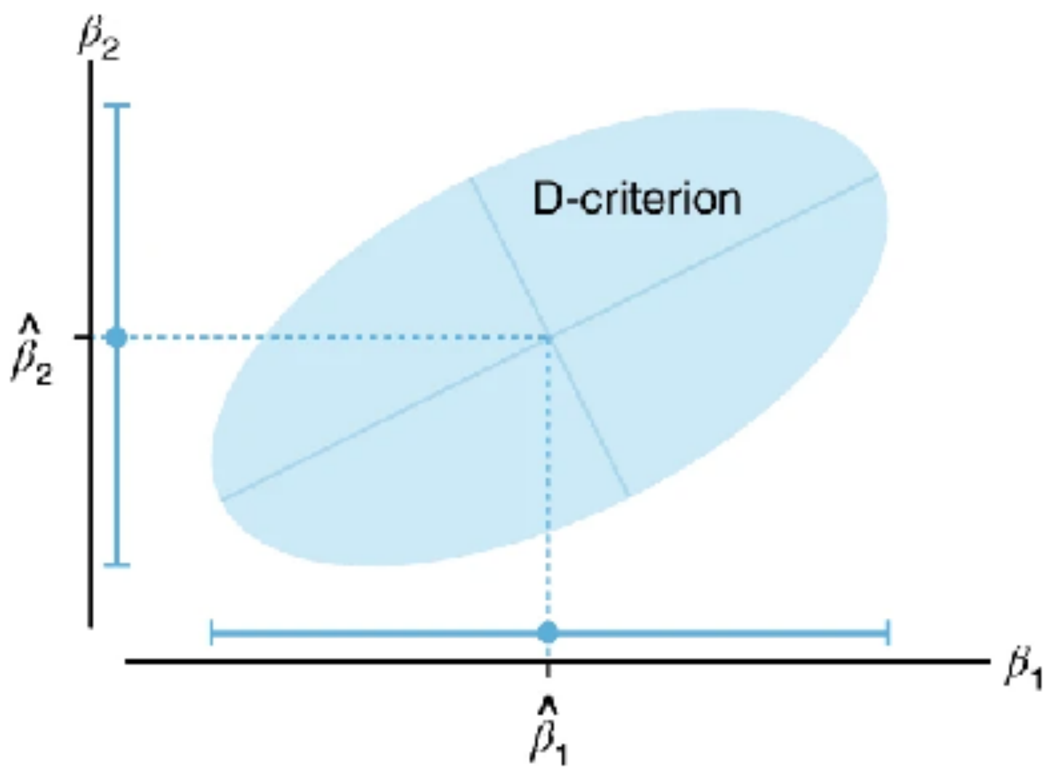


Data selection/curation

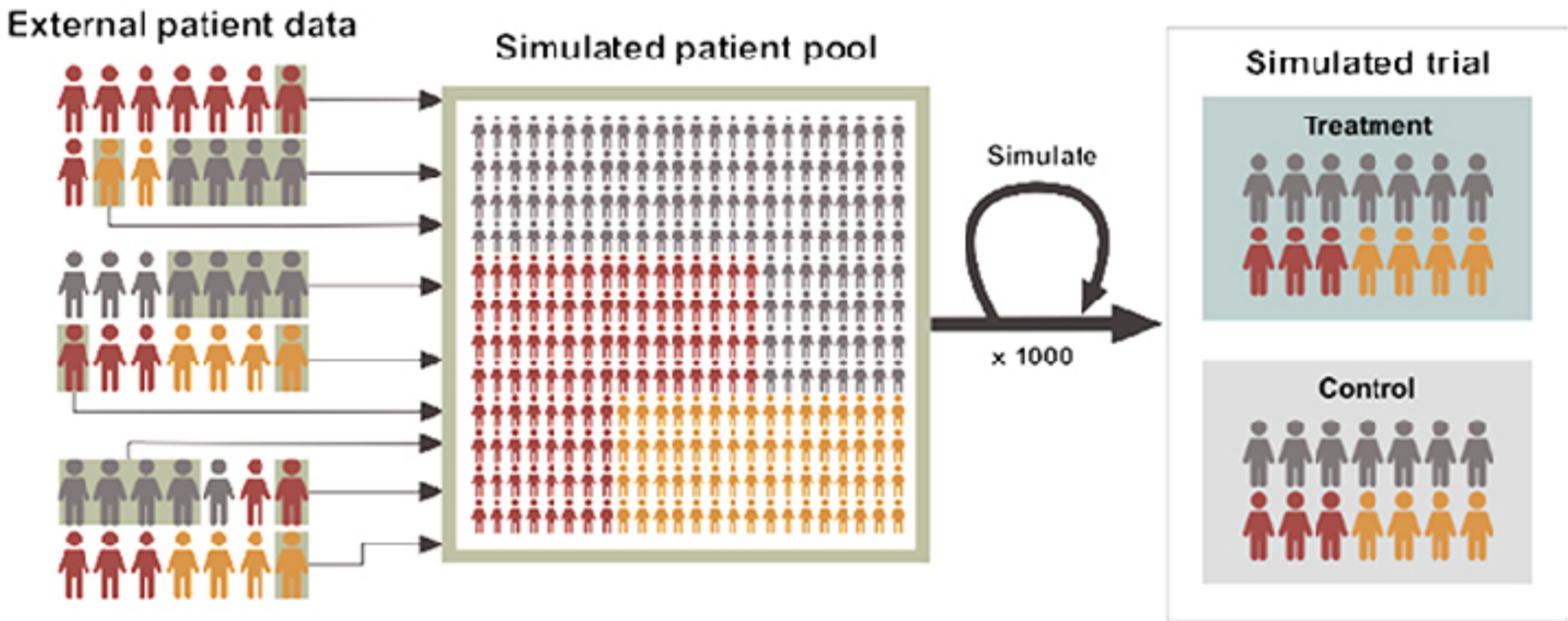
Classical analogs: Optimal experiment design, active learning, sample reweighing (e.g., in causal inference)

Which data minimizes the size of the resulting CI?

$$\max_{X} \det(X^{\top}X)$$



How do I combine data to make a valid inference?



In deep learning, (a) train and test distributions do not match (b) parameters are meaningless (c) data is huge-scale & models are “black-box”

Data selection/curation: Motivation

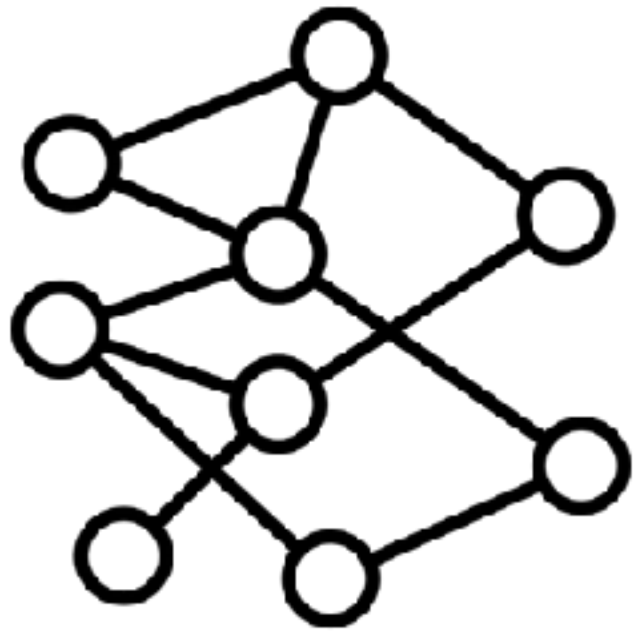
Data selection/curation: Motivation

Scraped internet data

Data selection/curation: Motivation

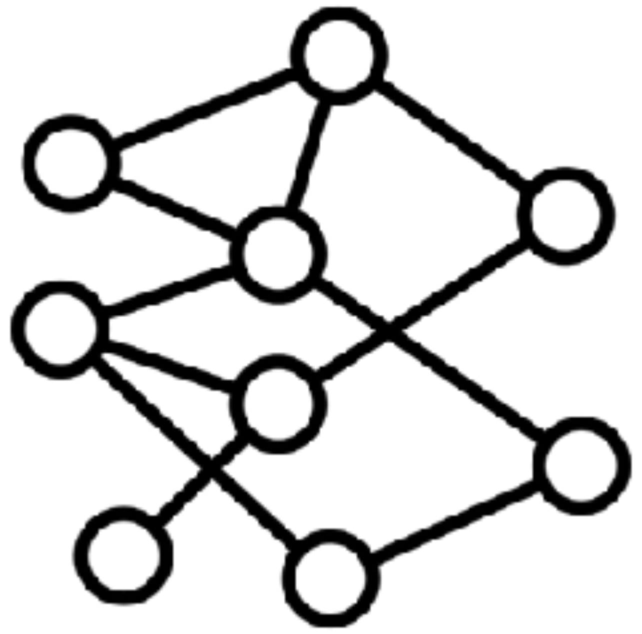
Scraped internet data

+

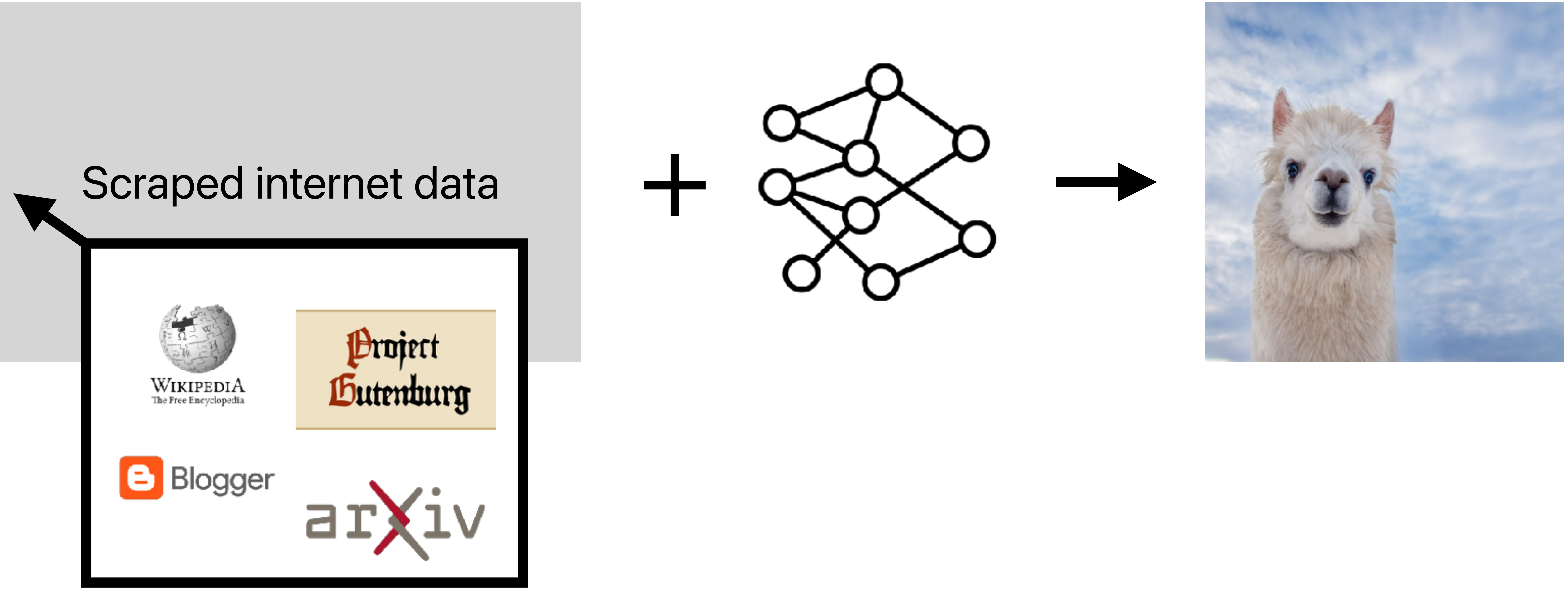


Data selection/curation: Motivation

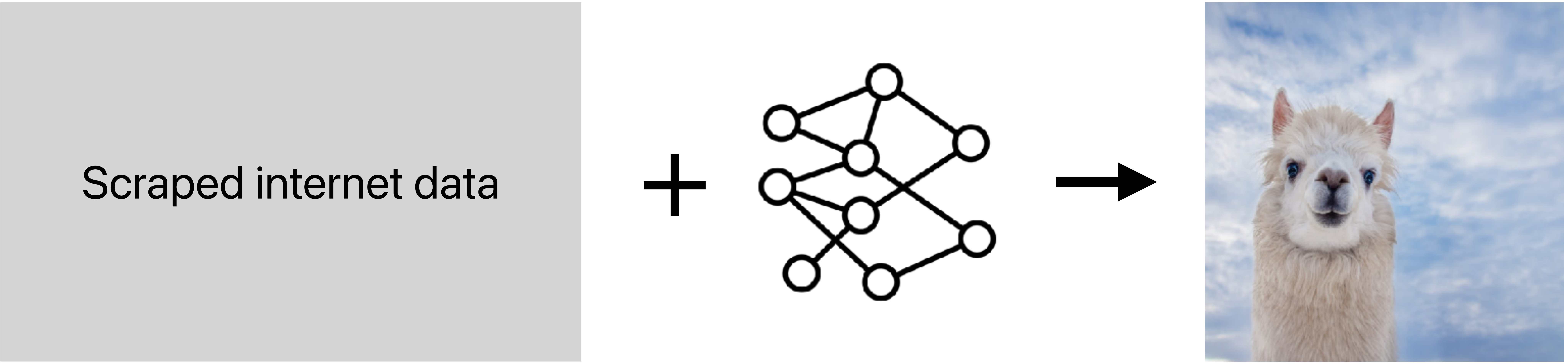
Scraped internet data



Data selection/curation: Motivation



Data selection/curation: Motivation

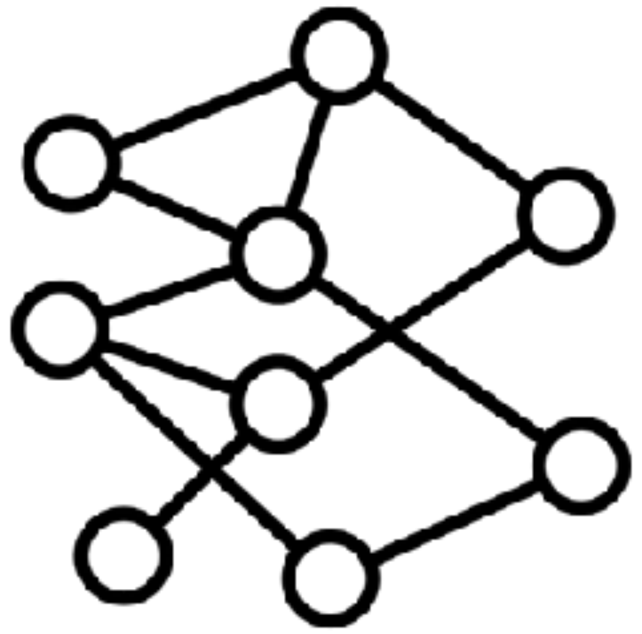


Data selection/curation: Motivation

electroniccigarettereviewed.info
prestigedentalproducts.com
brain-dumps.us

Scraped internet data

+



→

Data selection/cu

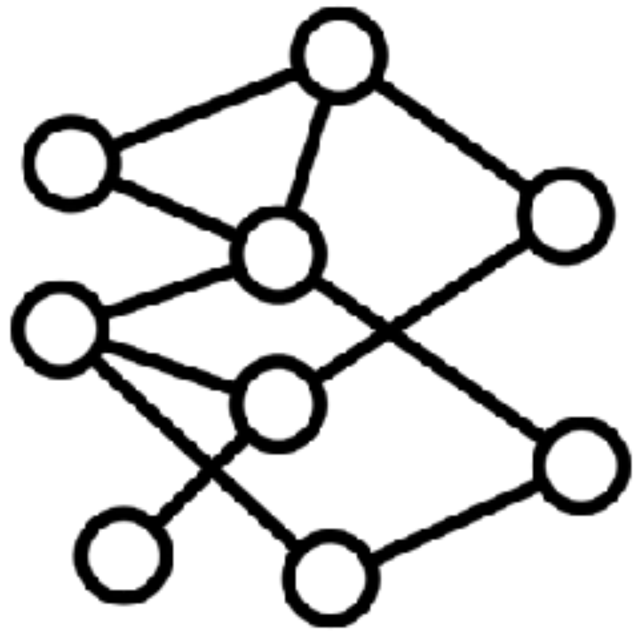
http://ufdc.ufl.edu/AA00010883/00095

ac omplishmnl 'o the mi alon and welR I the men. U (SR 600-17b-I) are clear enough and cover everybody. iM-be UeIrU d at home first. IDretewr.. NatM WLWr Of Panama. B Aleol B -w s sin,ambas to V '). .. &. .... iZ'." . '- .- '.'" ""5- ". ^ -*," ..^ -^:? ^~-. .... ; ,'- i ,? '. " ". 1 the eye to those Europe. like an orlontal: Mrt of the Ch-. te a ter ii wre n capw. to '. %e anti-trust sutt "I made let of money, but . mg C1 Y ay With ne shot of a Lot d Aieles mining and. petrlce Wymore is fac9 aur rtl 1 Dalton. Gr at gag tUdS 't come! Mand Jobaim AIMauccessful- M ." y uw? ie House ar&rkd, "Hold A mtllon yeas? A trillion? t Arrival." Or are they ageless?

electroniccigarettereviewed.info
prestigedentalproducts.com
brain-dumps.us

Scraped internet data

+



→

Data selection/curation: Methods

Data selection/curation: Methods

Taxonomy:

Goal: {targeted, untargeted}

Are we maximizing a target metric, or trying to simulate training?

Granularity: {sources, samples}

Are we combining/weighting datasets or filtering individual samples?

Distribution shift: {biased, unbiased}

Does the test distribution match train?

Data selection/curation: Methods

$$D^* = \max_{D \subset S} f(A(D))$$

Filter S based on a pre-defined "quality" function ϕ

Taxonomy:

Goal: {targeted, **untargeted**}

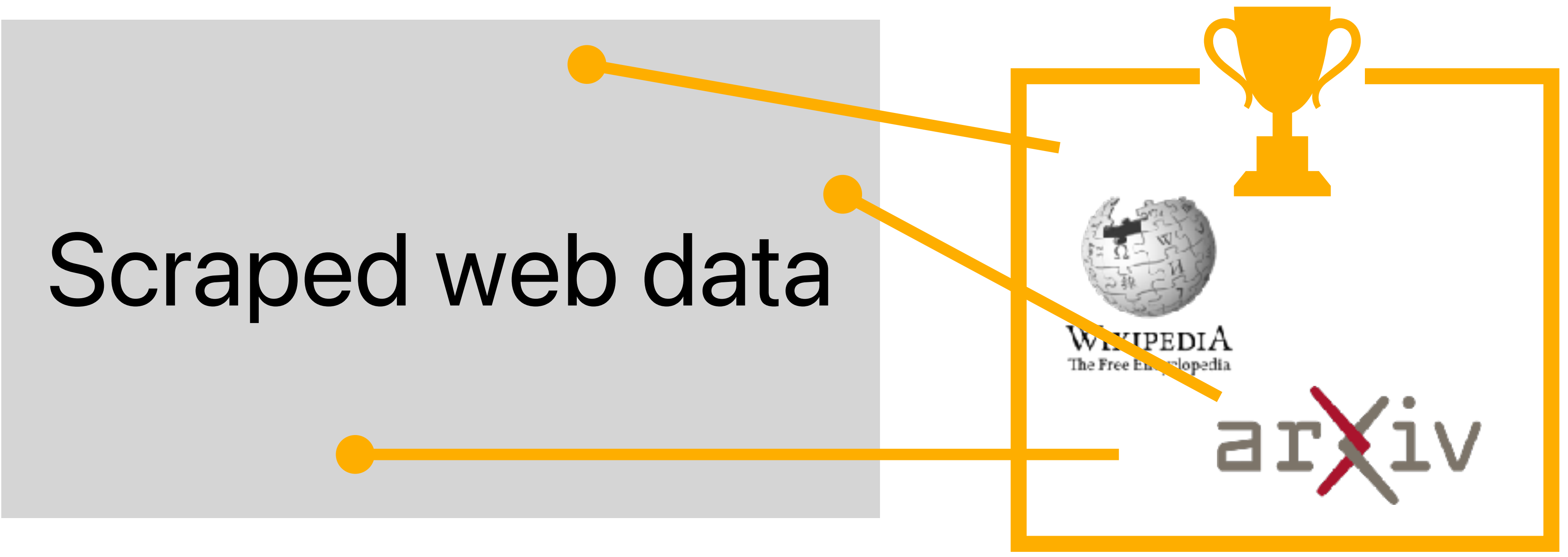
Are we maximizing a target metric, or trying to simulate training?

Granularity: {sources, **samples**}

Are we combining/weighting datasets or filtering individual samples?

Distribution shift: {biased, **unbiased**}

Does the test distribution match train?



(Deduplication, lexical mining, data cleaning...)

Data selection/curation: Methods

Taxonomy:

Goal: {targeted, **untargeted**}

Are we maximizing a target metric, or trying to simulate training?

Granularity: {**sources**, samples}

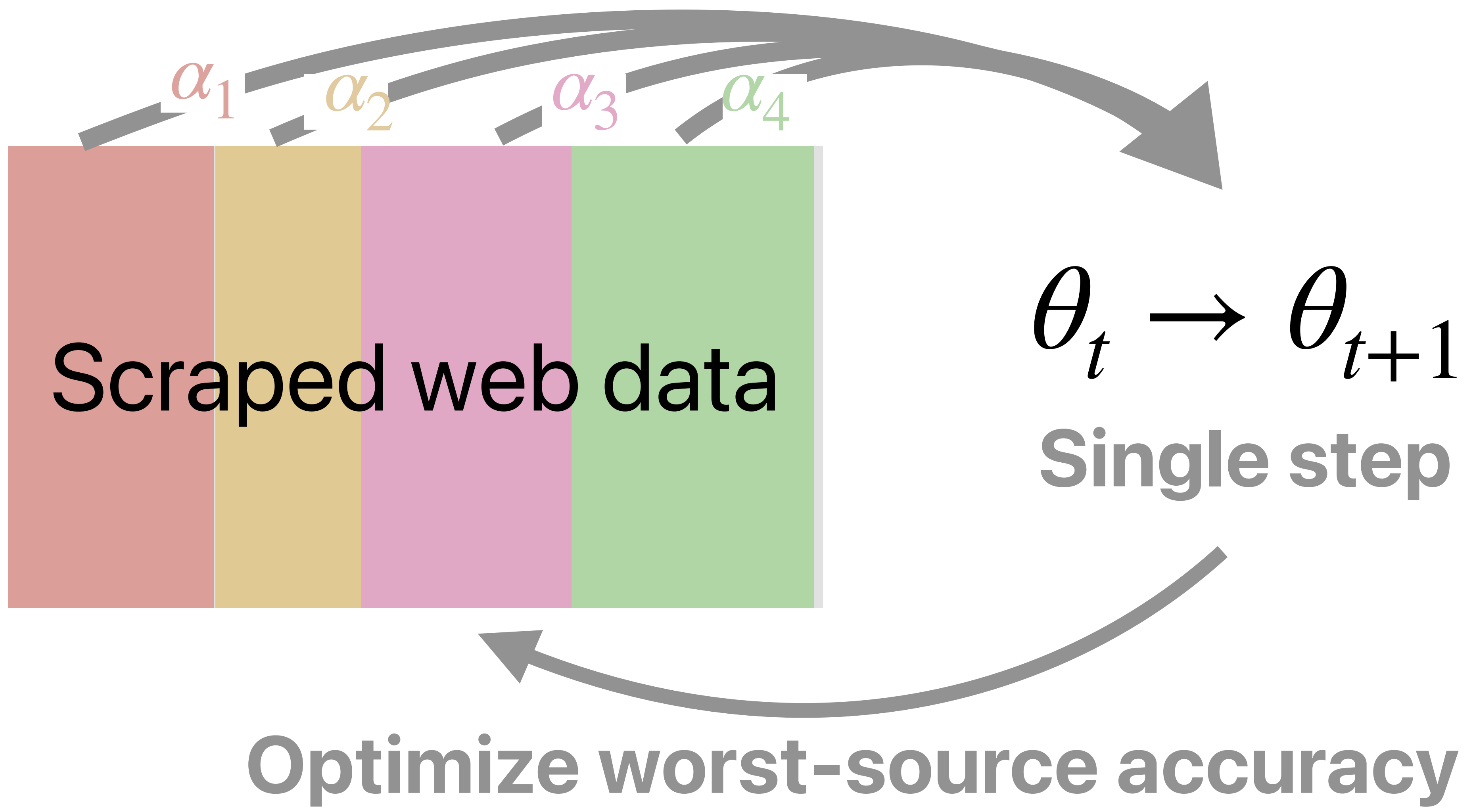
Are we combining/weighting datasets or filtering individual samples?

Distribution shift: {**biased**, unbiased}

Does the test distribution match train?

$$D^* = \max_{D \subset S} f(A(D))$$

Restrict D to mixture of pre-defined sources



(Distributionally robust optimization, bilevel optimization, …)

Data selection/curation: Methods

Taxonomy:

Goal: {targeted, untargeted}

Are we maximizing a target metric, or trying to simulate training?

Granularity: {sources, samples}

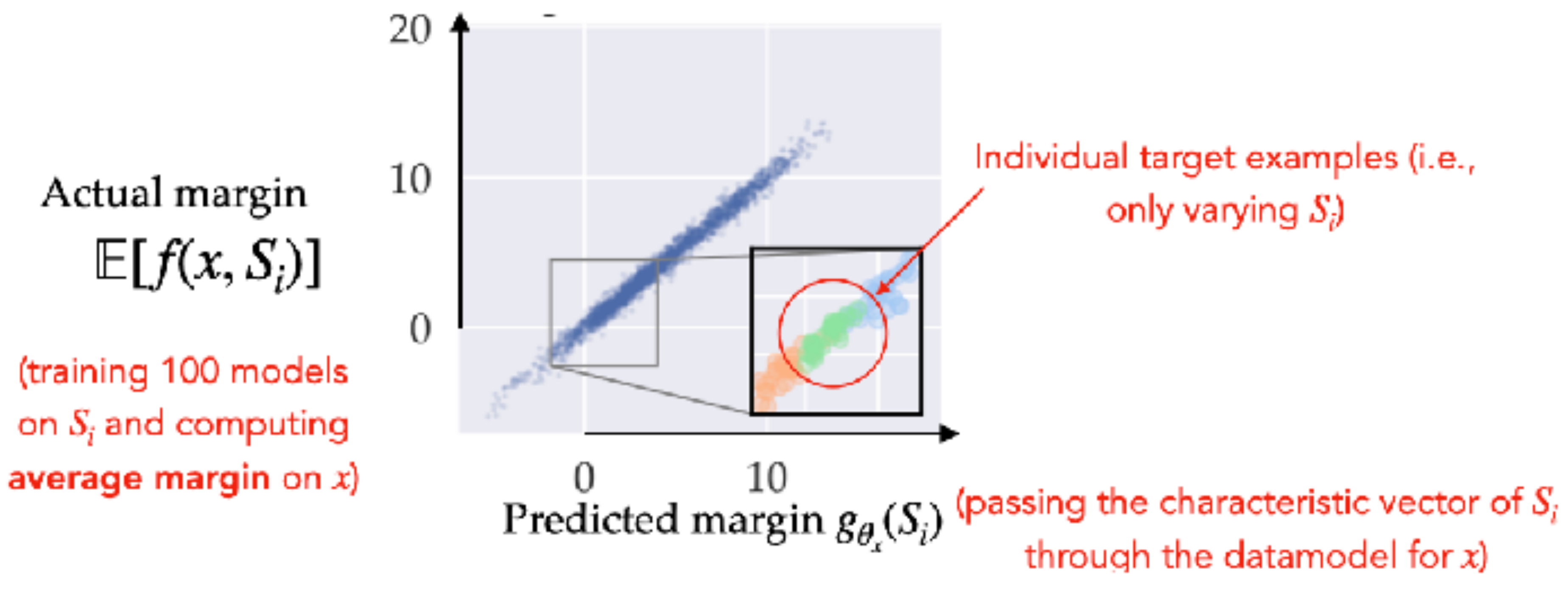
Are we combining/weighting datasets or filtering individual samples?

Distribution shift: {biased, unbiased}

Does the test distribution match train?

$$D^* = \max_{D \subset S} f(A(D))$$

Learn a model $\hat{f}$ from $D \to f(A(D))$ directly, then maximize $\hat{f}$



(Influence-based selection, data valuation, ...)

Data selection/curation: Methods

Taxonomy:

Goal: {targeted, untargeted}

Are we maximizing a target metric, or trying to simulate training?

Granularity: {sources, samples}

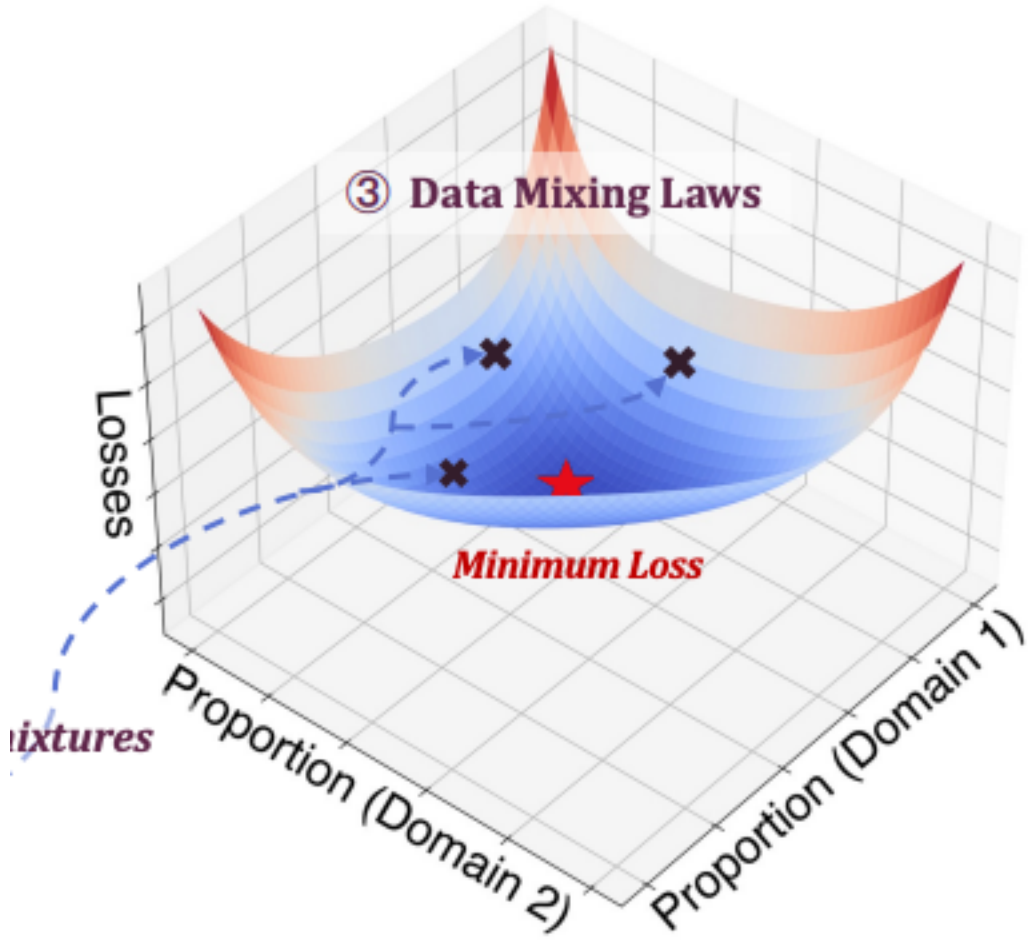
Are we combining/weighting datasets or filtering individual samples?

Distribution shift: {biased, unbiased}

Does the test distribution match train?

$$D* = \max_{D \subset S} f(A(D))$$

Learn or model mixture $\rightarrow f(A(D))$ directly



(Source-specific scaling laws, data mixing laws, ...)

Thank you (and please sign up!)

Sign-up sheet: https://tinyurl.com/reform-ml-signup

Mailing list: **reform-ml-list@stanford.edu**

Contact: **andrewi@stanford.edu**, **saberi@stanford.edu**

Tentative schedule:

1. 10/23 - Scaling laws 1 (Foundations)
2. 10/30 - Scaling laws 2 (Theoretical explanations)
3. 11/6 - Data selection 1 (Optimization-based methods)
4. 11/13 - Data selection 2 (Attribution-based methods)
5. 11/20 - Data selection 3 (Theoretical explanations)
6. 11/27 - Thanksgiving
7. 12/4 - Reserved for an extra lecture on one of the topics (or on another!)