Reproducible Statistical Inference

Jonathan Huggins
Department of Mathematics & Statistics
+ Faculty of Computing & Data Sciences
Boston University

Based on work with Jeff Miller and Jiawei Li

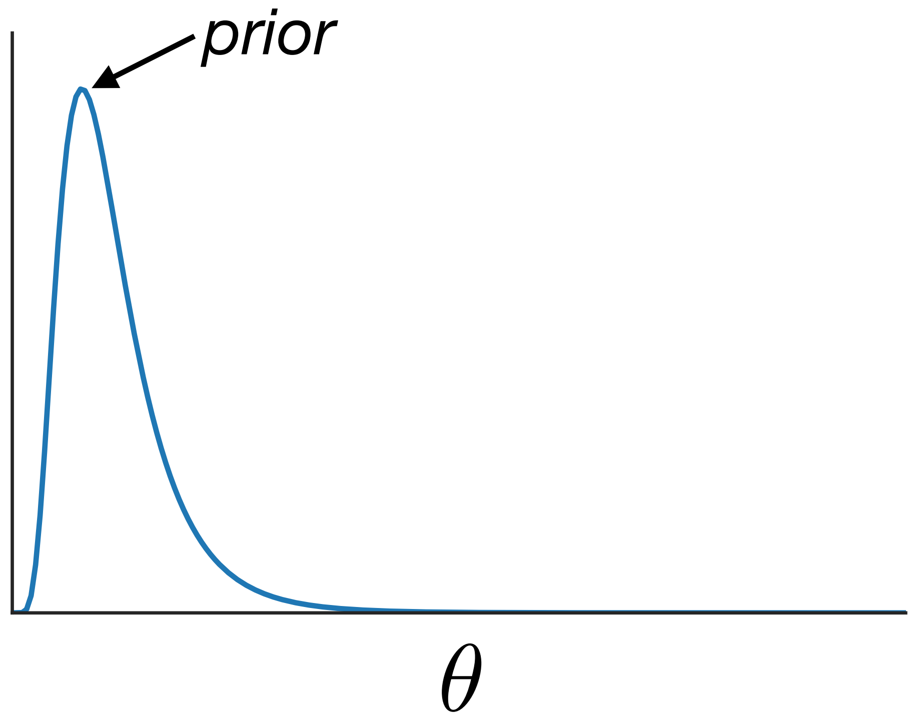
Bayesian inference

Bayesian inference

- **Goal:** learn about unobserved phenomenon (parameter) of interest θ **[e.g., "skill" of a baseball player]**

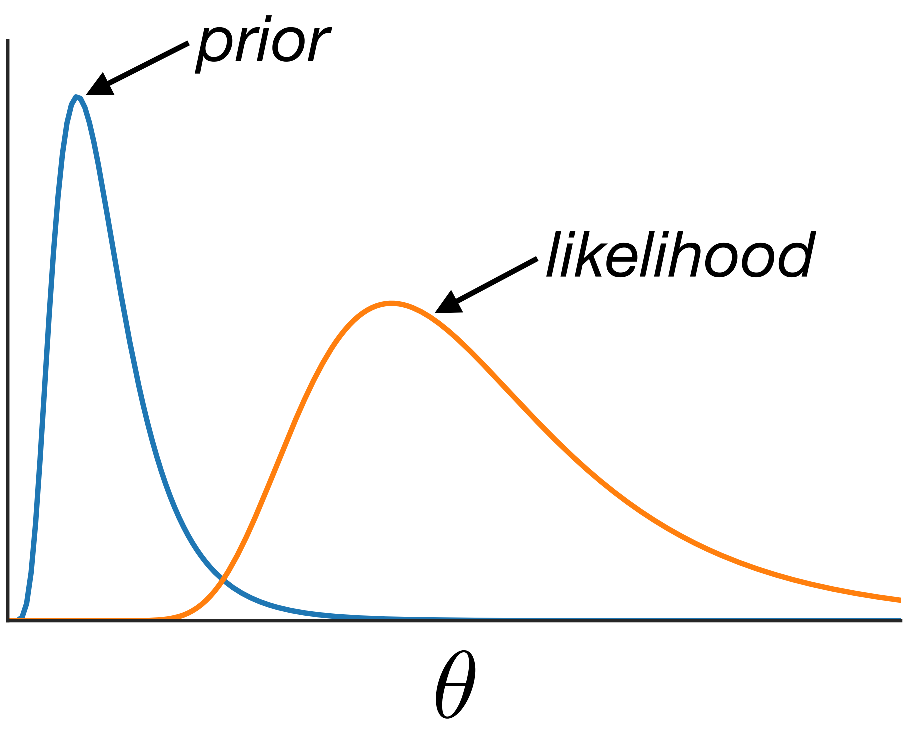
Bayesian inference

- **Goal:** learn about unobserved phenomenon (parameter) of interest θ **[e.g., "skill" of a baseball player]**
- **Prior** beliefs $\pi_0(\theta)$ about the phenomenon



Bayesian inference

- **Goal:** learn about unobserved phenomenon (parameter) of interest θ **[e.g., "skill" of a baseball player]**
- **Prior** beliefs $\pi_0(\theta)$ about the phenomenon
- Observe data x **[e.g., player's performance in some games]**



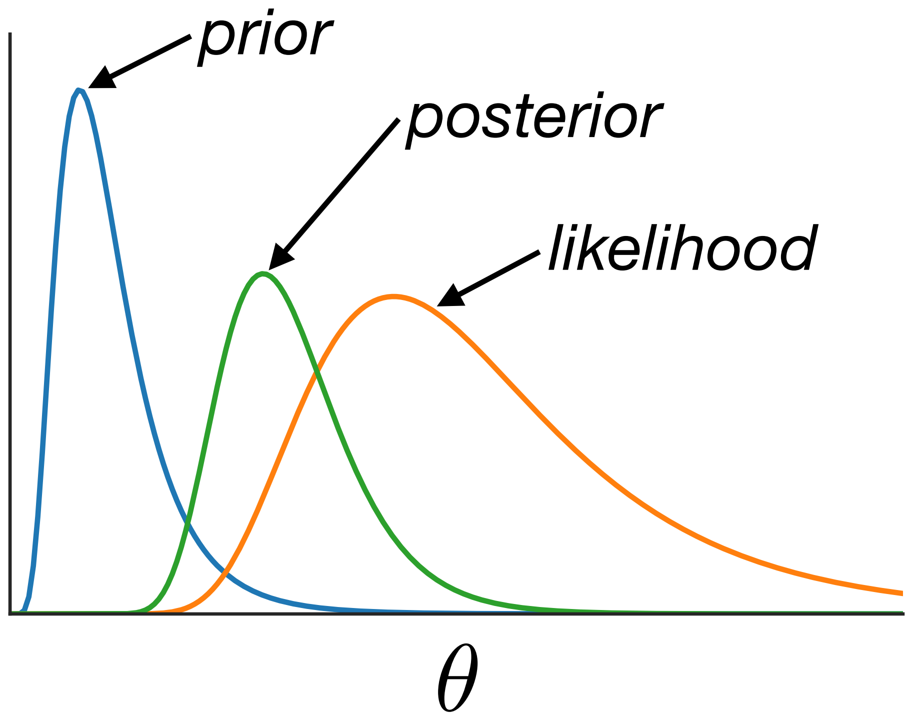
Bayesian inference

- **Goal:** learn about unobserved phenomenon (parameter) of interest θ **[e.g., “skill” of a baseball player]**

- **Prior** beliefs $\pi_0(\theta)$ about the phenomenon

- Observe data x **[e.g., player’s performance in some games]**

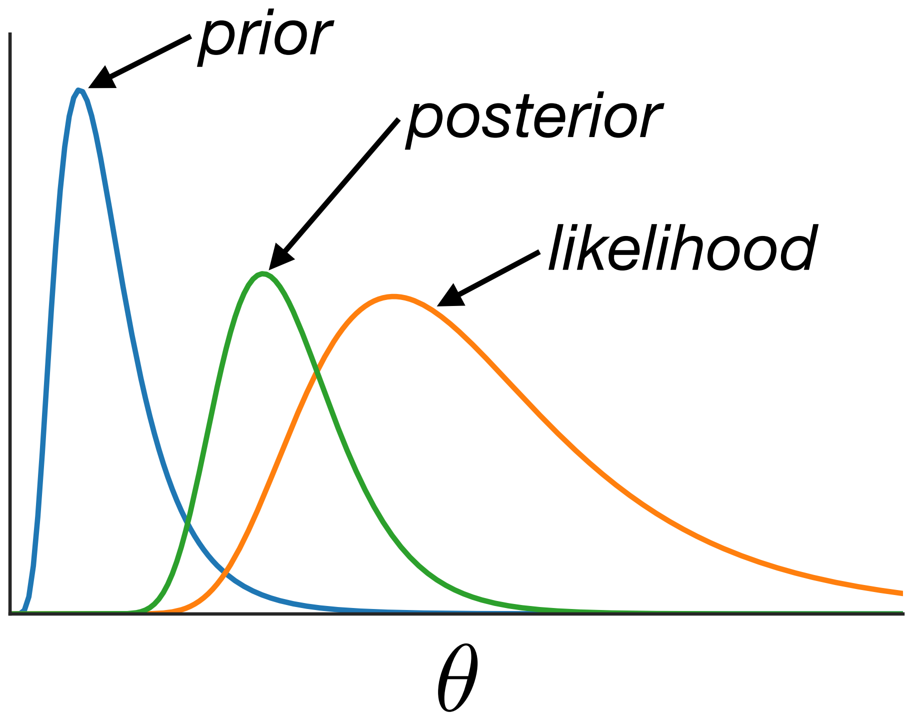
- Assume a **probabilistic model** $p(x \mid \theta)$



Bayesian inference

- **Goal:** learn about unobserved phenomenon (parameter) of interest θ **[e.g., "skill" of a baseball player]**
- **Prior** beliefs $\pi_0(\theta)$ about the phenomenon
- Observe data x **[e.g., player's performance in some games]**
- Assume a **probabilistic model** $p(x \mid \theta)$
- Combine prior & likelihood to form **posterior**:

$$\pi(\theta \mid x) = \frac{p(x \mid \theta)\pi_0(\theta)}{p(x)}$$



Bayesian inference

- **Goal:** learn about unobserved phenomenon (parameter) of interest θ **[e.g., “skill” of a baseball player]**

- **Prior** beliefs $\pi_0(\theta)$ about the phenomenon

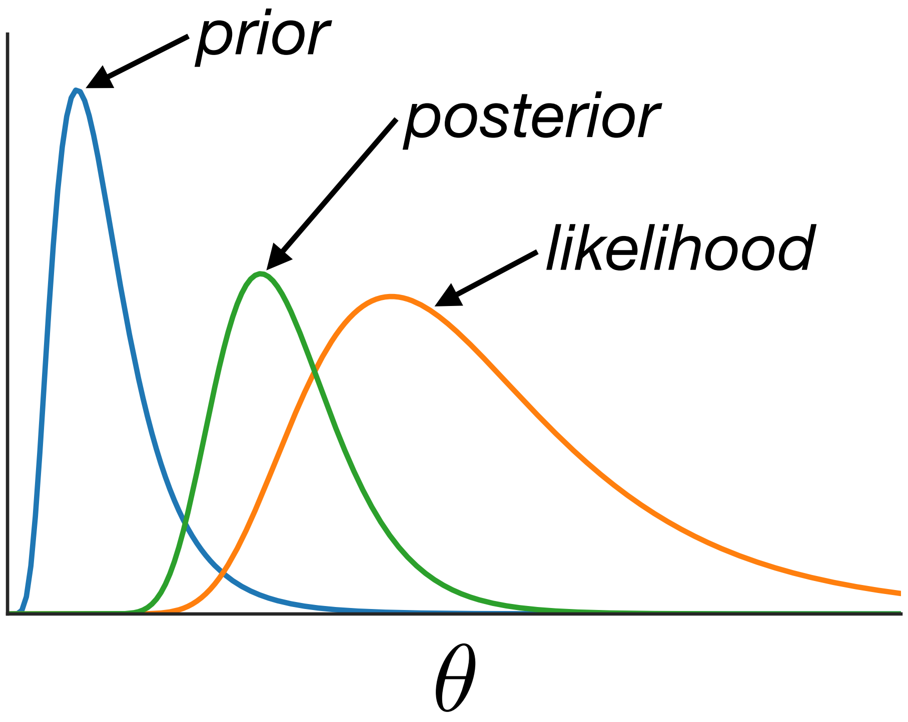
- Observe data x **[e.g., player’s performance in some games]**

- Assume a **probabilistic model** $p(x \mid \theta)$

- Combine prior & likelihood to form **posterior**:

$$\pi(\theta \mid x) = \frac{p(x \mid \theta)\pi_0(\theta)}{p(x)}$$

- **Benefits:** coherent belief updates, uncertainty quantification, flexible modeling, and more



Bayesian inference

- **Goal:** learn about unobserved phenomenon (parameter) of interest θ **[e.g., "skill" of a baseball player]**

- **Prior** beliefs $\pi_0(\theta)$ about the phenomenon

- Observe data x **[e.g., player's performance in some games]**

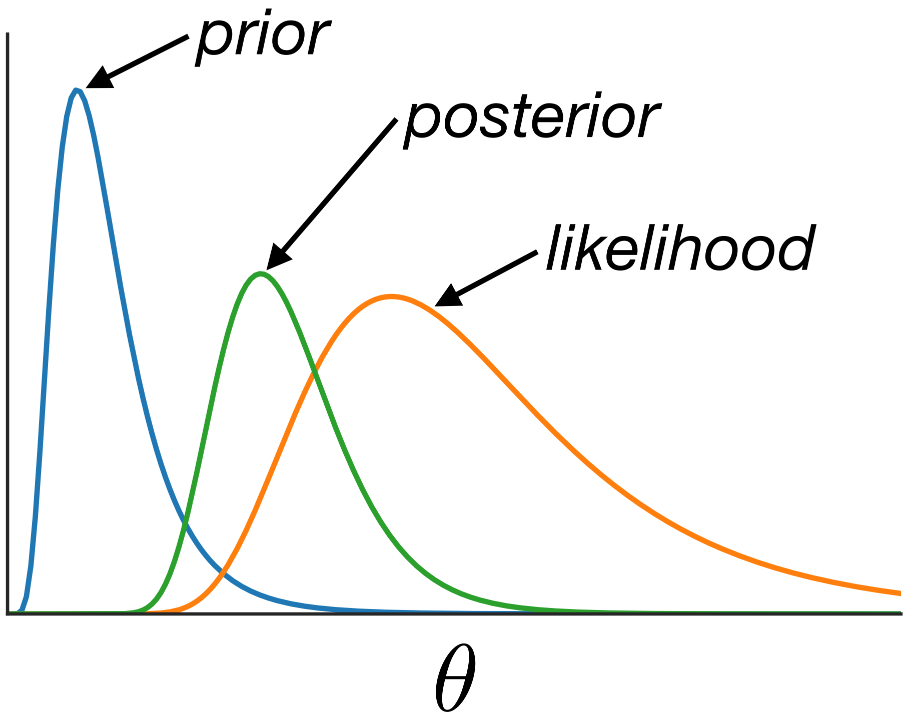
- Assume a **probabilistic model** $p(x \mid \theta)$

- Combine prior & likelihood to form **posterior**:

$$\pi(\theta \mid x) = \frac{p(x \mid \theta)\pi_0(\theta)}{p(x)}$$

- **Benefits:** coherent belief updates, uncertainty quantification, flexible modeling, and more

- **Assumption:** measurement model correct: *observed* x has distribution $p(x \mid \theta_{\text{true}})$



Motivating example: Bayesian model selection

Motivating example: Bayesian model selection

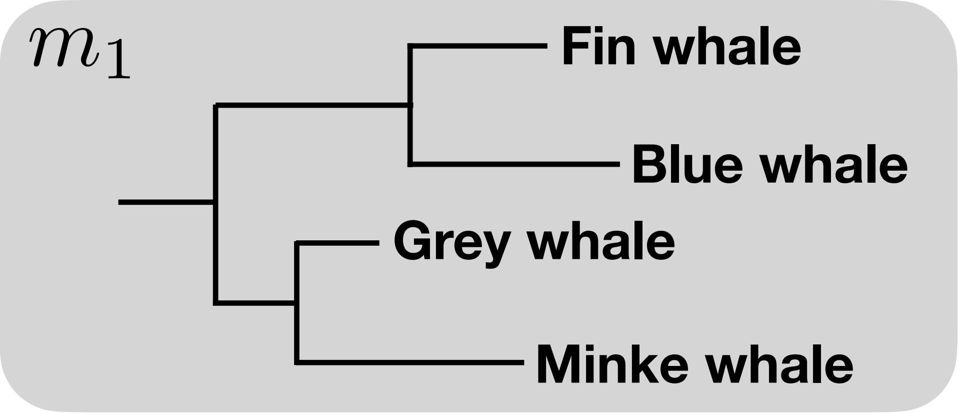
- Given data x, select between a (finite or countable) set of models $m_1, m_2, \ldots$

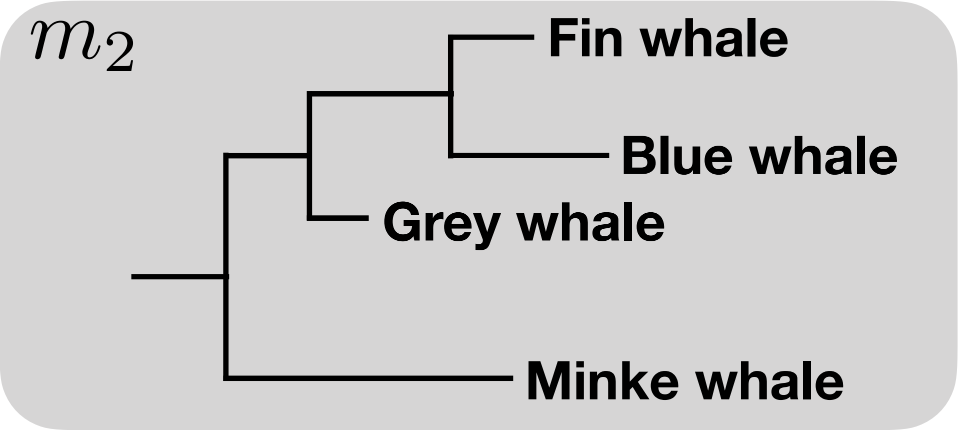
Motivating example: Bayesian model selection

- Given data x, select between a (finite or countable) set of models $m_1, m_2, \ldots$
- **Running example:** systematics

Motivating example: Bayesian model selection

- Given data x, select between a (finite or countable) set of models $m_1, m_2, \ldots$
- **Running example:** systematics
  - **Goal:** learn about evolutionary history of a group of species **[e.g., whales]**





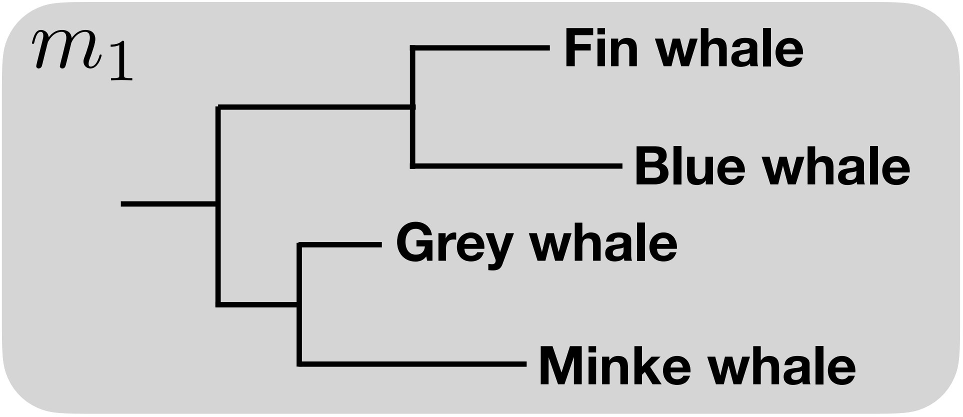
⋮

Motivating example: Bayesian model selection

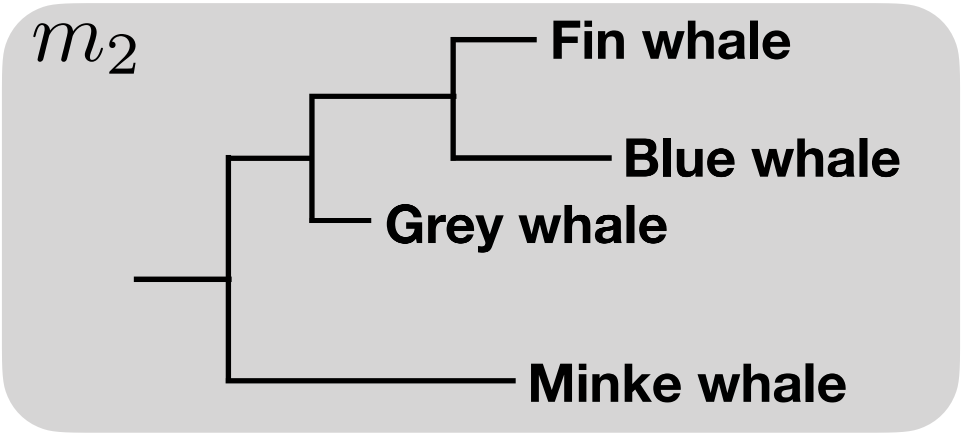
- Given data x, select between a (finite or countable) set of models $m_1, m_2, \ldots$
- **Running example:** systematics
  - **Goal:** learn about evolutionary history of a group of species **[e.g., whales]**
  - **Approach:** infer which phylogenetic trees are consistent with observed species characteristics x **[e.g., genetic data, physical features such as coloring and size]**

$$\pi(m_i \mid x) = \frac{p(x \mid m_i)\pi_0(m_i)}{\sum_j p(x \mid m_j)\pi_0(m_j)}$$

$\pi(m_1 \mid x) = 0.8$



$\pi(m_2 \mid x) = 0.1$



$\vdots$

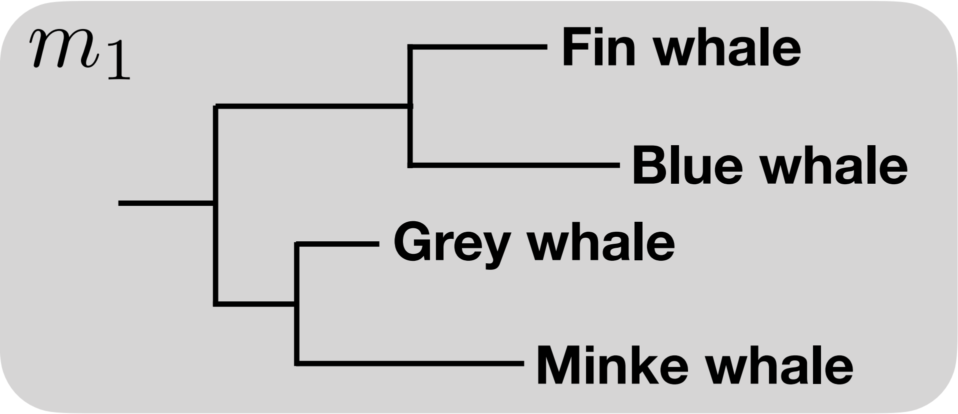
Motivating example: Bayesian model selection

- Given data x, select between a (finite or countable) set of models $m_1, m_2, \ldots$
- **Running example:** systematics
  - **Goal:** learn about evolutionary history of a group of species **[e.g., whales]**
  - **Approach:** infer which phylogenetic trees are consistent with observed species characteristics x **[e.g., genetic data, physical features such as coloring and size]**

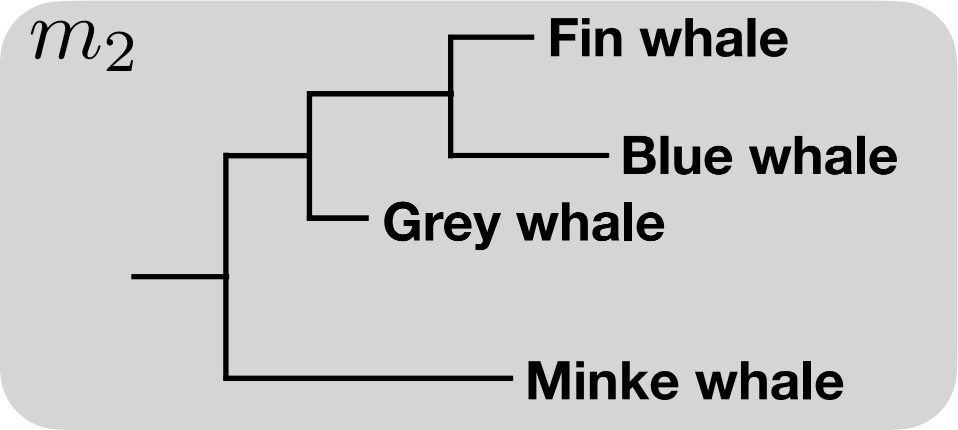
$$\pi(m_i \mid x) = \frac{p(x \mid m_i)\pi_0(m_i)}{\sum_j p(x \mid m_j)\pi_0(m_j)}$$

- **Problem:** (Bayesian) model selection doesn't always work as we might hope…

$\pi(m_1 \mid x) = 0.8$



$\pi(m_2 \mid x) = 0.1$



$\vdots$

Reproducibility in model selection

[H & Miller 2023]

Reproducibility in model selection

- **Problem:** infer phylogeny of 13 whale species from mitochondrial coding DNA

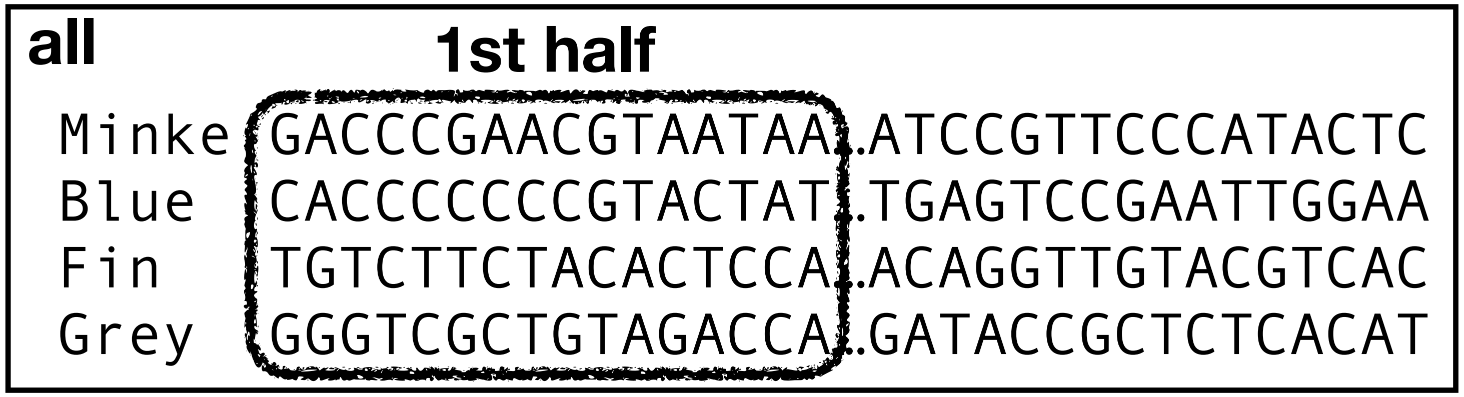
all

```
Minke GACCCGAACGTAATAA…ATCCGTTCCCATACTC
Blue  CACCCCCCCGTACTAT…TGAGTCCGAATTGGAA
Fin   TGTCTTCTACACTCCA…ACAGGTTGTACGTCAC
Grey  GGGTCGCTGTAGACCA…GATACCGCTCTCACAT
```

[H & Miller 2023]

Reproducibility in model selection

- **Problem:** infer phylogeny of 13 whale species from mitochondrial coding DNA

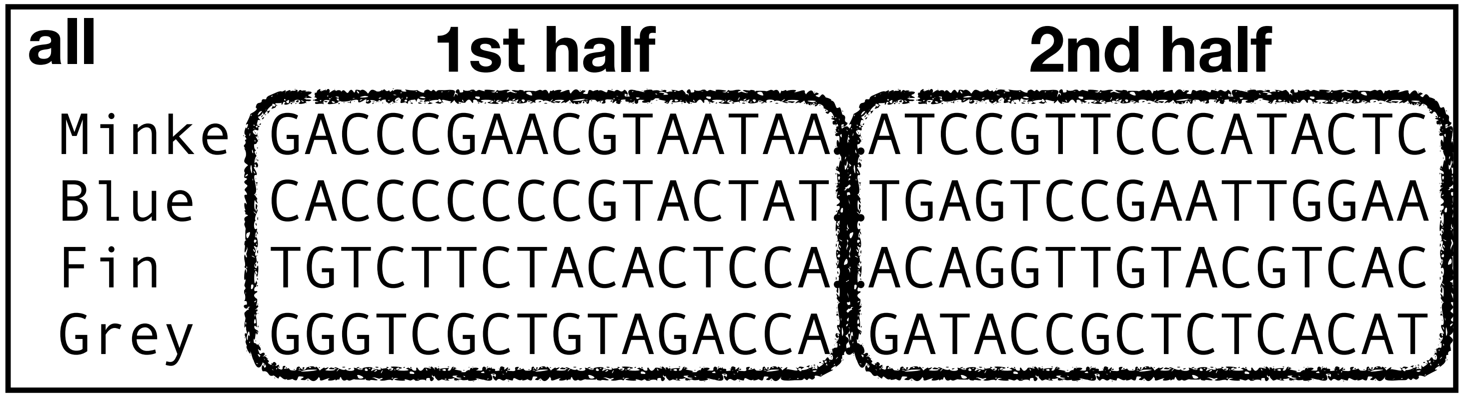


```
Minke GACCCGAACGTAATAA...ATCCGTTCCCATACTC
Blue  CACCCCCCCGTACTAT...TGAGTCCGAATTGGAA
Fin   TGTCTTCTACACTCCA...ACAGGTTGTACGTCAC
Grey  GGGTCGCTGTAGACCA...GATACCGCTCTCACAT
```

[H & Miller 2023]

Reproducibility in model selection

- **Problem:** infer phylogeny of 13 whale species from mitochondrial coding DNA

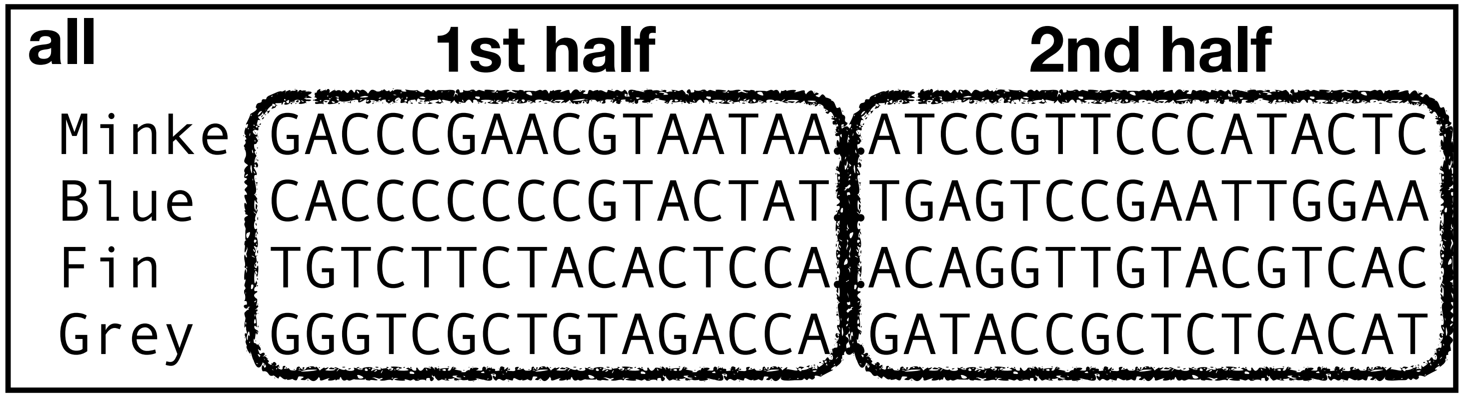


all	1st half	2nd half
Minke	GACCCGAACGTAATAA	ATCCGTTCCCATACTC
Blue	CACCCCCCCGTACTAT	TGAGTCCGAATTGGAA
Fin	TGTCTTCTACACTCCA	ACAGGTTGTACGTCAC
Grey	GGGTCGCTGTAGACCA	GATACCGCTCTCACAT

[H & Miller 2023]

Reproducibility in model selection

- **Problem:** infer phylogeny of 13 whale species from mitochondrial coding DNA
- Compute posterior tree probabilities based on **all**, **1st half**, and **2nd half**



all **1st half** **2nd half**

```
Minke GACCCGAACGTAATAA ATCCGTTCCCATACTC
Blue  CACCCCCCCGTACTAT TGAGTCCGAATTGGAA
Fin   TGTCTTCTACACTCCA ACAGGTTGTACGTCAC
Grey  GGGTCGCTGTAGACCA GATACCGCTCTCACAT
```

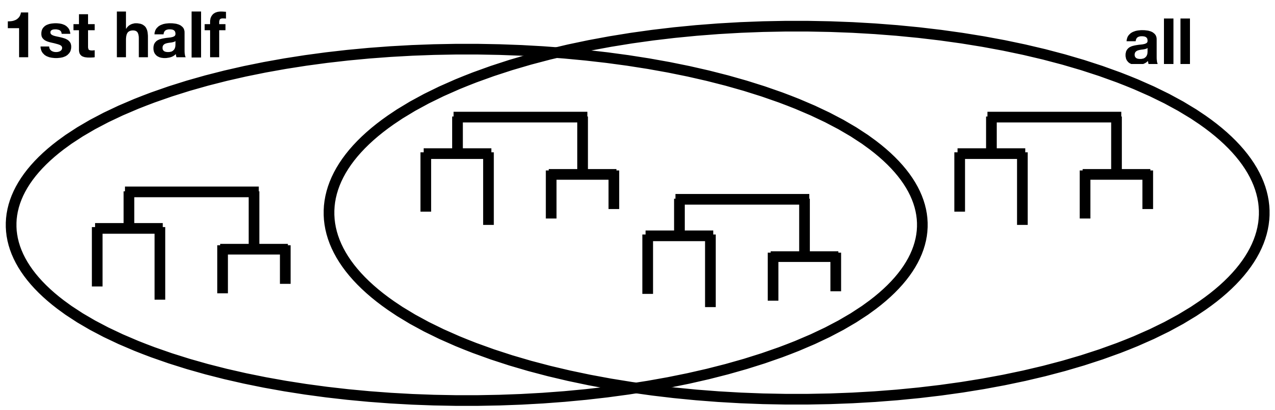
[H & Miller 2023]

Reproducibility in model selection

- **Problem:** infer phylogeny of 13 whale species from mitochondrial coding DNA
- Compute posterior tree probabilities based on **all**, **1st half**, and **2nd half**
- Compute overlap of 99% high probability regions

all **1st half** **2nd half**

```
Minke GACCCGAACGTAATAA ATCCGTTCCCATACTC
Blue  CACCCCCCCGTACTAT TGAGTCCGAATTGGAA
Fin   TGTCTTCTACACTCCA ACAGGTTGTACGTCAC
Grey  GGGTCGCTGTAGACCA GATACCGCTCTCACAT
```

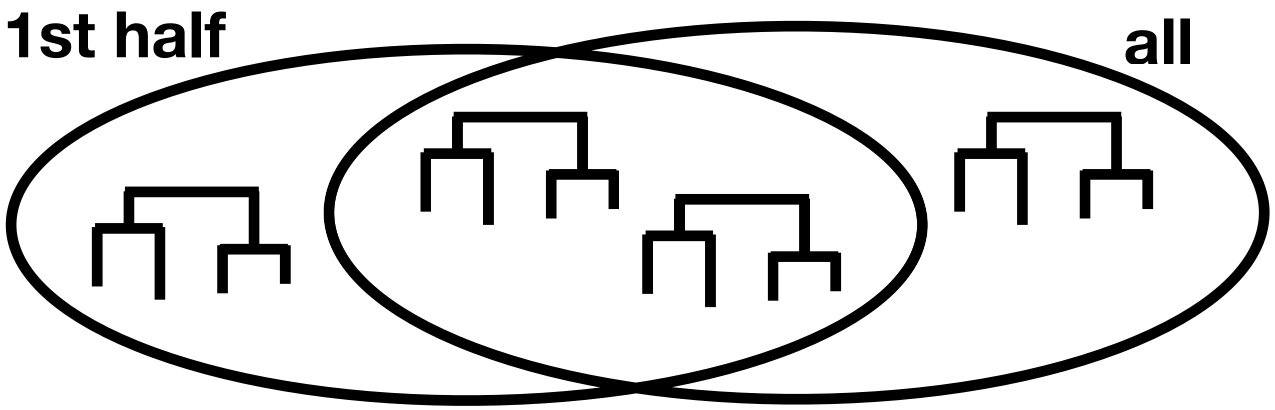


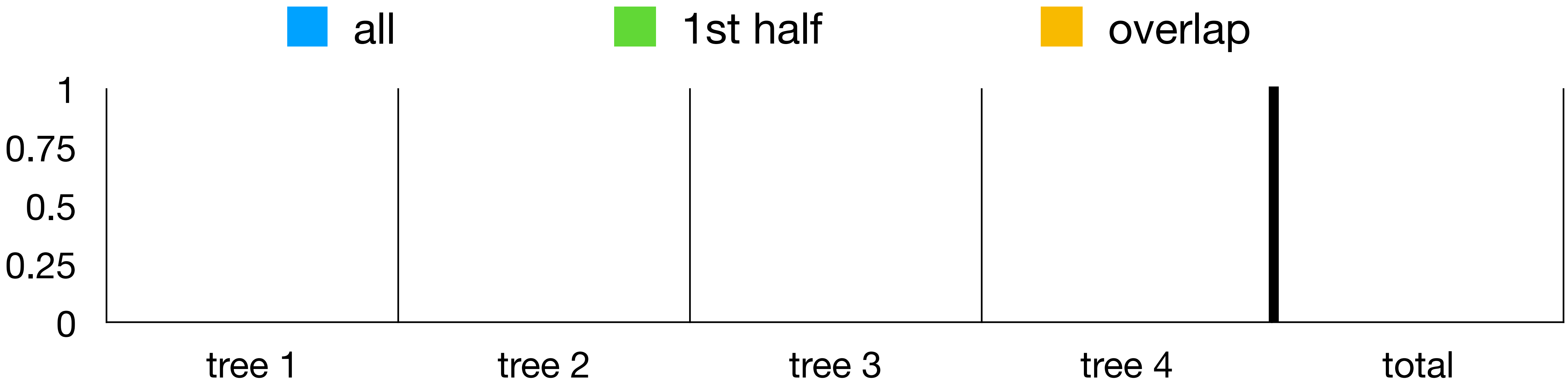
[H & Miller 2023]

Reproducibility in model selection

- **Problem:** infer phylogeny of 13 whale species from mitochondrial coding DNA
- Compute posterior tree probabilities based on **all**, **1st half**, and **2nd half**
- Compute overlap of 99% high probability regions

all	1st half	2nd half
Minke	GACCCGAACGTAATAA	ATCCGTTCCCATACTC
Blue	CACCCCCCCGTACTAT	TGAGTCCGAATTGGAA
Fin	TGTCTTCTACACTCCA	ACAGGTTGTACGTCAC
Grey	GGGTCGCTGTAGACCA	GATACCGCTCTCACAT



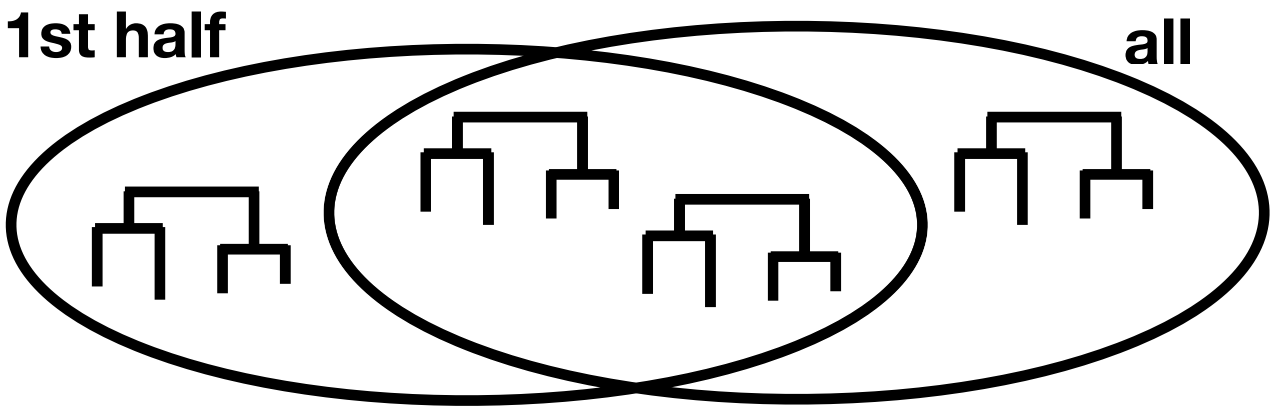


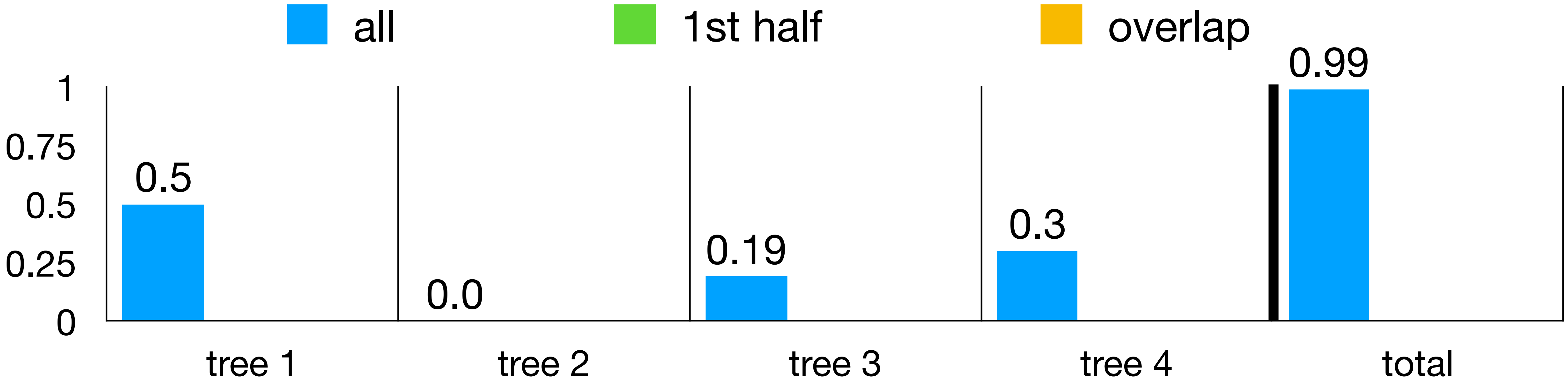
[H & Miller 2023]

Reproducibility in model selection

- **Problem:** infer phylogeny of 13 whale species from mitochondrial coding DNA
- Compute posterior tree probabilities based on **all**, **1st half**, and **2nd half**
- Compute overlap of 99% high probability regions

all	1st half	2nd half
Minke	GACCCGAACGTAATAA	ATCCGTTCCCATACTC
Blue	CACCCCCCCGTACTAT	TGAGTCCGAATTGGAA
Fin	TGTCTTCTACACTCCA	ACAGGTTGTACGTCAC
Grey	GGGTCGCTGTAGACCA	GATACCGCTCTCACAT

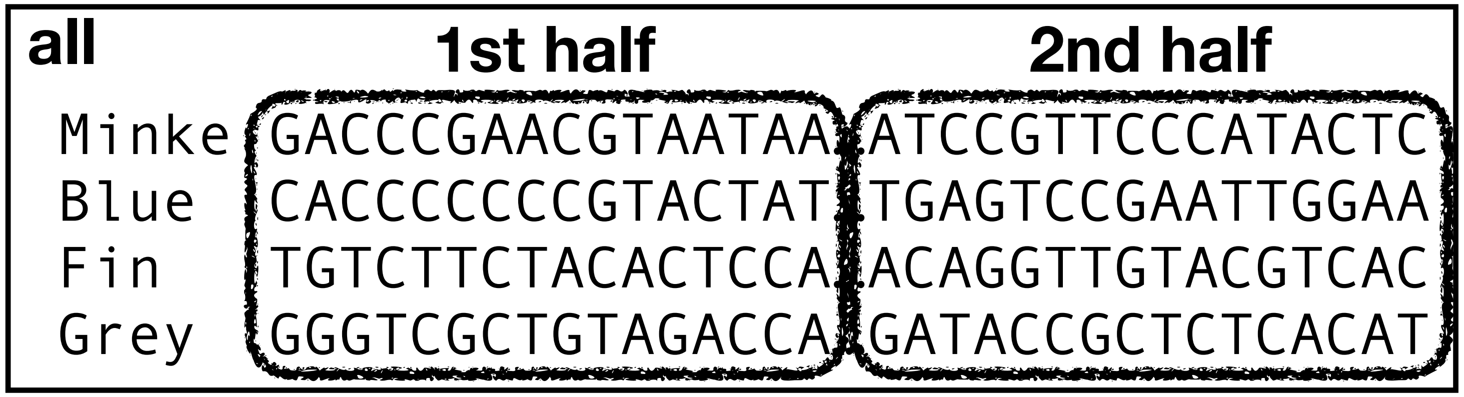




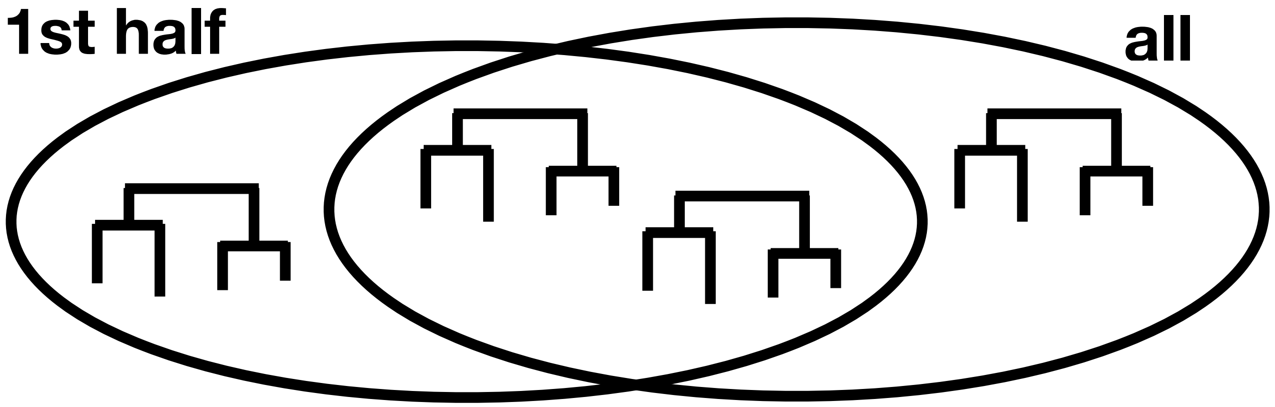
[H & Miller 2023]

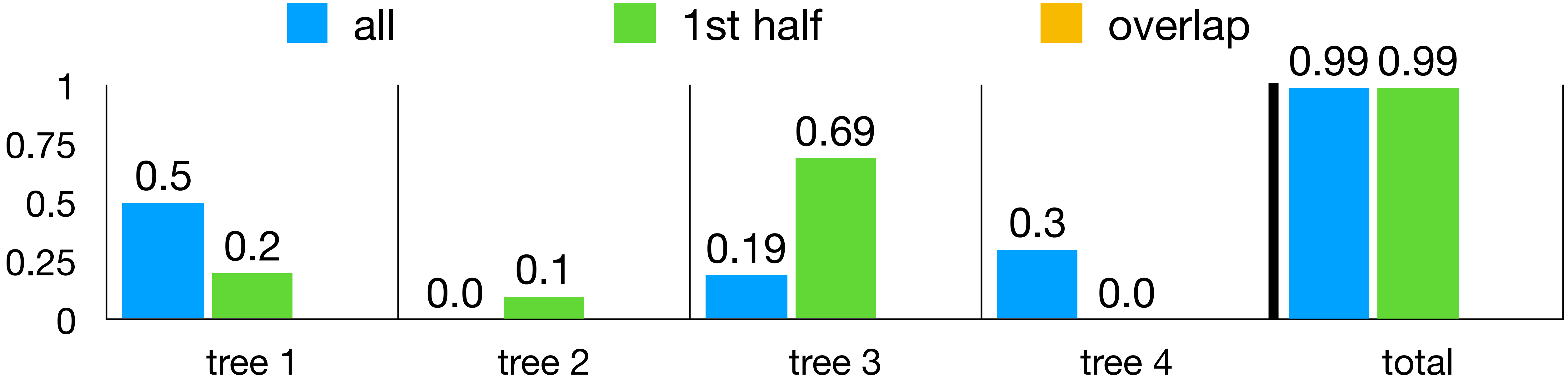
Reproducibility in model selection

- **Problem:** infer phylogeny of 13 whale species from mitochondrial coding DNA
- Compute posterior tree probabilities based on **all**, **1st half**, and **2nd half**
- Compute overlap of 99% high probability regions



```
Minke GACCCGAACGTAATAA ATCCGTTCCCATACTC
Blue  CACCCCCCCGTACTAT TGAGTCCGAATTGGAA
Fin   TGTCTTCTACACTCCA ACAGGTTGTACGTCAC
Grey  GGGTCGCTGTAGACCA GATACCGCTCTCACAT
```



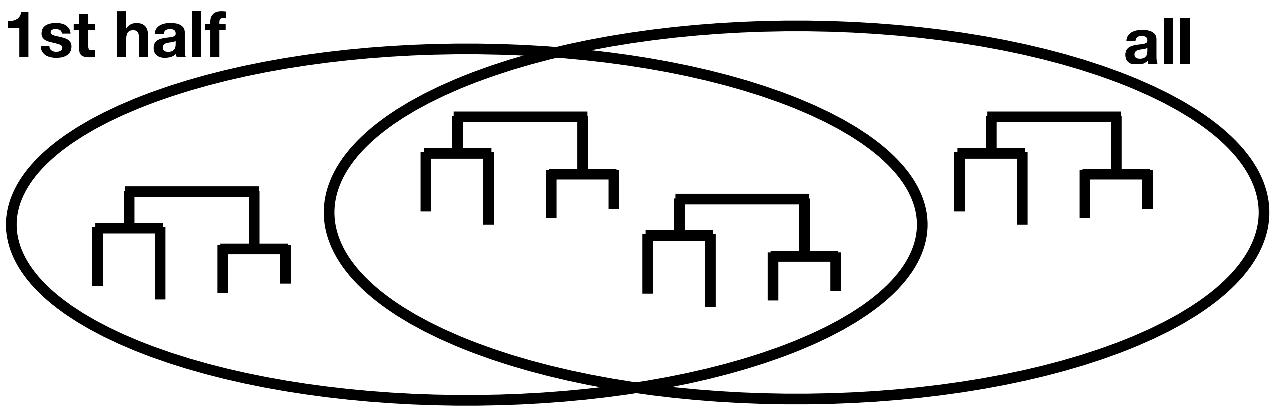


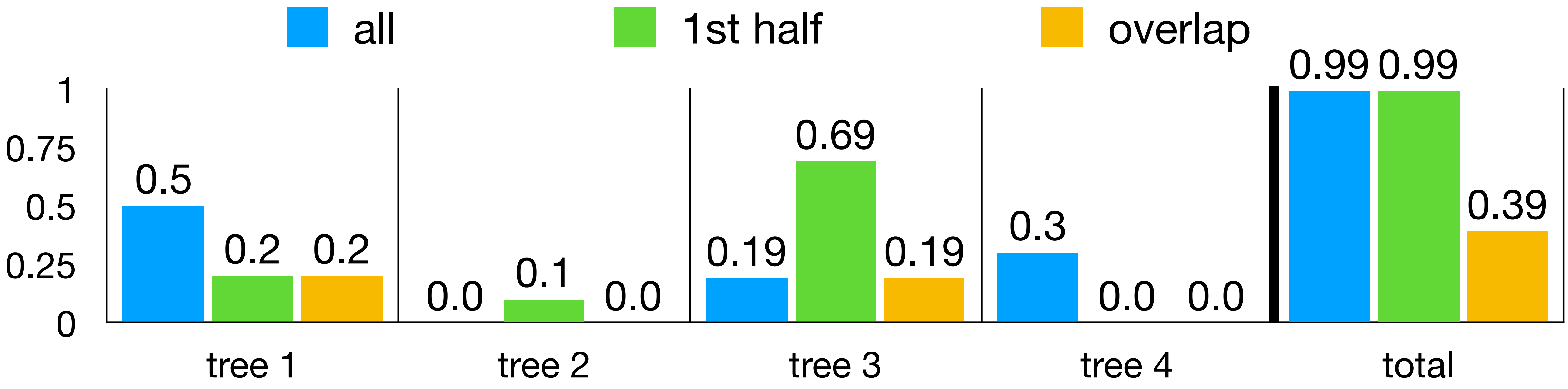
[**H** & Miller 2023]

Reproducibility in model selection

- **Problem:** infer phylogeny of 13 whale species from mitochondrial coding DNA
- Compute posterior tree probabilities based on **all**, **1st half**, and **2nd half**
- Compute overlap of 99% high probability regions

all	1st half	2nd half
Minke	GACCCGAACGTAATAA	ATCCGTTCCCATACTC
Blue	CACCCCCCCGTACTAT	TGAGTCCGAATTGGAA
Fin	TGTCTTCTACACTCCA	ACAGGTTGTACGTCAC
Grey	GGGTCGCTGTAGACCA	GATACCGCTCTCACAT



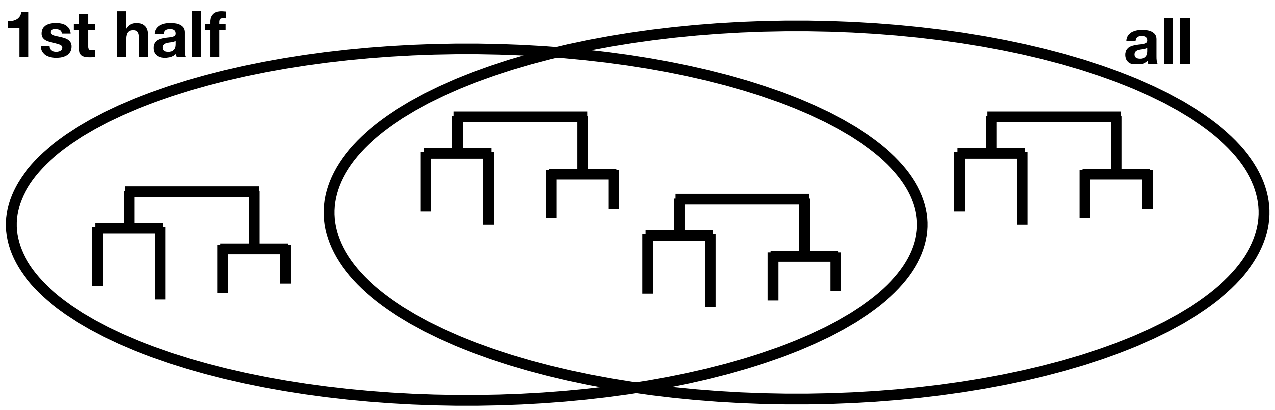


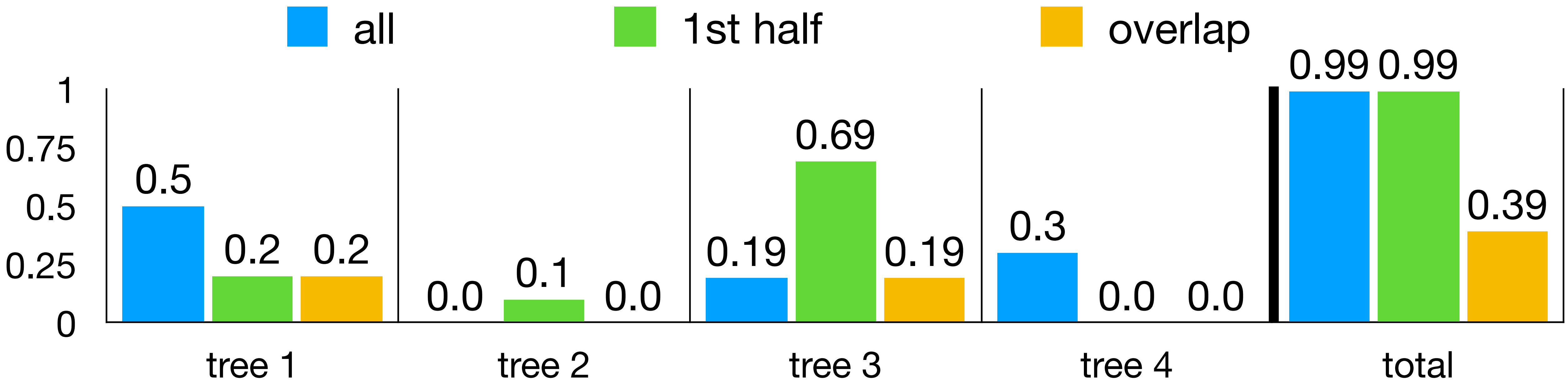
[H & Miller 2023]

Reproducibility in model selection

- **Problem:** infer phylogeny of 13 whale species from mitochondrial coding DNA
- Compute posterior tree probabilities based on **all**, **1st half**, and **2nd half**
- Compute overlap of 99% high probability regions
- 0% overlap = contradiction

all	1st half	2nd half
Minke	GACCCGAACGTAATAA	ATCCGTTCCCATACTC
Blue	CACCCCCCCGTACTAT	TGAGTCCGAATTGGAA
Fin	TGTCTTCTACACTCCA	ACAGGTTGTACGTCAC
Grey	GGGTCGCTGTAGACCA	GATACCGCTCTCACAT





[H & Miller 2023]

Bayesian phylogenetic inference may not be reproducible

Problem: infer phylogeny of 13 whale species

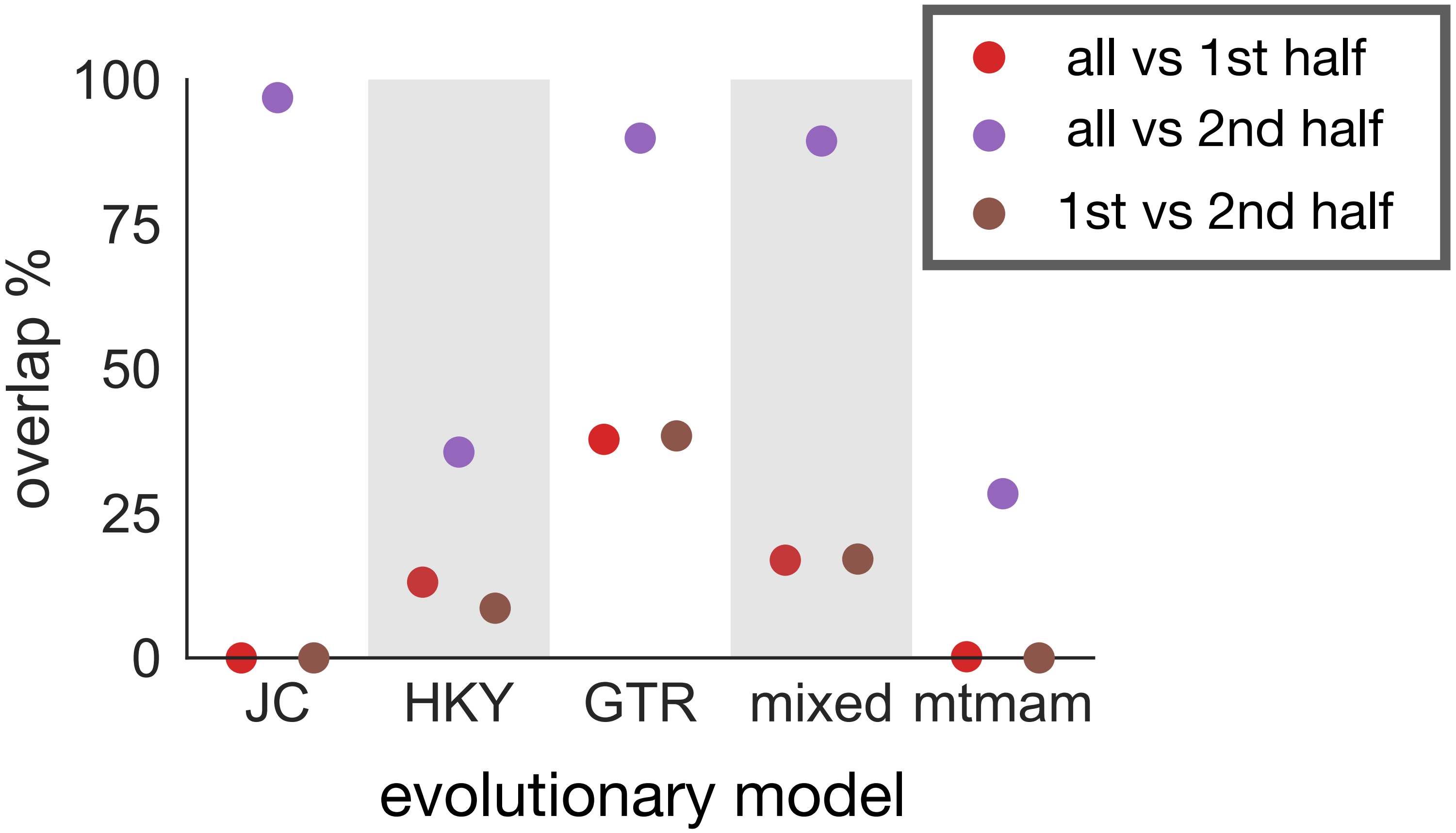
[**H** & Miller 2023]

Bayesian phylogenetic inference may not be reproducible

Problem: infer phylogeny of 13 whale species

- For some evolutionary models, little to no overlap

Cross-data consistency



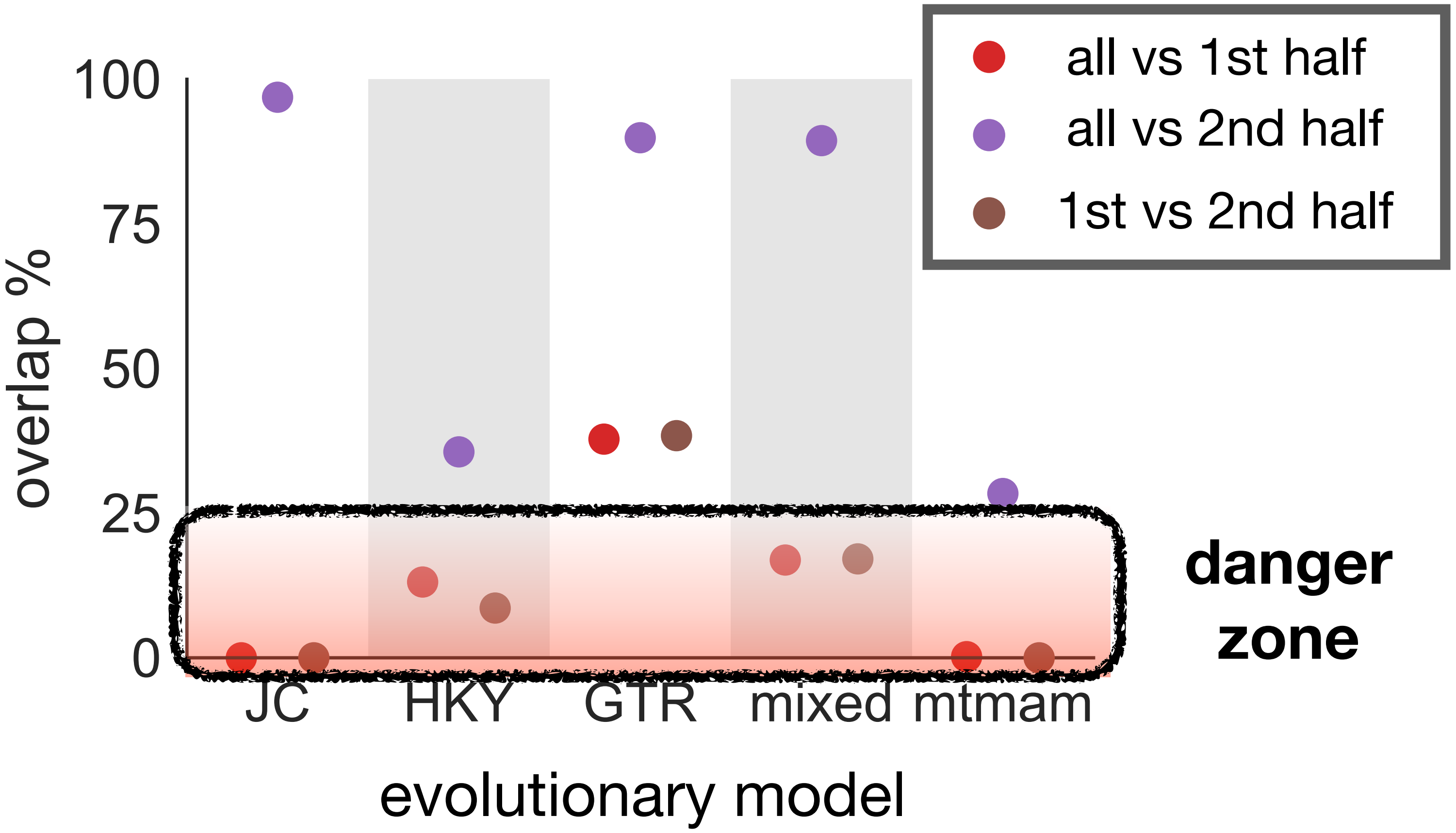
[H & Miller 2023]

Bayesian phylogenetic inference may not be reproducible

Problem: infer phylogeny of 13 whale species

- For some evolutionary models, little to no overlap

Cross-data consistency



danger zone

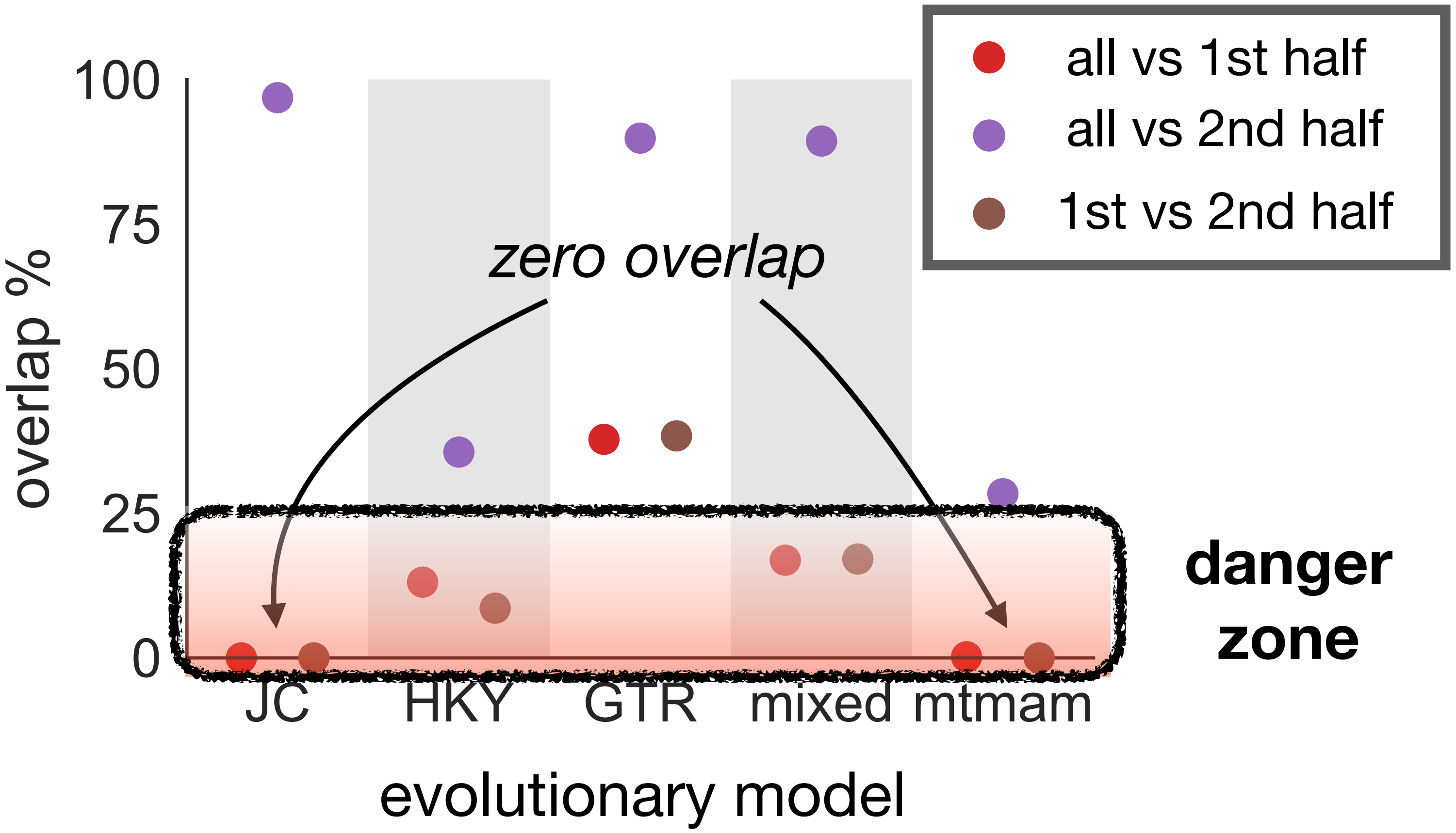
[H & Miller 2023]

Bayesian phylogenetic inference may not be reproducible

Problem: infer phylogeny of 13 whale species

- For some evolutionary models, little to no overlap

Cross-data consistency



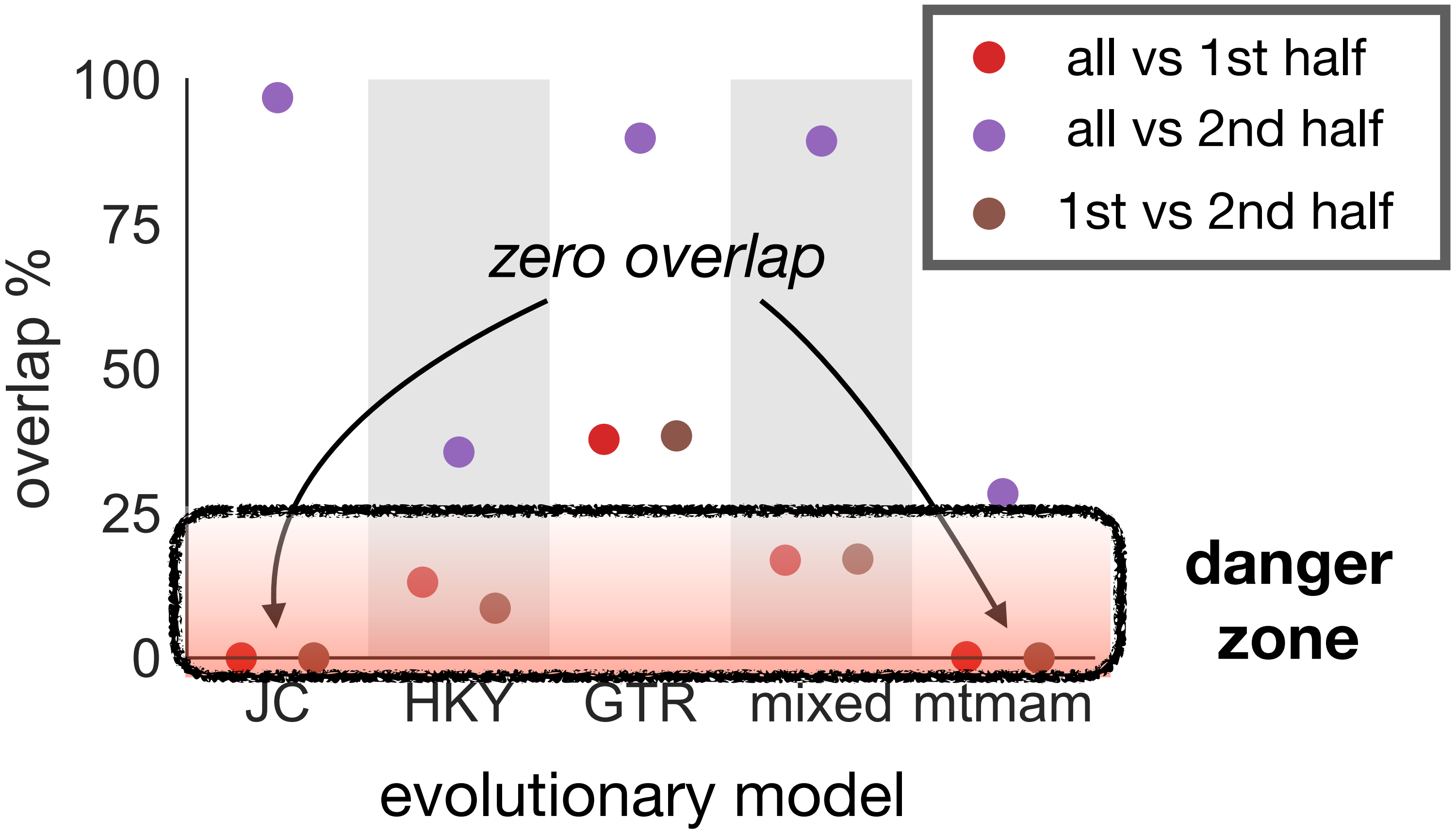
[H & Miller 2023]

Bayesian phylogenetic inference may not be reproducible

Problem: infer phylogeny of 13 whale species

- For some evolutionary models, little to no overlap
- (Bayesian) model selection is **unstable** and **not reproducible** [Wilcox et al. 2002, Alfaro et al. 2003, Douady et al. 2003, …]

Cross-data consistency



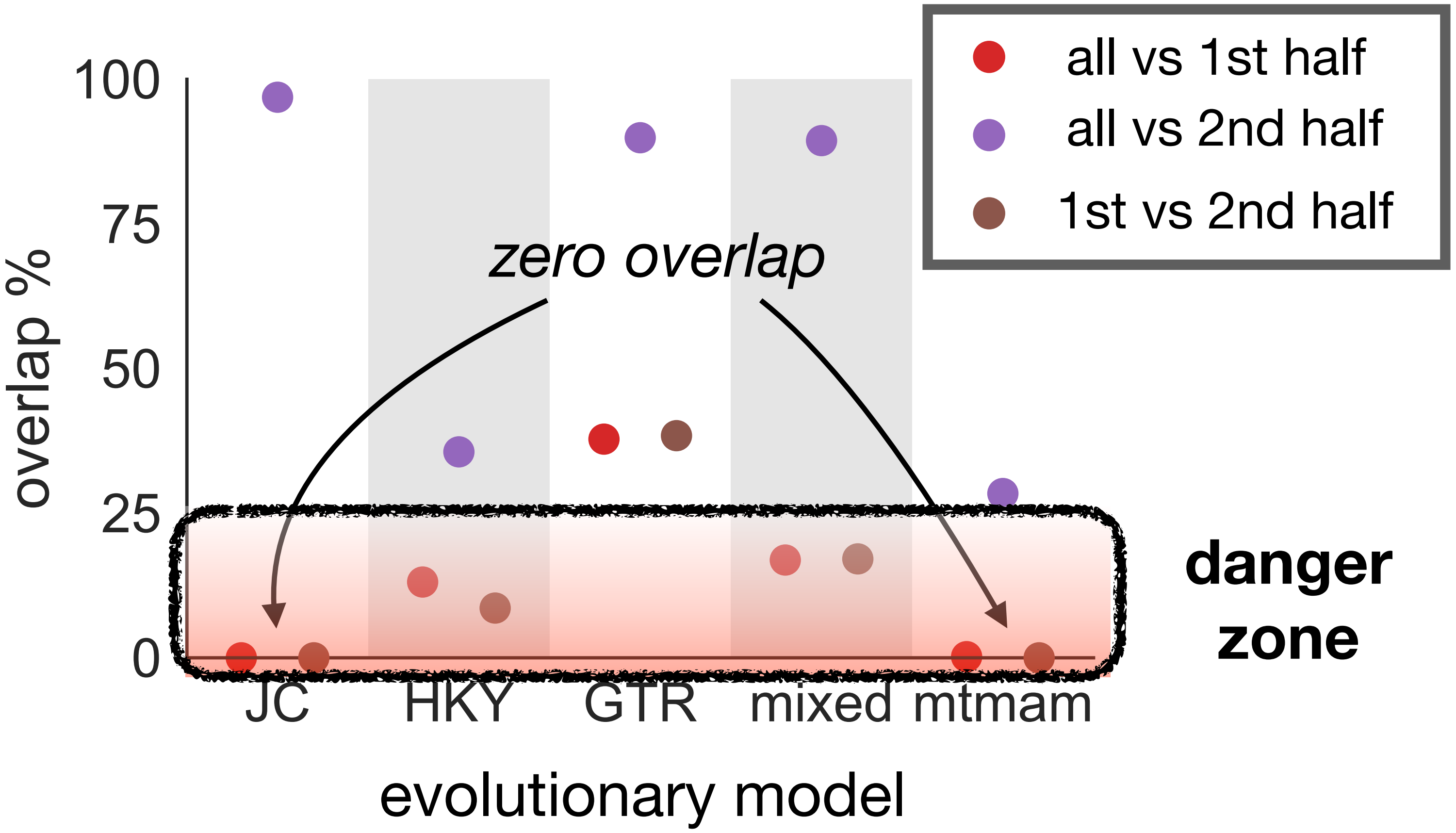
[H & Miller 2023]

Bayesian phylogenetic inference may not be reproducible

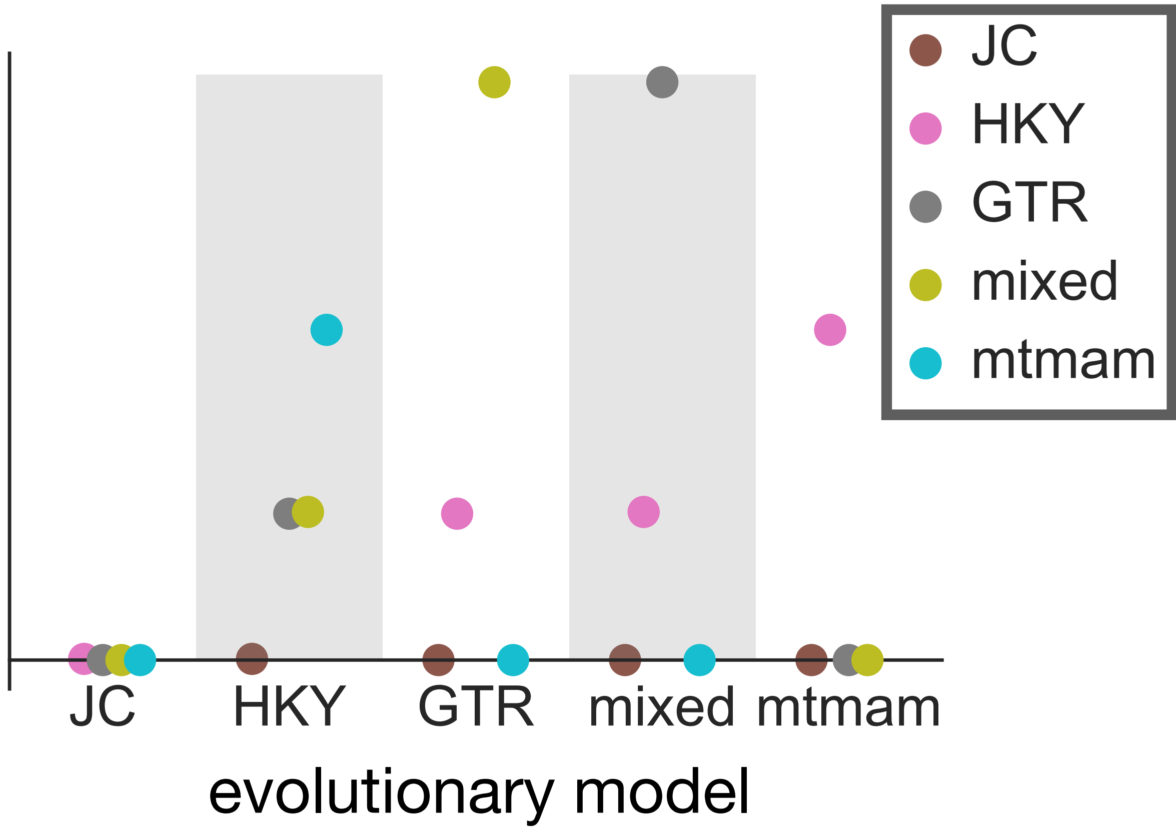
Problem: infer phylogeny of 13 whale species

- For some evolutionary models, little to no overlap
- (Bayesian) model selection is **unstable** and **not reproducible** [Wilcox et al. 2002, Alfaro et al. 2003, Douady et al. 2003, …]
- Same problem comparing evolutionary models with data fixed

Cross-data consistency



Cross-model consistency



[H & Miller 2023]

Overview

Overview

- **Minimal goal:** avoid methods that lead to contradictory/non-reproducible inferences

Overview

- **Minimal goal:** avoid methods that lead to contradictory/non-reproducible inferences
- For example, when…
  - getting more data
  - slightly tweaking model
  - using different data from same generative process

Overview

- **Minimal goal:** avoid methods that lead to contradictory/non-reproducible inferences
- For example, when…
  - getting more data
  - slightly tweaking model
  - using different data from same generative process
- **This talk:** when and how contradictions can arise in
  1. model selection
  2. prediction with high-dimensional models
  3. unsupervised learning

Overview

- **Minimal goal:** avoid methods that lead to contradictory/non-reproducible inferences
- For example, when…
  - getting more data
  - slightly tweaking model
  - using different data from same generative process
- **This talk:** when and how contradictions can arise in
  1. model selection
  2. prediction with high-dimensional models
  3. unsupervised learning
- **Takeaways:**
  - A. Non-reproducibility can be **subtle** (and is **problem-dependent**)
  - B. Not specific to Bayes
  - C. Need default, low-cost **protective methods** that remain **statistically efficient**

Part 1: Model Selection

[Yang & Zhu 2018, **H** & Miller 2023]

Part 1: Model Selection

- Assume data x and models m_1 and m_2

[Yang & Zhu 2018, **H** & Miller 2023]

Part 1: Model Selection

- Assume data x and models m_1 and m_2
- Assume they explain the data-generating distribution equally well: $\mathbb{E}\{\log p(x \mid m_1)\} = \mathbb{E}\{\log p(x \mid m_2)\}$

[Yang & Zhu 2018, **H** & Miller 2023]

Part 1: Model Selection

- Assume data x and models m_1 and m_2

- Assume they explain the data-generating distribution equally well: $\mathbb{E}\{\log p(x \mid m_1)\} = \mathbb{E}\{\log p(x \mid m_2)\}$

- We'd **hope** models have equal posterior probability as $N \to \infty$: $\pi(m_1 \mid x) = \pi(m_2 \mid x) = 1/2$

[Yang & Zhu 2018, **H** & Miller 2023]

Part 1: Model Selection

- Assume data x and models m_1 and m_2
- Assume they explain the data-generating distribution equally well:
  $\mathbb{E}\{\log p(x \mid m_1)\} = \mathbb{E}\{\log p(x \mid m_2)\}$
- We'd **hope** models have equal posterior probability as $N \to \infty$:
  $\pi(m_1 \mid x) = \pi(m_2 \mid x) = 1/2$
- However…

Theorem **[Yang & Zhu 2019, H & Miller 2023]:**
As $N \to \infty$, $\pi(m_1 \mid x) \overset{d}{\to} \text{Bernoulli}(0.5)$

[Yang & Zhu 2018, **H** & Miller 2023]

Part 1: Model Selection

- Assume data x and models m_1 and m_2

- Assume they explain the data-generating distribution equally well:
  $\mathbb{E}\{\log p(x \mid m_1)\} = \mathbb{E}\{\log p(x \mid m_2)\}$

- We'd **hope** models have equal posterior probability as $N \to \infty$:
  $\pi(m_1 \mid x) = \pi(m_2 \mid x) = 1/2$

- However…

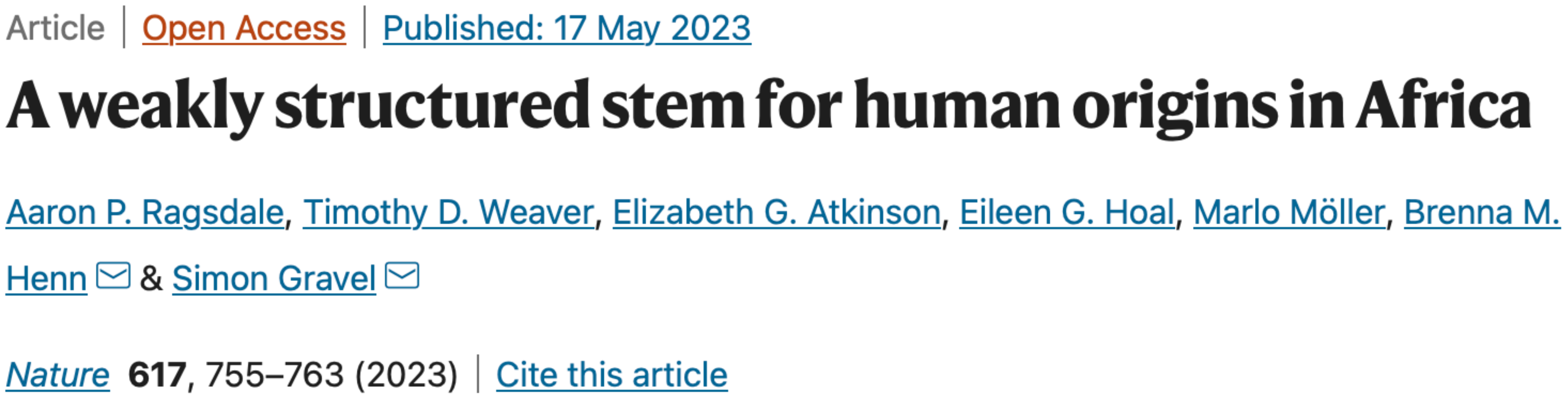
Theorem **[Yang & Zhu 2019, H & Miller 2023]:**

As $N \to \infty$, $\pi(m_1 \mid x) \overset{d}{\to} \text{Bernoulli}(0.5)$

all posterior mass on a single, arbitrary model

[Yang & Zhu 2018, **H** & Miller 2023]

This is not an abstract concern…



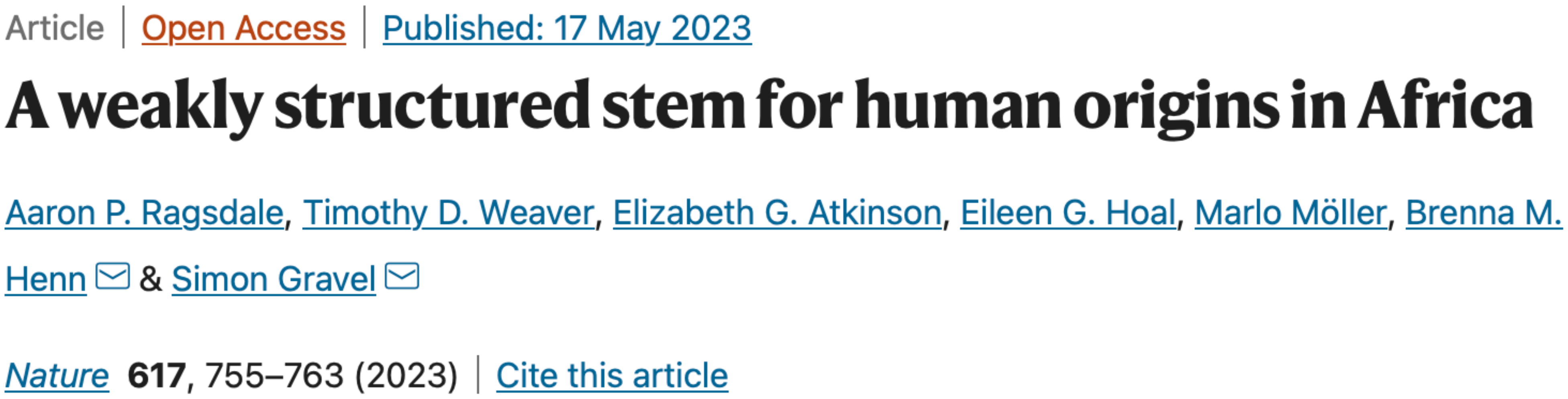
Article | Open Access | Published: 17 May 2023

A weakly structured stem for human origins in Africa

Aaron P. Ragsdale, Timothy D. Weaver, Elizabeth G. Atkinson, Eileen G. Hoal, Marlo Möller, Brenna M. Henn & Simon Gravel

Nature **617**, 755–763 (2023) | Cite this article

This is not an abstract concern…



Article | Open Access | Published: 17 May 2023

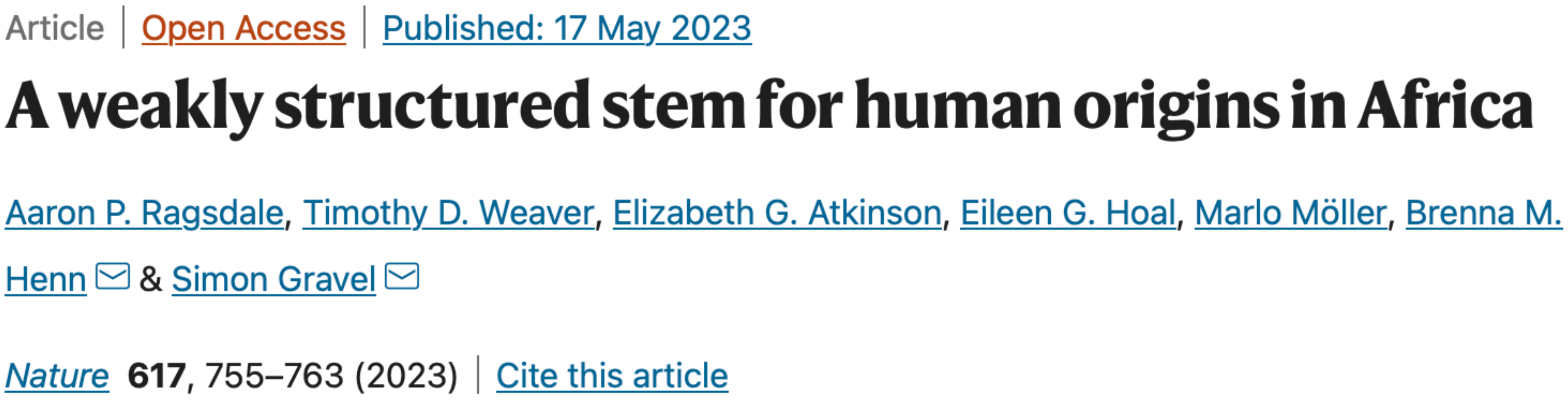
A weakly structured stem for human origins in Africa

Aaron P. Ragsdale, Timothy D. Weaver, Elizabeth G. Atkinson, Eileen G. Hoal, Marlo Möller, Brenna M. Henn ✉ & Simon Gravel ✉

Nature **617**, 755–763 (2023) | Cite this article

Despite broad agreement that Homo sapiens originated in Africa, ***considerable uncertainty surrounds specific models of divergence and migration*** across the continent…

This is not an abstract concern…



Article | Open Access | Published: 17 May 2023

A weakly structured stem for human origins in Africa

Aaron P. Ragsdale, Timothy D. Weaver, Elizabeth G. Atkinson, Eileen G. Hoal, Marlo Möller, Brenna M. Henn ✉ & Simon Gravel ✉

Nature **617**, 755–763 (2023) | Cite this article

Despite broad agreement that Homo sapiens originated in Africa, ***considerable uncertainty surrounds specific models of divergence and migration*** across the continent…

…Progress is hampered by a shortage of fossil and genomic data, as well as ***variability in previous estimates of divergence times.*** Here we seek to discriminate among such models…

This is not an abstract concern...

Article | Open Access | Published: 17 May 2023

A weakly structured stem for human origins in Africa

Aaron P. Ragsdale, Timothy D. Weaver, Elizabeth G. Atkinson, Eileen G. Hoal, Marlo Möller, Brenna M. Henn & Simon Gravel

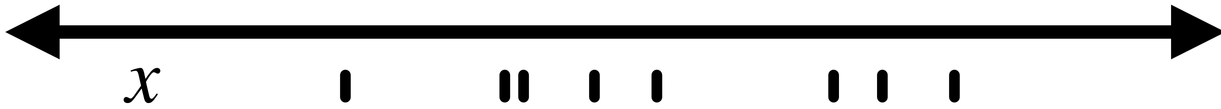
Nature **617**, 755–763 (2023) | Cite this article

Despite broad agreement that Homo sapiens originated in Africa, ***considerable uncertainty surrounds specific models of divergence and migration*** across the continent…

…Progress is hampered by a shortage of fossil and genomic data, as well as ***variability in previous estimates of divergence times.*** Here we seek to discriminate among such models…

…We show that ***model misspecification explains the variation in previous estimates of divergence time***

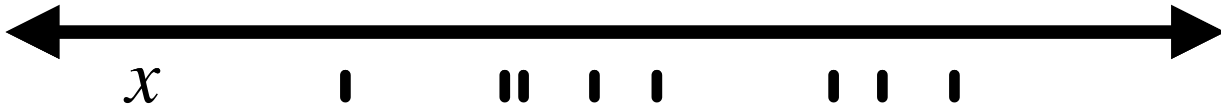
A solution: the bagged posterior



[Douady et al. 2003, Bühlmann 2014, **H** & Miller 2023]

A solution: the bagged posterior

- Have data $x = (x_1, \ldots, x_N)$

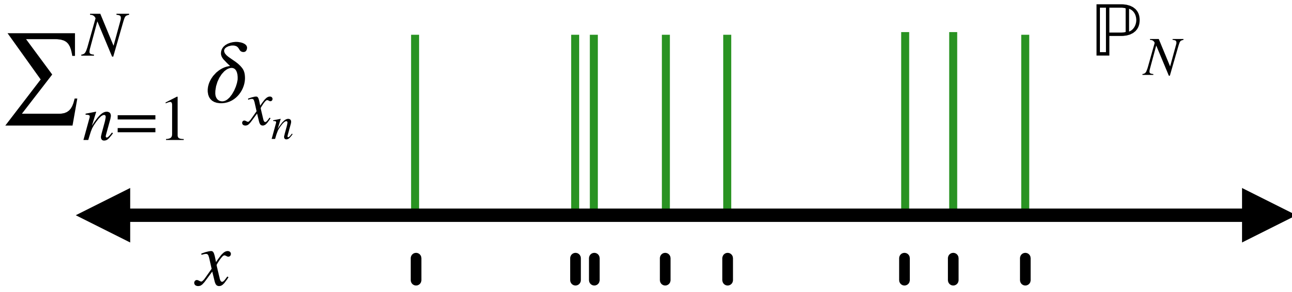


[Douady et al. 2003, Bühlmann 2014, **H** & Miller 2023]

A solution: the bagged posterior

- Have data $x = (x_1, \ldots, x_N)$

- Empirical data distribution $\mathbb{P}_N = N^{-1} \sum_{n=1}^{N} \delta_{x_n}$



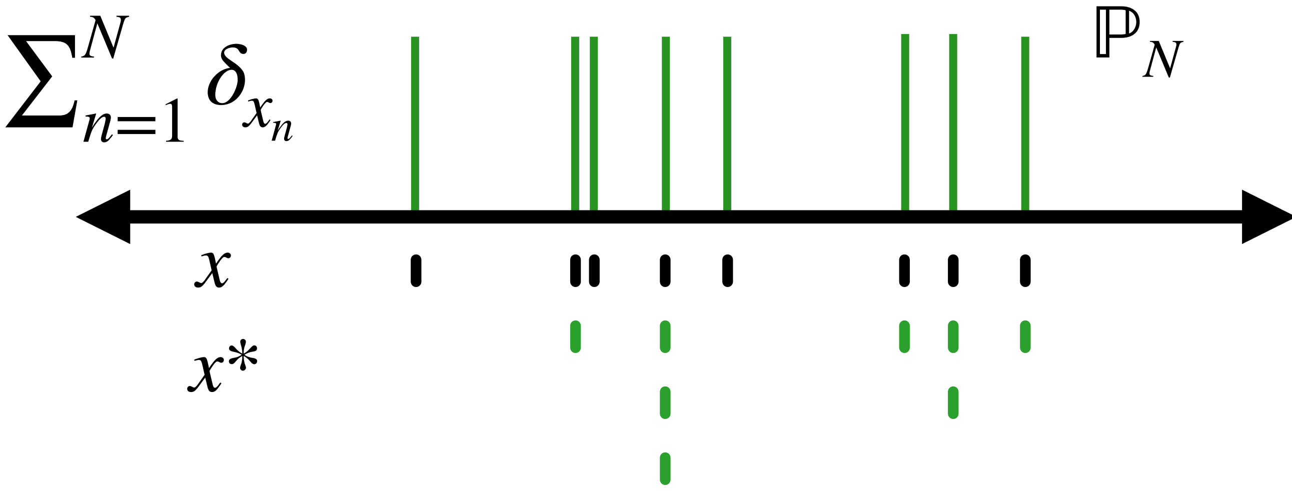
[Douady et al. 2003, Bühlmann 2014, **H** & Miller 2023]

A solution: the bagged posterior

- Have data $x = (x_1, \ldots, x_N)$

- Empirical data distribution $\mathbb{P}_N = N^{-1} \sum_{n=1}^{N} \delta_{x_n}$

- Bootstrap dataset $x^* = (x^*_1, \ldots, x^*_M)$, where x^*_m i.i.d. $\sim \mathbb{P}_N$



[Douady et al. 2003, Bühlmann 2014, **H** & Miller 2023]

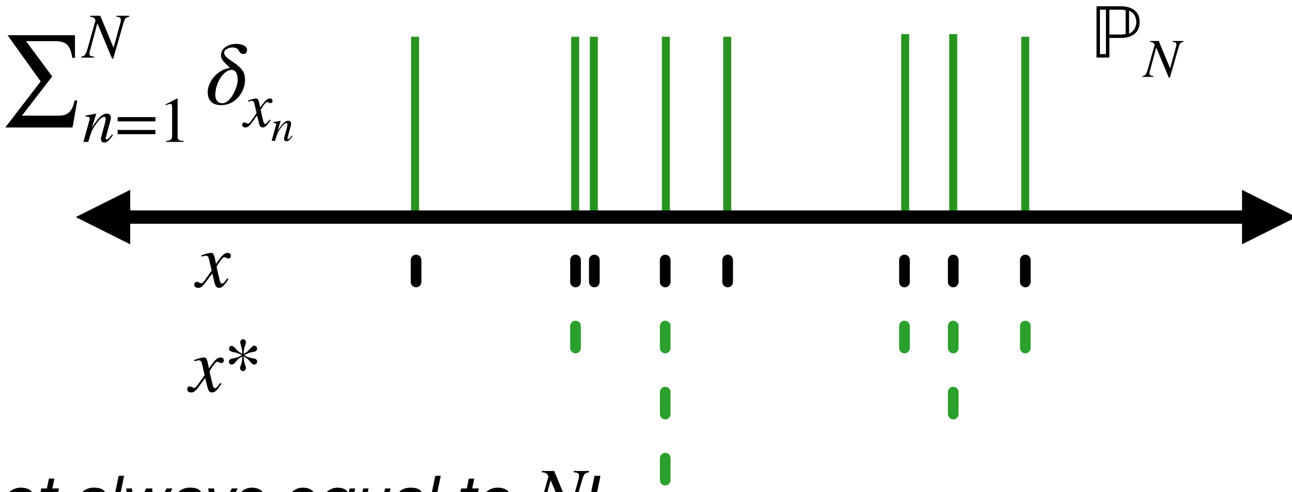
A solution: the bagged posterior

- Have data $x = (x_1, \ldots, x_N)$

- Empirical data distribution $\mathbb{P}_N = N^{-1} \sum_{n=1}^{N} \delta_{x_n}$

- Bootstrap dataset $x^* = (x_1^*, \ldots, x_M^*)$,

  where x_m^* i.i.d. $\sim \mathbb{P}_N$



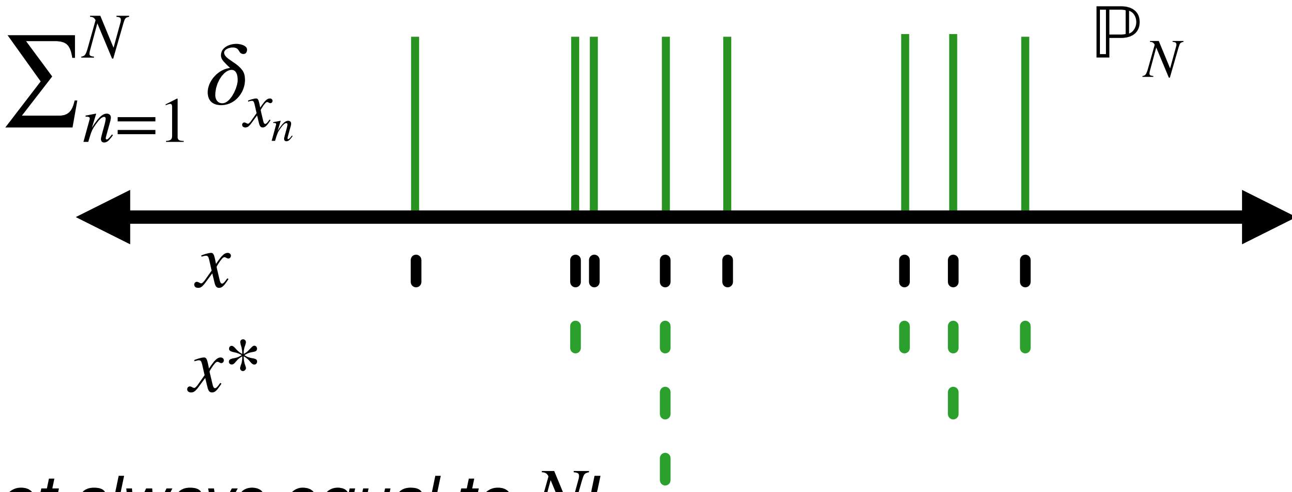
not always equal to N!

[Douady et al. 2003, Bühlmann 2014, **H** & Miller 2023]

A solution: the bagged posterior

- Have data $x = (x_1, \ldots, x_N)$

- Empirical data distribution $\mathbb{P}_N = N^{-1} \sum_{n=1}^{N} \delta_{x_n}$



- Bootstrap dataset $x^* = (x_1^*, \ldots, x_M^*)$, where x_m^* i.i.d. $\sim \mathbb{P}_N$

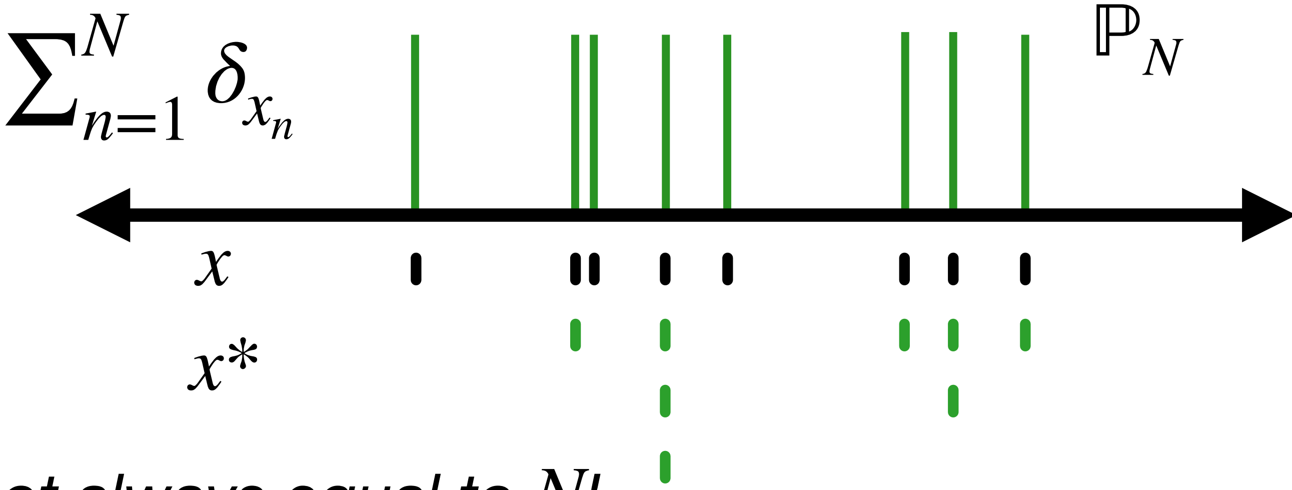
not always equal to N!

- **Bagged posterior** (a.k.a. **BayesBag**): $\pi^*(\theta \mid x) = \mathbb{E}\{\pi(\theta \mid x^*) \mid x\}$

[Douady et al. 2003, Bühlmann 2014, **H** & Miller 2023]

A solution: the bagged posterior

- Have data $x = (x_1, \ldots, x_N)$
- Empirical data distribution $\mathbb{P}_N = N^{-1} \sum_{n=1}^{N} \delta_{x_n}$



- Bootstrap dataset $x^* = (x_1^*, \ldots, x_M^*)$,

  where x_m^* i.i.d. $\sim \mathbb{P}_N$

not always equal to N!

- **Bagged posterior** (a.k.a. **BayesBag**):

  $\pi^*(\theta \mid x) = \mathbb{E}\{\pi(\theta \mid x^*) \mid x\}$

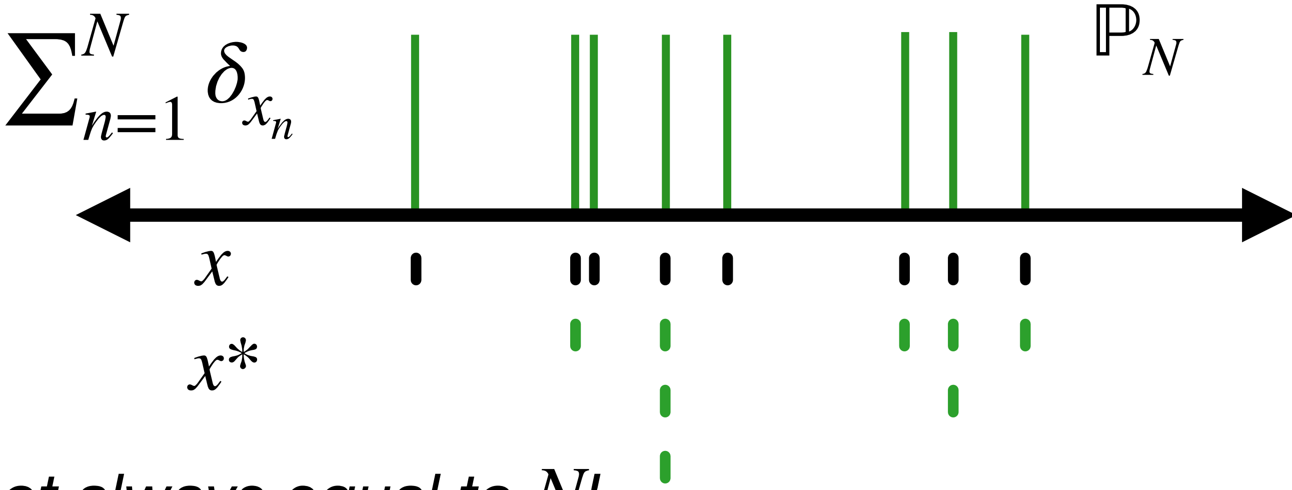
- In practice, sample B bootstrap datasets: $\pi^*(\theta \mid x) \approx \sum_{b=1}^{B} \pi(\theta \mid x_{(b)}^*)$

[Douady et al. 2003, Bühlmann 2014, **H** & Miller 2023]

A solution: the bagged posterior

- Have data $x = (x_1, \ldots, x_N)$

- Empirical data distribution $\mathbb{P}_N = N^{-1} \sum_{n=1}^{N} \delta_{x_n}$



- Bootstrap dataset $x^* = (x^*_1, \ldots, x^*_M)$,
  where x^*_m i.i.d. $\sim \mathbb{P}_N$

not always equal to N!

- **Bagged posterior** (a.k.a. **BayesBag**):
  $\pi^*(\theta \mid x) = \mathbb{E}\{\pi(\theta \mid x^*) \mid x\}$

- In practice, sample B bootstrap datasets: $\pi^*(\theta \mid x) \approx \sum_{b=1}^{B} \pi(\theta \mid x^*_{(b)})$
  - Suffices to take $B = 50$ or 100

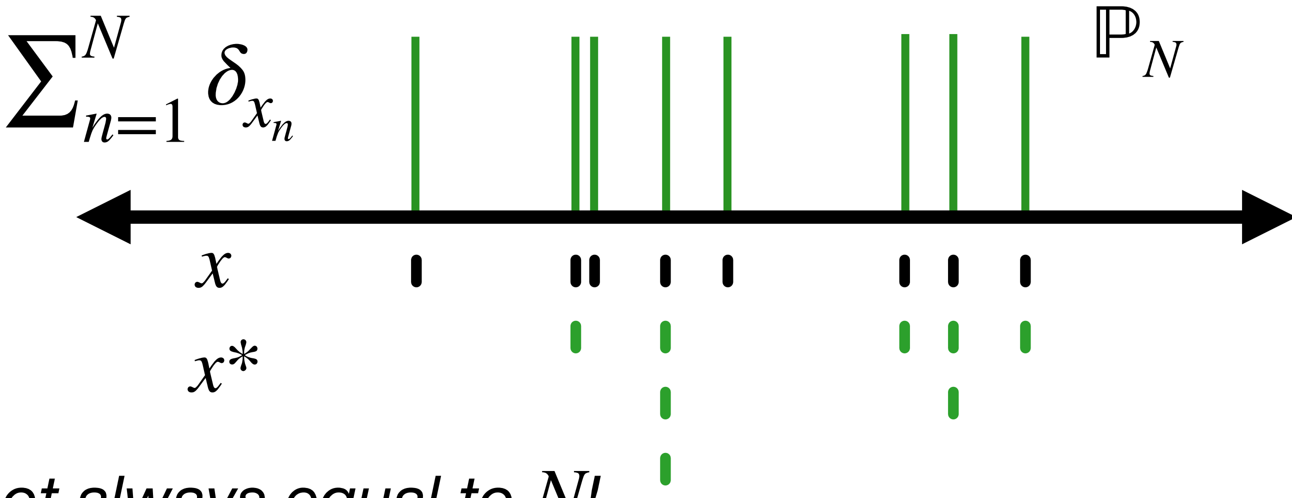
[Douady et al. 2003, Bühlmann 2014, **H** & Miller 2023]

A solution: the bagged posterior

- Have data $x = (x_1, \ldots, x_N)$

- Empirical data distribution $\mathbb{P}_N = N^{-1} \sum_{n=1}^{N} \delta_{x_n}$

- Bootstrap dataset $x^* = (x^*_1, \ldots, x^*_M)$,
  where x^*_m i.i.d. $\sim \mathbb{P}_N$



not always equal to N!

- **Bagged posterior** (a.k.a. **BayesBag**):
  $\pi^*(\theta \mid x) = \mathbb{E}\{\pi(\theta \mid x^*) \mid x\}$

- In practice, sample B bootstrap datasets: $\pi^*(\theta \mid x) \approx \sum_{b=1}^{B} \pi(\theta \mid x^*_{(b)})$
  - Suffices to take $B = 50$ or 100
  - Benefits: **easy to use**, can **parallelize** across B

[Douady et al. 2003, Bühlmann 2014, **H** & Miller 2023]

The bagged posterior is stable

- Assume data x and two models m_1 and m_2
- Assume they explain the data-generating distribution equally well:
  $\mathbb{E}\{\log p(x \mid m_1)\} = \mathbb{E}\{\log p(x \mid m_2)\}$
- We'd **hope** models have equal poster probability as $N \to \infty$:
  $\pi(m_1 \mid Y) = \pi(m_2 \mid Y) = 1/2$
- However…

Theorem [Yang & Zhu 2019, H & Miller 2023]:
As $N \to \infty$, $\pi(m_1 \mid x) \xrightarrow{d} \text{Bernoulli}(0.5)$

all posterior mass on a single, arbitrary model

[Yang & Zhu 2018, **H** & Miller 2023]

The bagged posterior is stable

- Assume data x and two models m_1 and m_2
- Assume they explain the data-generating distribution equally well:
  $\mathbb{E}\{\log p(x \mid m_1)\} = \mathbb{E}\{\log p(x \mid m_2)\}$
- We'd **hope** models have equal poster probability as $N \to \infty$:
  $\pi(m_1 \mid Y) = \pi(m_2 \mid Y) = 1/2$
- However…

Theorem [Yang & Zhu 2019, H & Miller 2023]:
As $N \to \infty$, $\pi(m_1 \mid x) \overset{d}{\to} \text{Bernoulli}(0.5)$

all posterior mass on a single, arbitrary model

Theorem [H & Miller 2023]: As $N \to \infty$

1. If $M = N$, then $\pi^*(m_1 \mid x) \overset{d}{\to} \text{Uniform}(0,1)$

[Yang & Zhu 2018, H & Miller 2023]

The bagged posterior is stable

- Assume data x and two models m_1 and m_2

- Assume they explain the data-generating distribution equally well:
  $\mathbb{E}\{\log p(x \mid m_1)\} = \mathbb{E}\{\log p(x \mid m_2)\}$

- We'd **hope** models have equal poster probability as $N \to \infty$:
  $\pi(m_1 \mid Y) = \pi(m_2 \mid Y) = 1/2$

- However…

Theorem [Yang & Zhu 2019, H & Miller 2023]:
As $N \to \infty$, $\pi(m_1 \mid x) \xrightarrow{d} \text{Bernoulli}(0.5)$

all posterior mass on a single, arbitrary model

Theorem [H & Miller 2023]: As $N \to \infty$

1. If $M = N$, then $\pi^*(m_1 \mid x) \xrightarrow{d} \text{Uniform}(0{,}1)$
2. If $M/N \to 0$, then $\pi^*(m_1 \mid x) \xrightarrow{P} 1/2$

[Yang & Zhu 2018, H & Miller 2023]

The bagged posterior is stable

- Assume data x and two models m_1 and m_2

- Assume they explain the data-generating distribution equally well:
  $\mathbb{E}\{\log p(x \mid m_1)\} = \mathbb{E}\{\log p(x \mid m_2)\}$

- We'd **hope** models have equal poster probability as $N \to \infty$:
  $\pi(m_1 \mid Y) = \pi(m_2 \mid Y) = 1/2$

- However…

Theorem [Yang & Zhu 2019, H & Miller 2023]:
As $N \to \infty$, $\pi(m_1 \mid x) \xrightarrow{d} \mathrm{Bernoulli}(0.5)$

all posterior mass on a single, arbitrary model

Theorem [H & Miller 2023]: As $N \to \infty$

1. If $M = N$, then $\pi^*(m_1 \mid x) \xrightarrow{d} \mathrm{Uniform}(0,1)$
2. If $M/N \to 0$, then $\pi^*(m_1 \mid x) \xrightarrow{P} 1/2$

Recommendation:

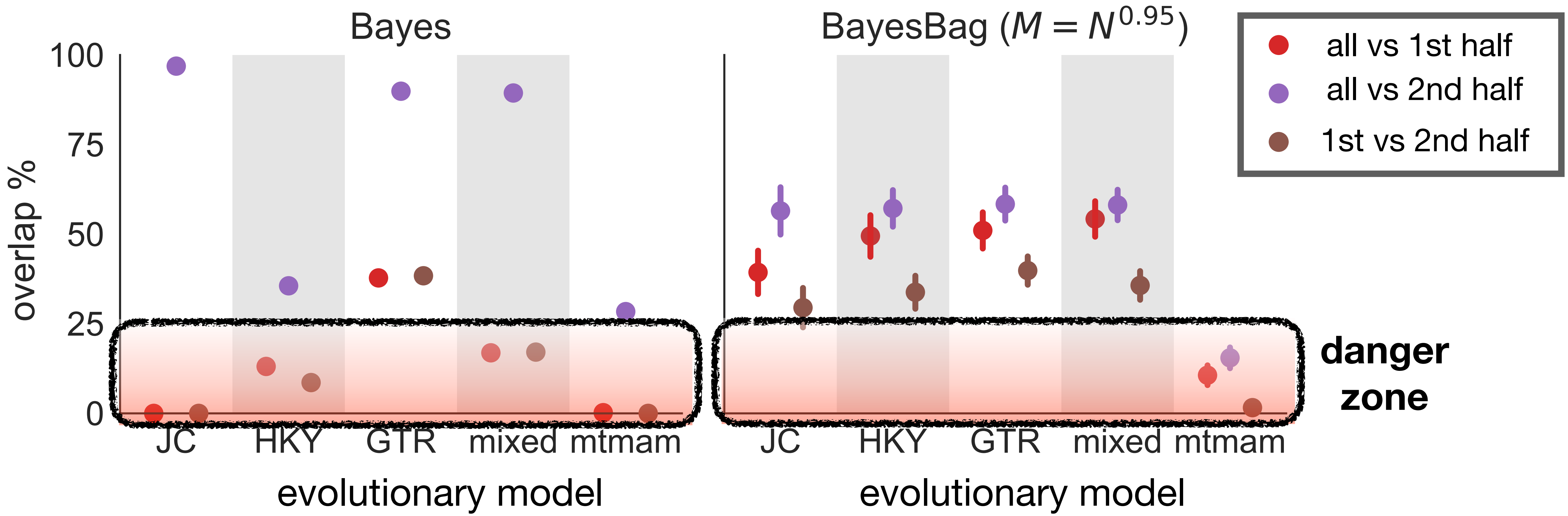
- $M = N^{0.95}$ default
- $M = N^{0.75}$ if significant misspecification and/or many models

[Yang & Zhu 2018, H & Miller 2023]

Bagged posterior improves stability

Goal: infer phylogeny of 13 whale species

Cross-data consistency

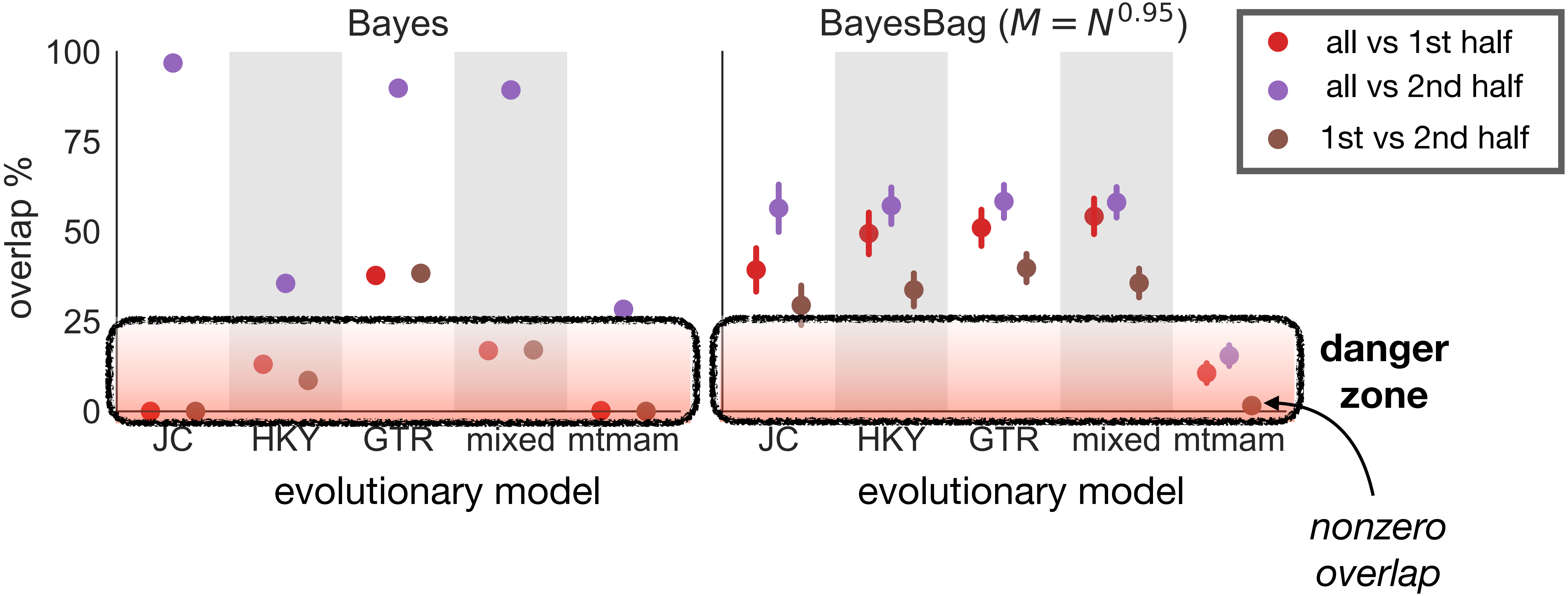


[H & Miller 2023]

Bagged posterior improves stability

Goal: infer phylogeny of 13 whale species

Cross-data consistency

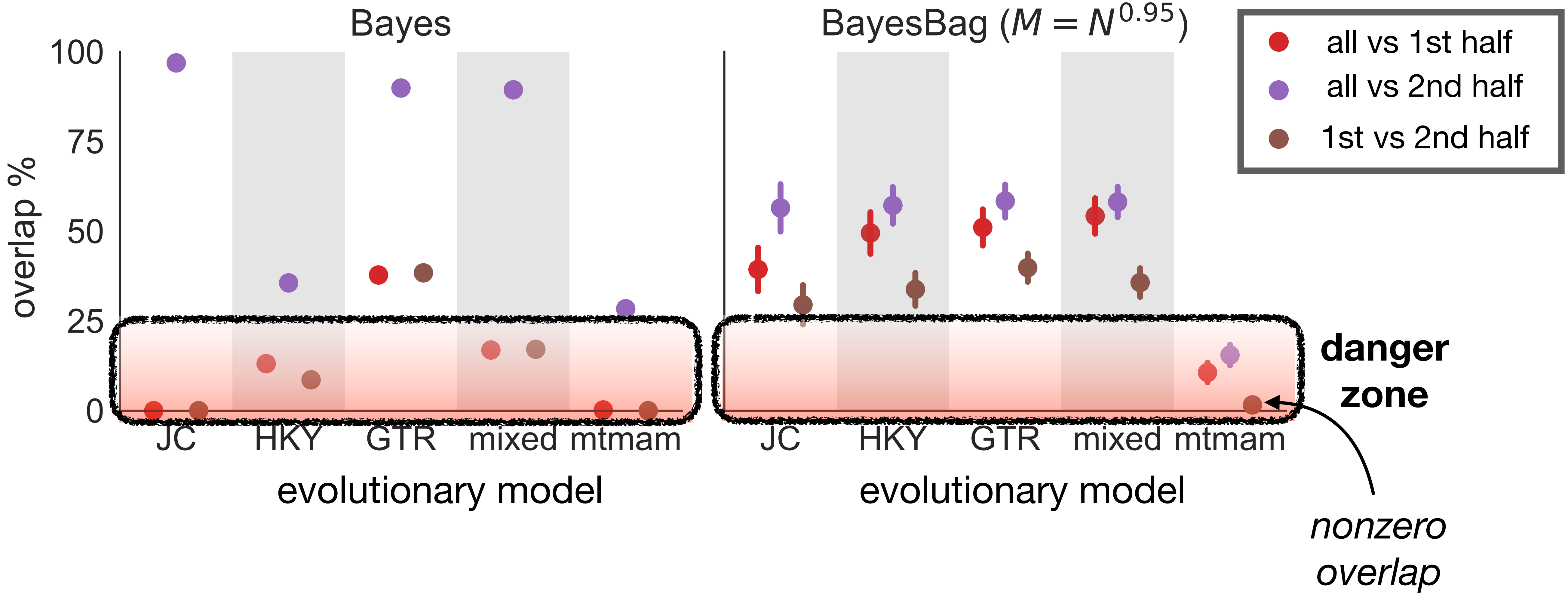


[**H** & Miller 2023]

Bagged posterior improves stability

Goal: infer phylogeny of 13 whale species

Cross-data consistency



- Significant overlap between non-MTMAM BayesBag and mixed Bayes
- BayesBag dramatically improves cross-model consistency too

[H & Miller 2023]

Part 2: Prediction with High-Dimensional Models

Part 2: Prediction with High-Dimensional Models

- Part 1: Bayesian model selection may not be self-consistent (i.e., reproducible) if all models are misspecified

Part 2: Prediction with High-Dimensional Models

- Part 1: Bayesian model selection may not be self-consistent (i.e., reproducible) if all models are misspecified
- But what about parameter estimation?

Part 2: Prediction with High-Dimensional Models

- Part 1: Bayesian model selection may not be self-consistent (i.e., reproducible) if all models are misspecified

- But what about parameter estimation?

- Usual focus is on pseudo-true parameter (i.e., KL-minimizing parameter) $\theta^\star := \arg\max_\theta \mathbb{E}\{\log p(x_1 \mid \theta)\}$ [Kleijn & van der Vaart 2012, De Blasi & Walker 2012, Walker 2013, **H** & Miller 2019, Hoff & Wakefield 2021]

Part 2: Prediction with High-Dimensional Models

- Part 1: Bayesian model selection may not be self-consistent (i.e., reproducible) if all models are misspecified

- But what about parameter estimation?

- Usual focus is on pseudo-true parameter (i.e., KL-minimizing parameter) $\theta^\star := \arg\max_\theta \mathbb{E}\{\log p(x_1 \mid \theta)\}$ [Kleijn & van der Vaart 2012, De Blasi & Walker 2012, Walker 2013, **H** & Miller 2019, Hoff & Wakefield 2021]

- This choice is somewhat arbitrary if model misspecified

Part 2: Prediction with High-Dimensional Models

- Part 1: Bayesian model selection may not be self-consistent (i.e., reproducible) if all models are misspecified

- But what about parameter estimation?

- Usual focus is on pseudo-true parameter (i.e., KL-minimizing parameter) $\theta^\star := \arg\max_\theta \mathbb{E}\{\log p(x_1 \mid \theta)\}$ [Kleijn & van der Vaart 2012, De Blasi & Walker 2012, Walker 2013, **H** & Miller 2019, Hoff & Wakefield 2021]

- This choice is somewhat arbitrary if model misspecified

- **We propose an alternative approach** (details later if there's time/interest):

Part 2: Prediction with High-Dimensional Models

- Part 1: Bayesian model selection may not be self-consistent (i.e., reproducible) if all models are misspecified

- But what about parameter estimation?

- Usual focus is on pseudo-true parameter (i.e., KL-minimizing parameter) $\theta^\star := \arg\max_\theta \mathbb{E}\{\log p(x_1 \mid \theta)\}$ [Kleijn & van der Vaart 2012, De Blasi & Walker 2012, Walker 2013, **H** & Miller 2019, Hoff & Wakefield 2021]

- This choice is somewhat arbitrary if model misspecified

- **We propose an alternative approach** (details later if there's time/interest):
  - Formalize a generally applicable self-consistency criterion based on overlap probabilities

Part 2: Prediction with High-Dimensional Models

- Part 1: Bayesian model selection may not be self-consistent (i.e., reproducible) if all models are misspecified

- But what about parameter estimation?

- Usual focus is on pseudo-true parameter (i.e., KL-minimizing parameter) $\theta^\star := \arg\max_\theta \mathbb{E}\{\log p(x_1 \mid \theta)\}$ [Kleijn & van der Vaart 2012, De Blasi & Walker 2012, Walker 2013, **H** & Miller 2019, Hoff & Wakefield 2021]

- This choice is somewhat arbitrary if model misspecified

- **We propose an alternative approach** (details later if there's time/interest):
  - Formalize a generally applicable self-consistency criterion based on overlap probabilities
  - The posterior can violate criterion (especially in high-dimensional settings)

Part 2: Prediction with High-Dimensional Models

- Part 1: Bayesian model selection may not be self-consistent (i.e., reproducible) if all models are misspecified

- But what about parameter estimation?

- Usual focus is on pseudo-true parameter (i.e., KL-minimizing parameter) $\theta^\star := \arg\max_\theta \mathbb{E}\{\log p(x_1 \mid \theta)\}$ [Kleijn & van der Vaart 2012, De Blasi & Walker 2012, Walker 2013, **H** & Miller 2019, Hoff & Wakefield 2021]

- This choice is somewhat arbitrary if model misspecified

- **We propose an alternative approach** (details later if there's time/interest):
  - Formalize a generally applicable self-consistency criterion based on overlap probabilities
  - The posterior can violate criterion (especially in high-dimensional settings)
  - The bagged posterior doesn't violate criterion

Part 3: Unsupervised Learning of *Bona Fide* Latent Structure

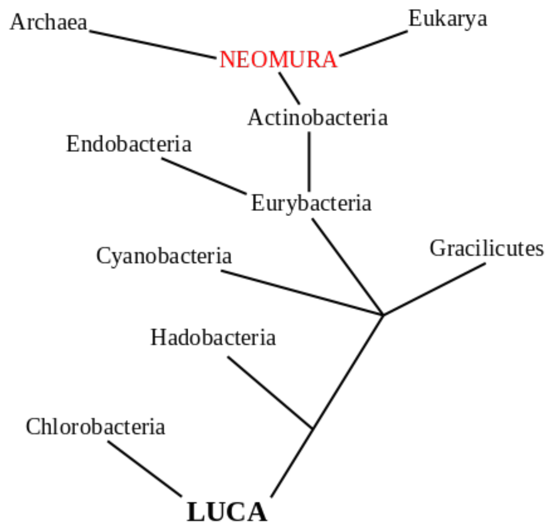
[Box & Hunter 1965, Box 1979, Cox 1990, Lehman 1990, Breiman 2001, Shmueli 2010, Miller & Dunson 2019]

Part 3: Unsupervised Learning of *Bona Fide* Latent Structure

- **Common task:** Unsupervised learning of **“latent structures”**

[Box & Hunter 1965, Box 1979, Cox 1990, Lehman 1990, Breiman 2001, Shmueli 2010, Miller & Dunson 2019]

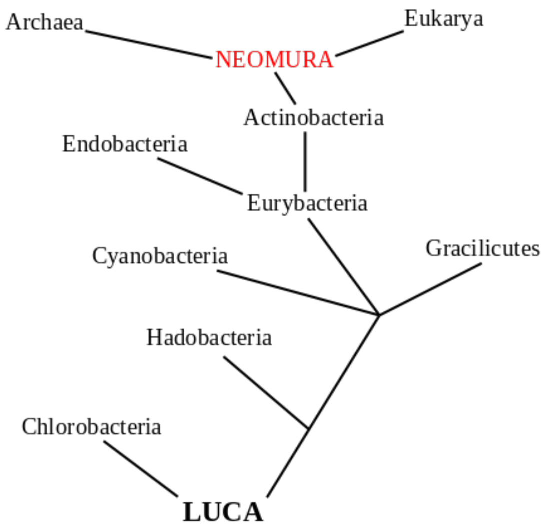
Part 3: Unsupervised Learning of *Bona Fide* Latent Structure

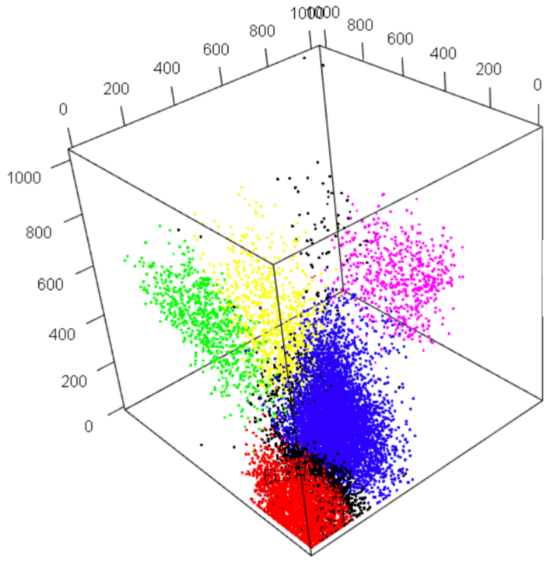


- **Common task:** Unsupervised learning of **“latent structures”**
    - **Phylogenetic inference** learns ancestries [Felsenstein 2004]

[Box & Hunter 1965, Box 1979, Cox 1990, Lehman 1990, Breiman 2001, Shmueli 2010, Miller & Dunson 2019]

Part 3: Unsupervised Learning of *Bona Fide* Latent Structure

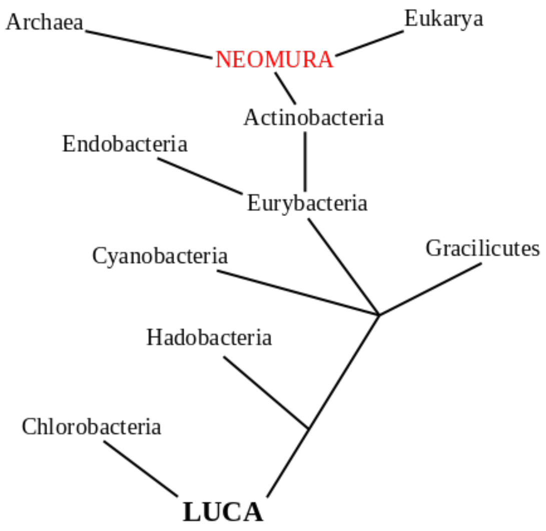


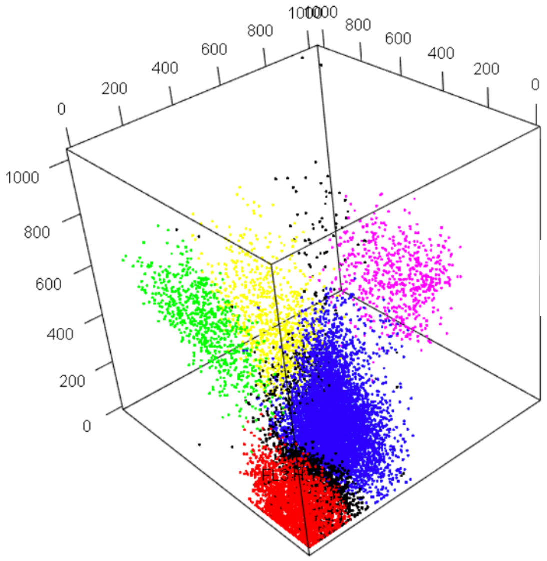


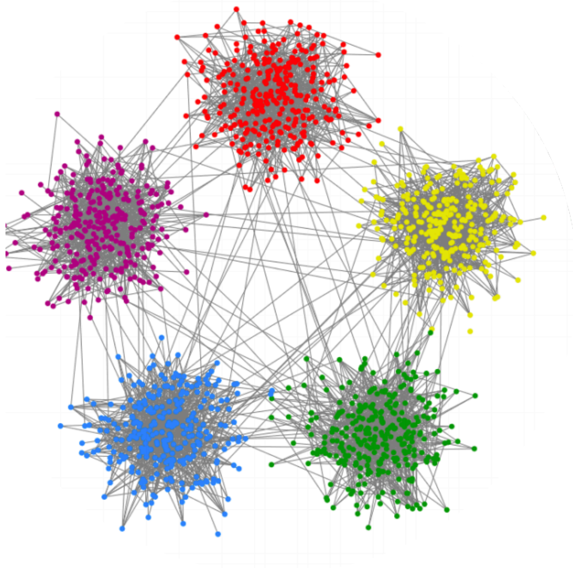
- **Common task:** Unsupervised learning of **"latent structures"**
    - **Phylogenetic inference** learns ancestries [Felsenstein 2004]
    - **Clustering** learns cell types [Lee & McLachlan 2014]

[Box & Hunter 1965, Box 1979, Cox 1990, Lehman 1990, Breiman 2001, Shmueli 2010, Miller & Dunson 2019]

Part 3: Unsupervised Learning of *Bona Fide* Latent Structure



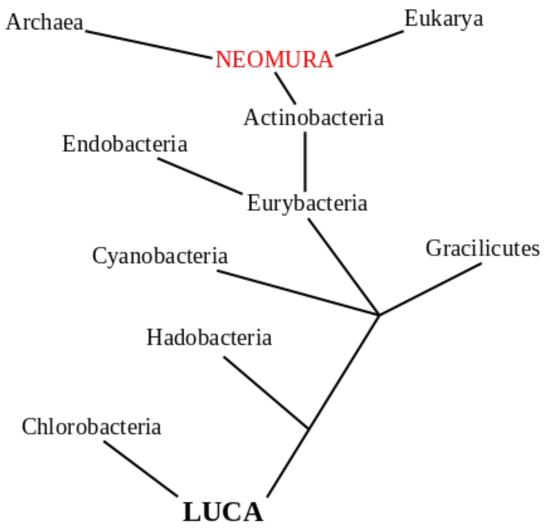


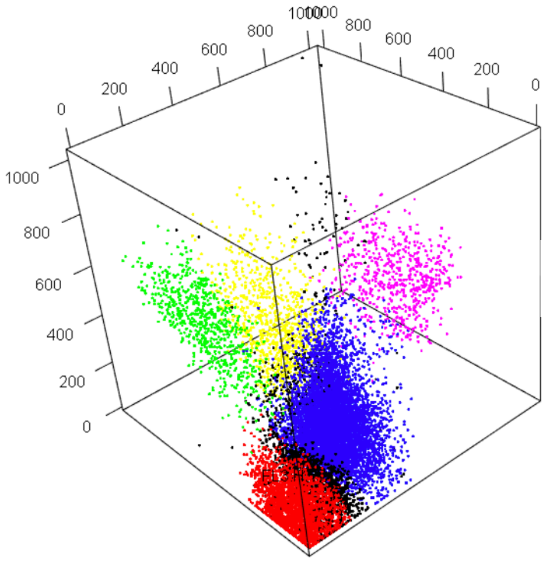


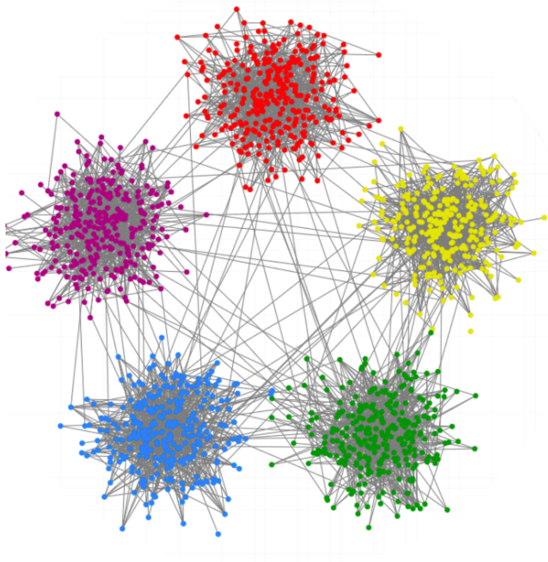
- **Common task:** Unsupervised learning of **“latent structures”**
  - **Phylogenetic inference** learns ancestries [Felsenstein 2004]
  - **Clustering** learns cell types [Lee & McLachlan 2014]
  - **Stochastic block model** learns communities [Abbe 2018]

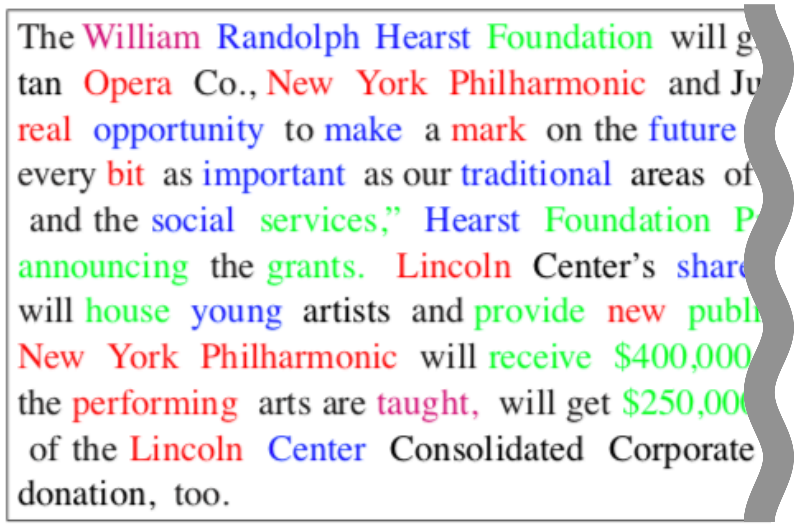
[Box & Hunter 1965, Box 1979, Cox 1990, Lehman 1990, Breiman 2001, Shmueli 2010, Miller & Dunson 2019]

Part 3: Unsupervised Learning of *Bona Fide* Latent Structure







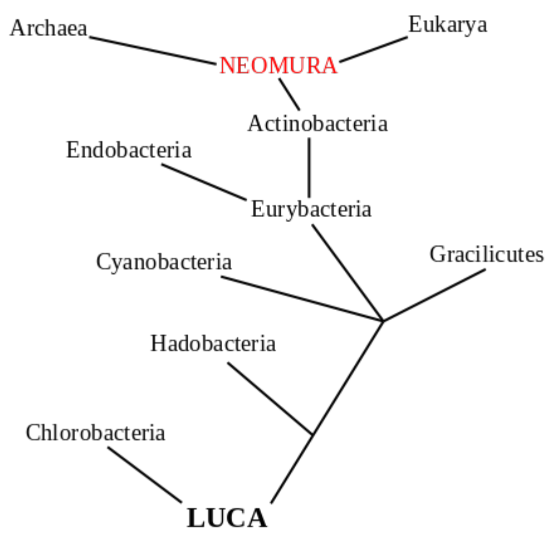


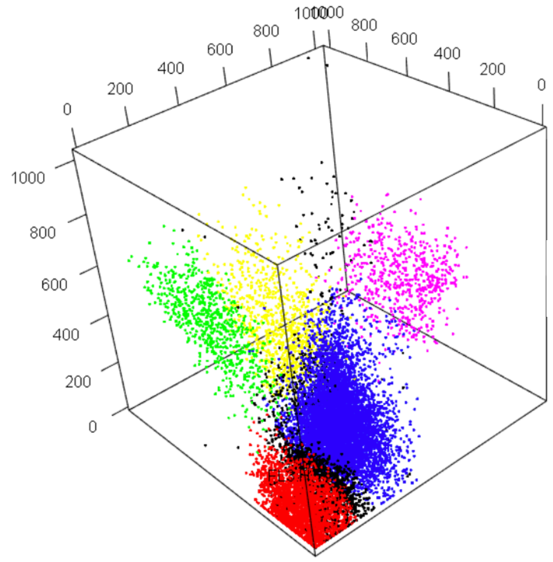
The William Randolph Hearst Foundation will g
tan Opera Co., New York Philharmonic and Ju
real opportunity to make a mark on the future
every bit as important as our traditional areas of
and the social services," Hearst Foundation P
announcing the grants. Lincoln Center's share
will house young artists and provide new publi
New York Philharmonic will receive $400,000
the performing arts are taught, will get $250,000
of the Lincoln Center Consolidated Corporate
donation, too.

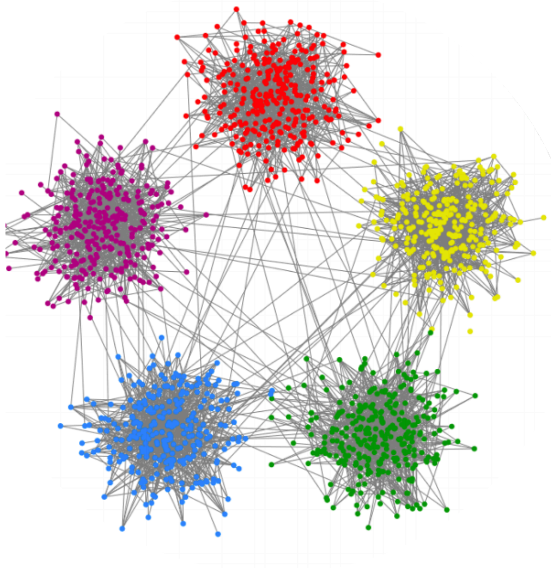
- **Common task:** Unsupervised learning of **"latent structures"**
  - **Phylogenetic inference** learns ancestries [Felsenstein 2004]
  - **Clustering** learns cell types [Lee & McLachlan 2014]
  - **Stochastic block model** learns communities [Abbe 2018]
  - **Latent Dirichlet allocation** learns document topics [Blei et al. 2003]

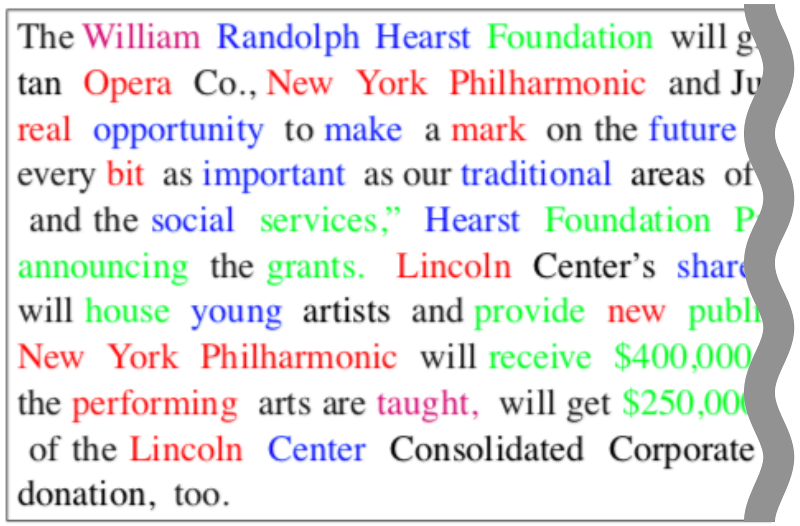
[Box & Hunter 1965, Box 1979, Cox 1990, Lehman 1990, Breiman 2001, Shmueli 2010, Miller & Dunson 2019]

Part 3: Unsupervised Learning of *Bona Fide* Latent Structure









The William Randolph Hearst Foundation will g
tan Opera Co., New York Philharmonic and Ju
real opportunity to make a mark on the future
every bit as important as our traditional areas of
and the social services," Hearst Foundation P
announcing the grants. Lincoln Center's shar
will house young artists and provide new publ
New York Philharmonic will receive $400,000
the performing arts are taught, will get $250,00
of the Lincoln Center Consolidated Corporate
donation, too.

- **Common task:** Unsupervised learning of **"latent structures"**

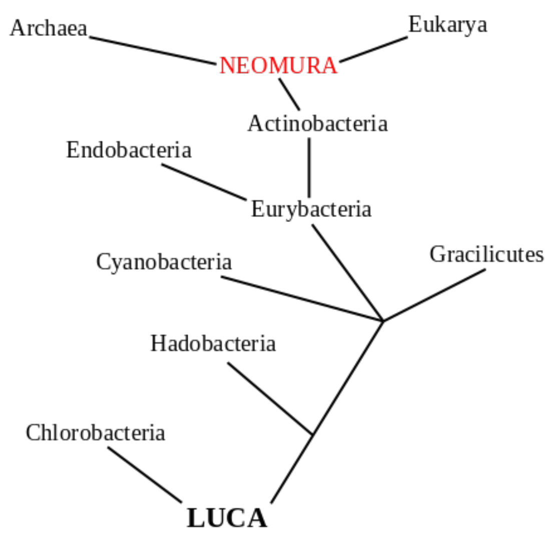
explanatory

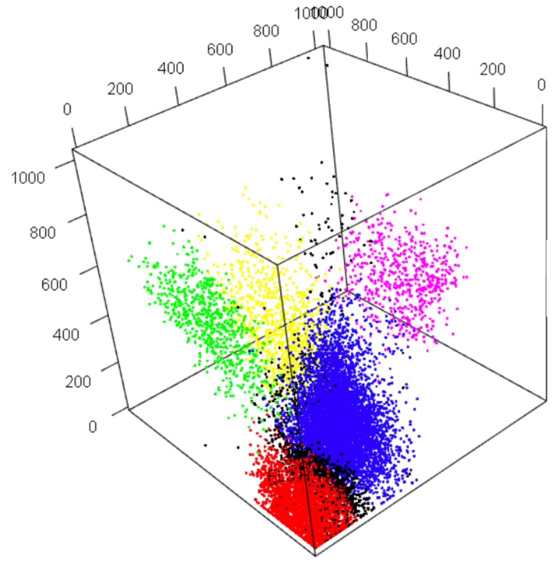
- **Phylogenetic inference** learns ancestries [Felsenstein 2004]
- **Clustering** learns cell types [Lee & McLachlan 2014]
- **Stochastic block model** learns communities [Abbe 2018]
- **Latent Dirichlet allocation** learns document topics [Blei et al. 2003]

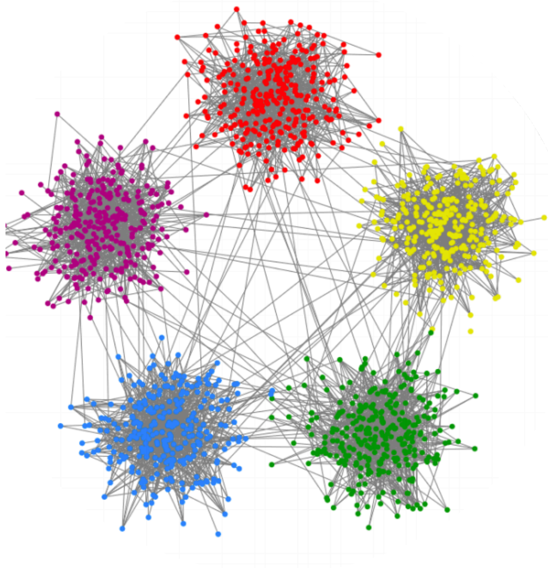
exploratory

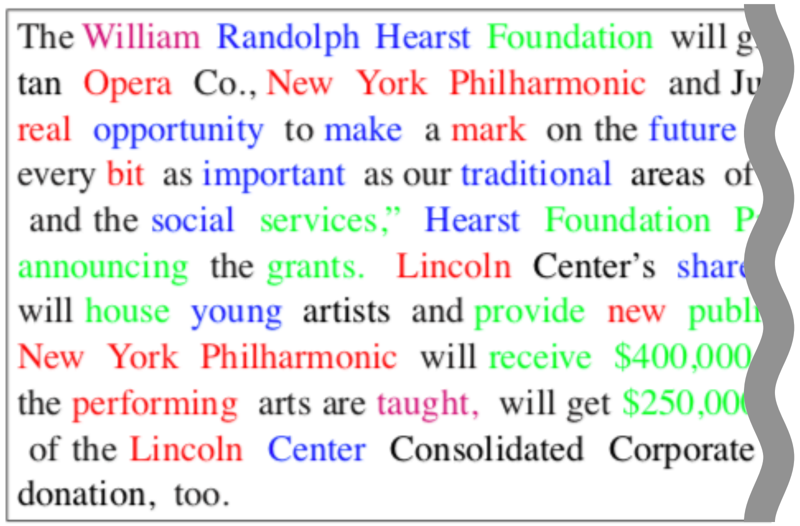
[Box & Hunter 1965, Box 1979, Cox 1990, Lehman 1990, Breiman 2001, Shmueli 2010, Miller & Dunson 2019]

Part 3: Unsupervised Learning of *Bona Fide* Latent Structure









The William Randolph Hearst Foundation will g
tan Opera Co., New York Philharmonic and Ju
real opportunity to make a mark on the future
every bit as important as our traditional areas of
and the social services," Hearst Foundation P
announcing the grants. Lincoln Center's shar
will house young artists and provide new publ
New York Philharmonic will receive $400,000
the performing arts are taught, will get $250,00
of the Lincoln Center Consolidated Corporate
donation, too.

- **Common task:** Unsupervised learning of **"latent structures"**

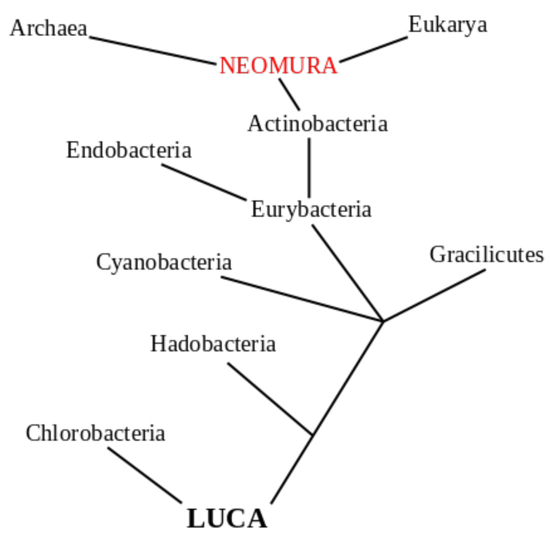
explanatory

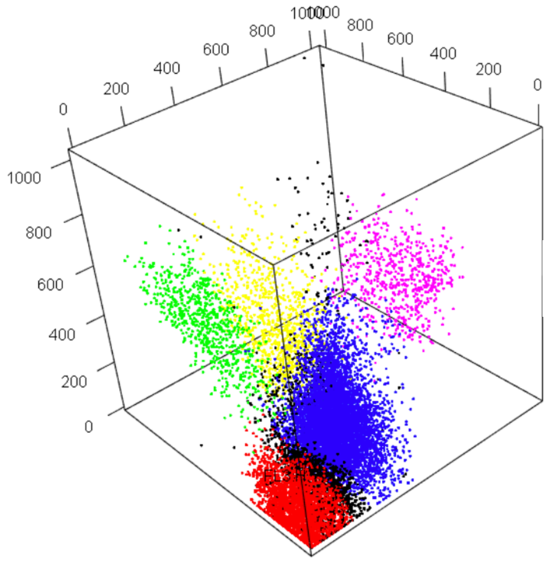
- **Phylogenetic inference** learns ancestries [Felsenstein 2004]
- **Clustering** learns cell types [Lee & McLachlan 2014]
- **Stochastic block model** learns communities [Abbe 2018]
- **Latent Dirichlet allocation** learns document topics [Blei et al. 2003]

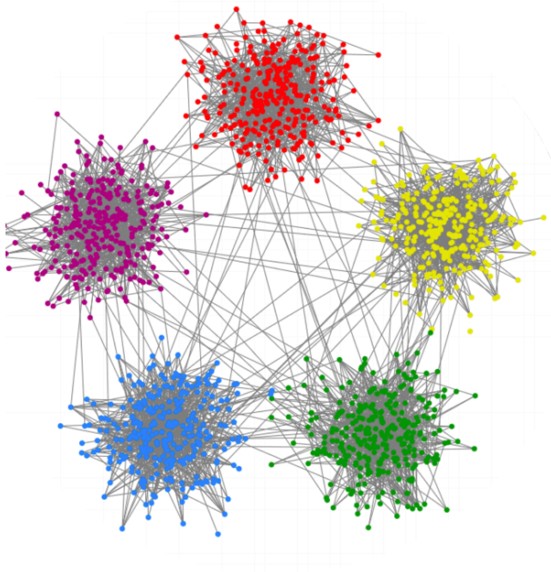
exploratory

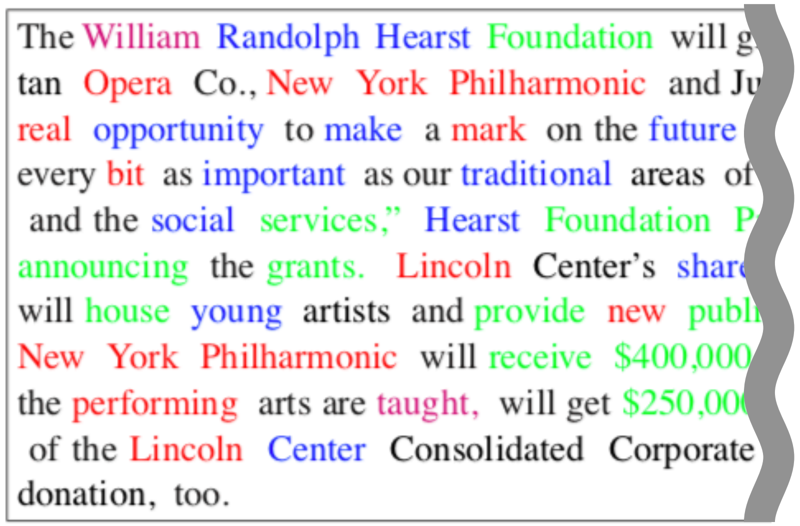
[Box & Hunter 1965, Box 1979, Cox 1990, Lehman 1990, Breiman 2001, Shmueli 2010, Miller & Dunson 2019]

Part 3: Unsupervised Learning of *Bona Fide* Latent Structure









The William Randolph Hearst Foundation will g
tan Opera Co., New York Philharmonic and Ju
real opportunity to make a mark on the future
every bit as important as our traditional areas of
and the social services," Hearst Foundation P
announcing the grants. Lincoln Center's share
will house young artists and provide new publ
New York Philharmonic will receive $400,000
the performing arts are taught, will get $250,00
of the Lincoln Center Consolidated Corporate
donation, too.

- **Common task:** Unsupervised learning of **"latent structures"**

explanatory

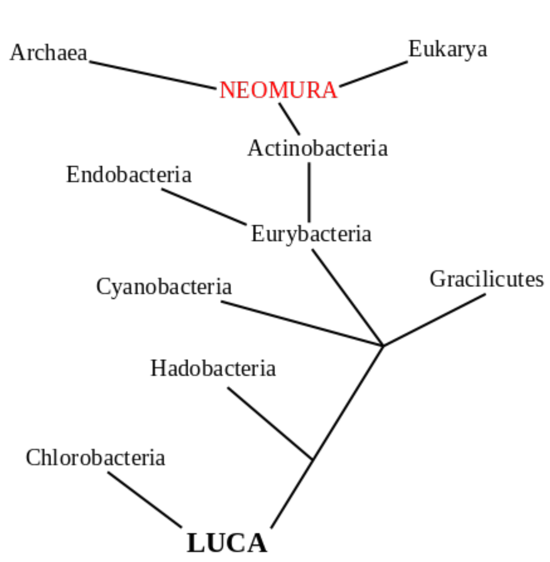
exploratory

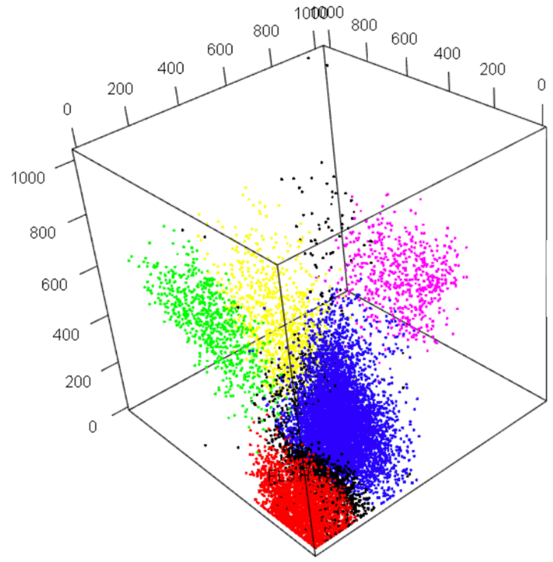
- **Phylogenetic inference** learns ancestries [Felsenstein 2004]
- **Clustering** learns cell types [Lee & McLachlan 2014]
- **Stochastic block model** learns communities [Abbe 2018]
- **Latent Dirichlet allocation** learns document topics [Blei et al. 2003]

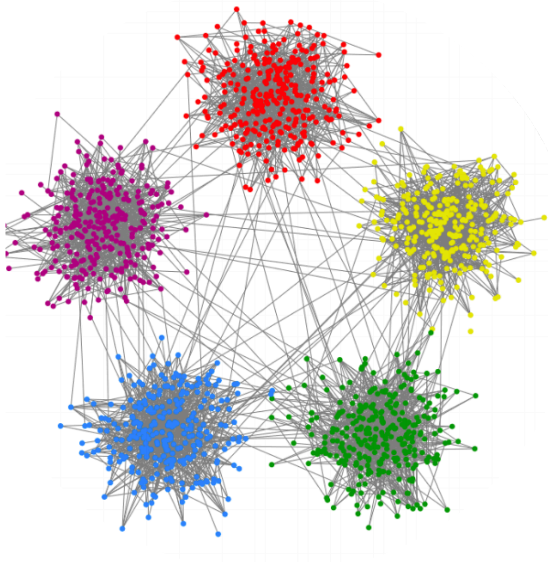
- **Challenge:** how do we recover "true" latent structure when the model is wrong?

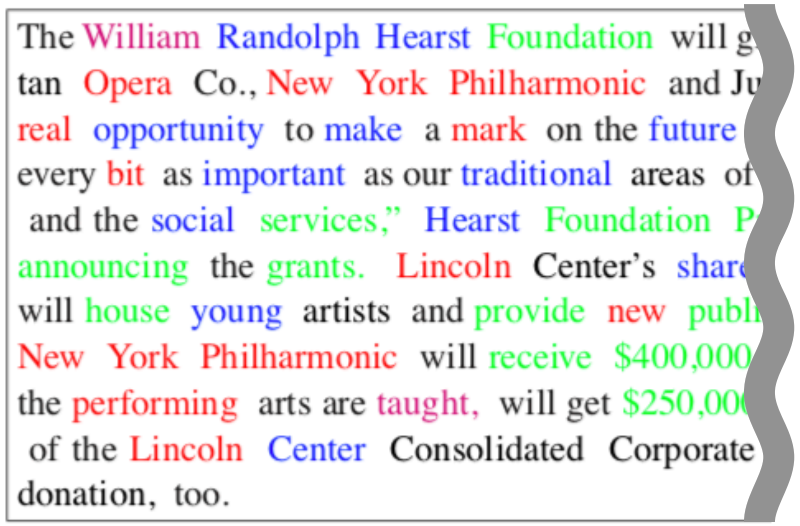
[Box & Hunter 1965, Box 1979, Cox 1990, Lehman 1990, Breiman 2001, Shmueli 2010, Miller & Dunson 2019]

Part 3: Unsupervised Learning of *Bona Fide* Latent Structure









The William Randolph Hearst Foundation will g
tan Opera Co., New York Philharmonic and Ju
real opportunity to make a mark on the future
every bit as important as our traditional areas of
and the social services," Hearst Foundation P
announcing the grants. Lincoln Center's share
will house young artists and provide new publi
New York Philharmonic will receive $400,000
the performing arts are taught, will get $250,00
of the Lincoln Center Consolidated Corporate
donation, too.

- **Common task:** Unsupervised learning of **"latent structures"**

explanatory

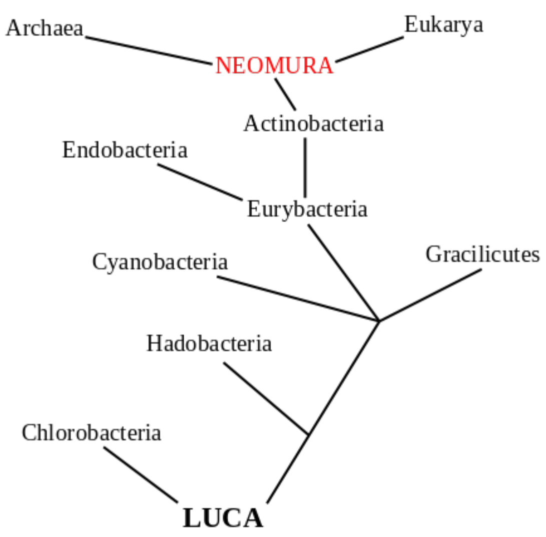
exploratory

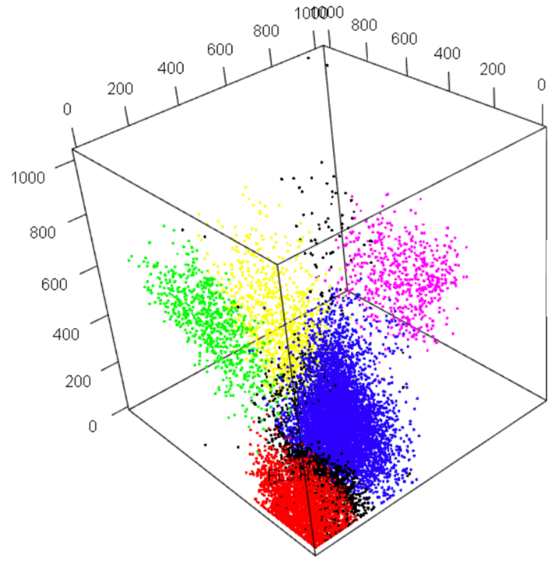
- **Phylogenetic inference** learns ancestries [Felsenstein 2004]
- **Clustering** learns cell types [Lee & McLachlan 2014]
- **Stochastic block model** learns communities [Abbe 2018]
- **Latent Dirichlet allocation** learns document topics [Blei et al. 2003]

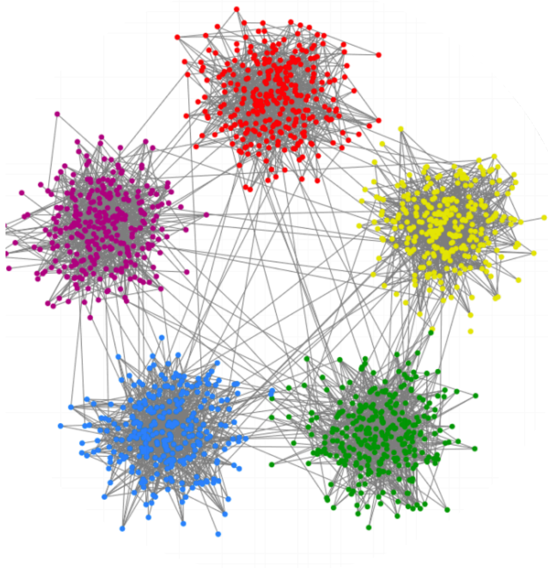
- **Challenge:** how do we recover "true" latent structure when the model is wrong?
- **This part:** learning *bona fide* clusters with mixture models

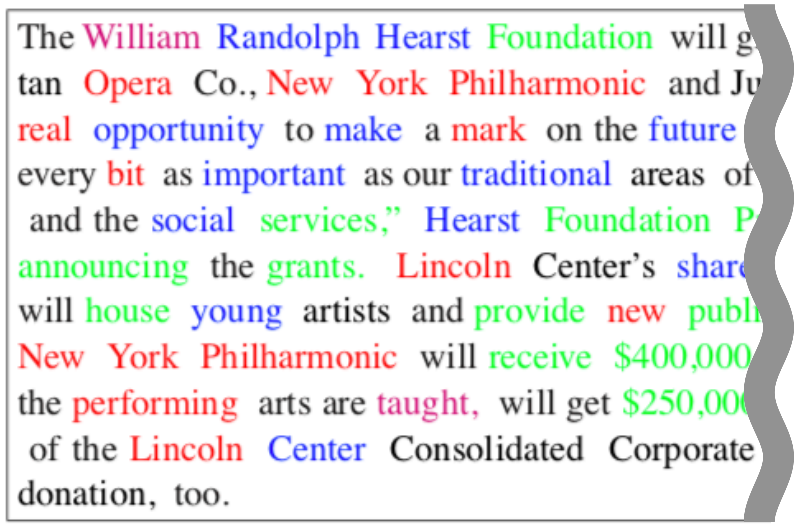
[Box & Hunter 1965, Box 1979, Cox 1990, Lehman 1990, Breiman 2001, Shmueli 2010, Miller & Dunson 2019]

Part 3: Unsupervised Learning of *Bona Fide* Latent Structure









The William Randolph Hearst Foundation will g
tan Opera Co., New York Philharmonic and Ju
real opportunity to make a mark on the future
every bit as important as our traditional areas of
and the social services," Hearst Foundation P
announcing the grants. Lincoln Center's shar
will house young artists and provide new publi
New York Philharmonic will receive $400,000
the performing arts are taught, will get $250,00
of the Lincoln Center Consolidated Corporate
donation, too.

- **Common task:** Unsupervised learning of **"latent structures"**

explanatory

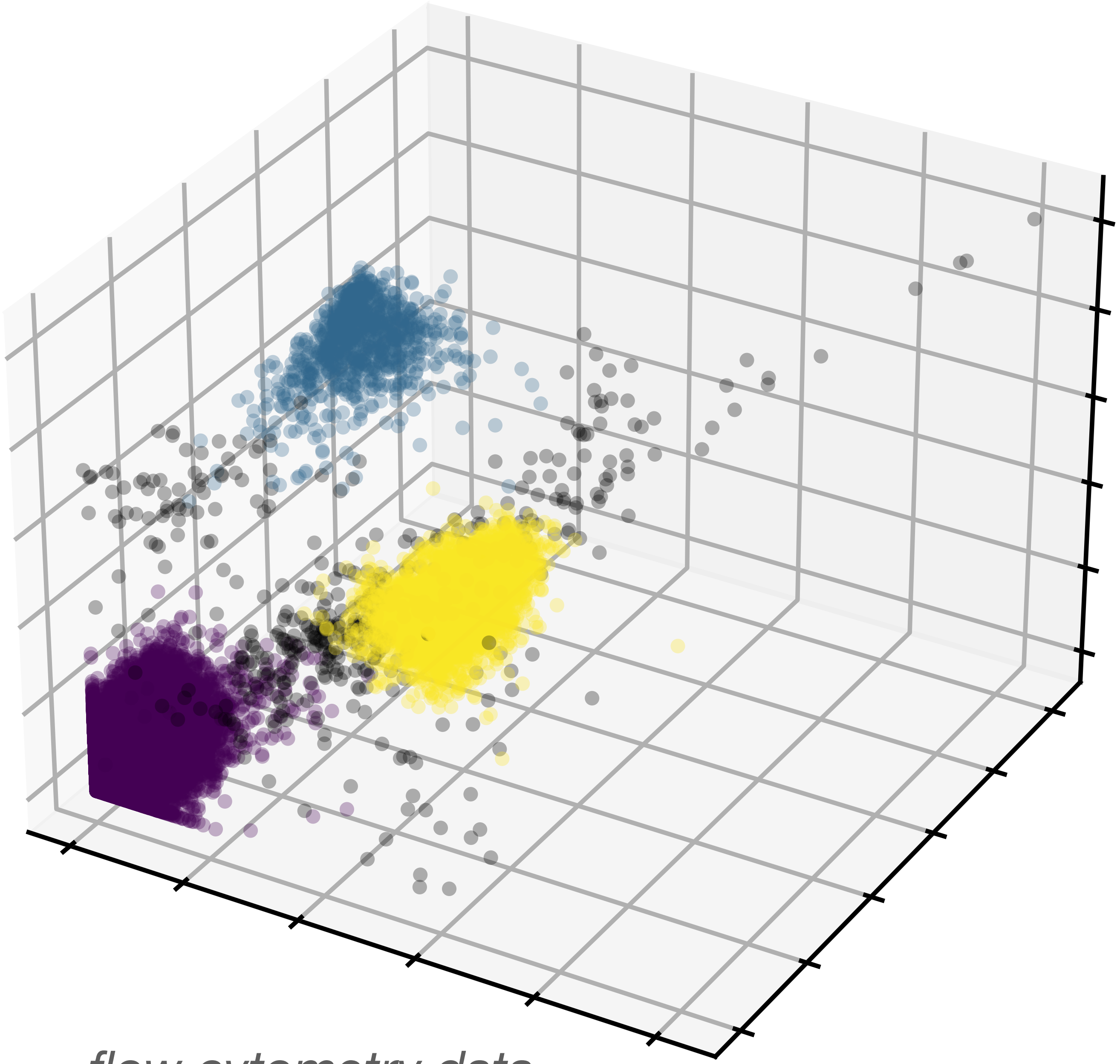
- **Phylogenetic inference** learns ancestries [Felsenstein 2004]
- **Clustering** learns cell types [Lee & McLachlan 2014]
- **Stochastic block model** learns communities [Abbe 2018]
- **Latent Dirichlet allocation** learns document topics [Blei et al. 2003]

exploratory

- **Challenge:** how do we recover "true" latent structure when the model is wrong?
- **This part:** learning *bona fide* clusters with mixture models
- **Key ideas:** use known causal structure and domain knowledge

[Box & Hunter 1965, Box 1979, Cox 1990, Lehman 1990, Breiman 2001, Shmueli 2010, Miller & Dunson 2019]

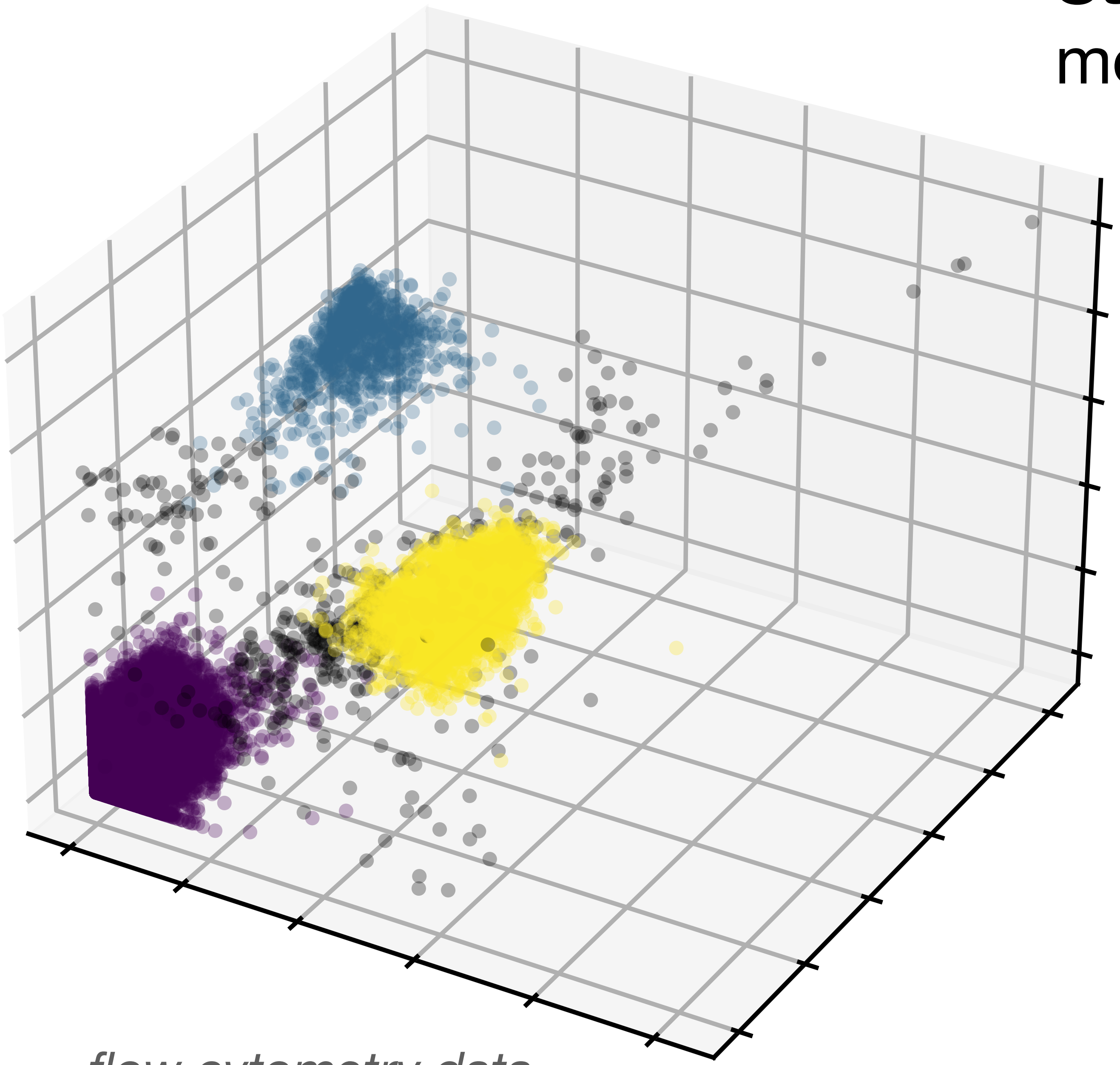
Clustering cells by type



flow cytometry data

[cf. Miller & Dunson 2018, Cai, Campbell & Broderick 2021]

Clustering cells by type

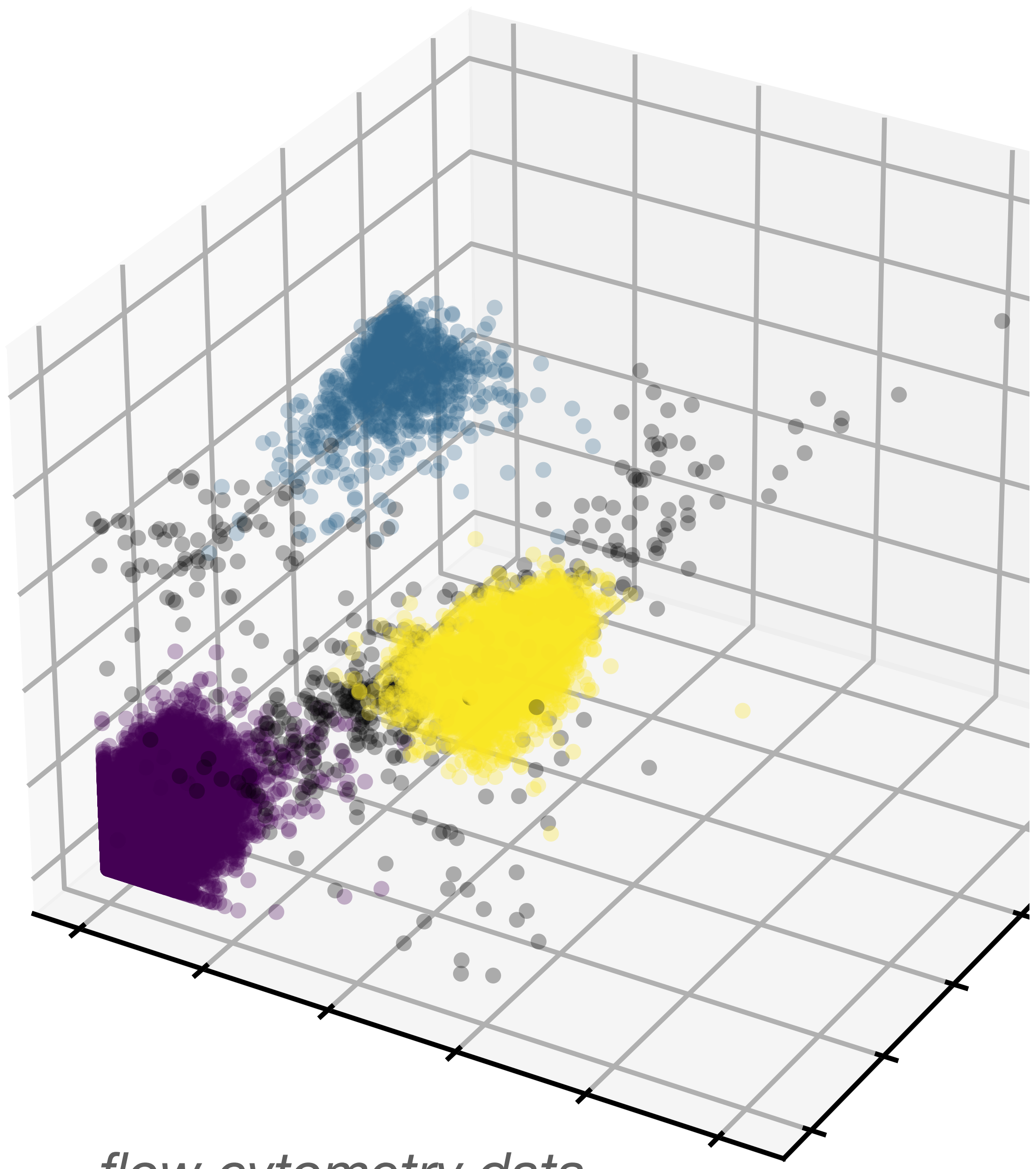


flow cytometry data

Standard approach: Gaussian mixture model with prior on # of components k

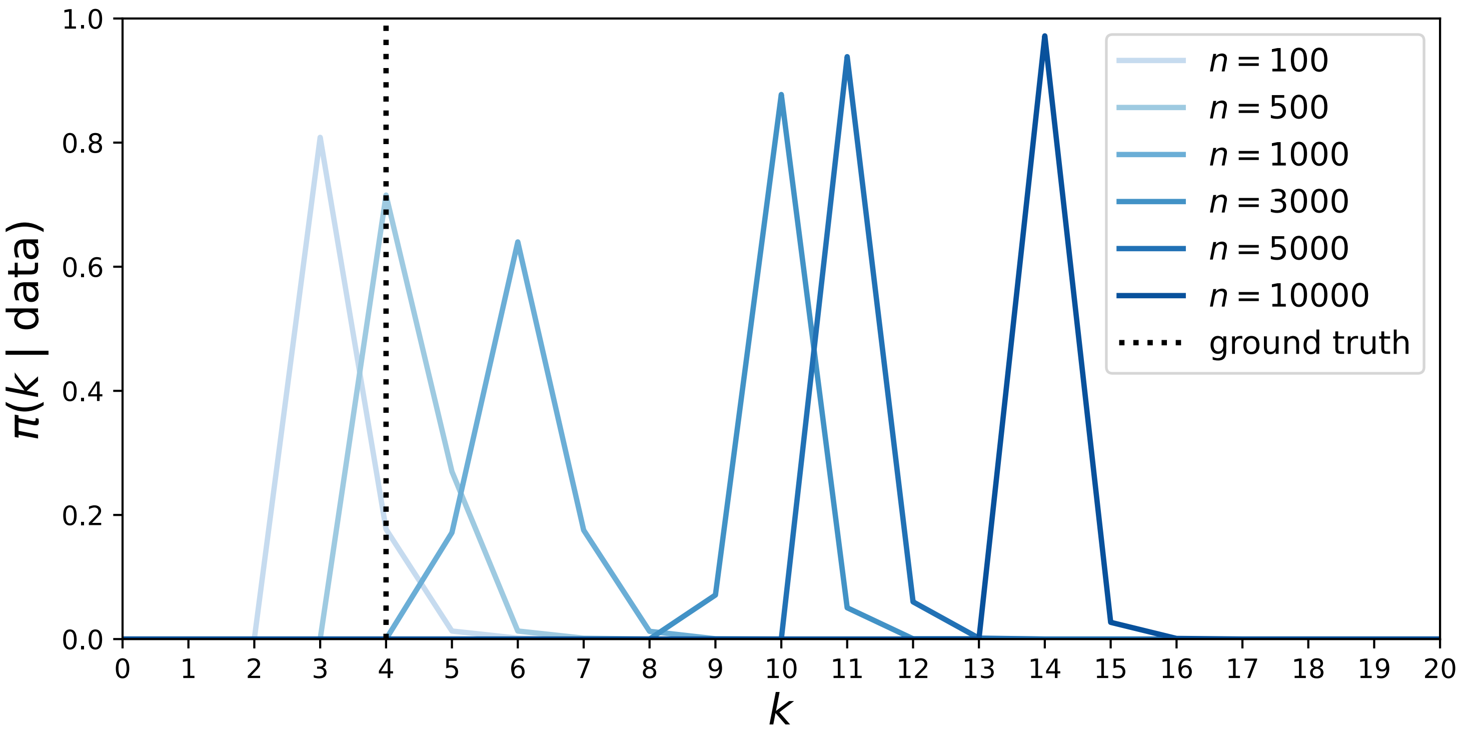
[cf. Miller & Dunson 2018, Cai, Campbell & Broderick 2021]

Clustering cells by type



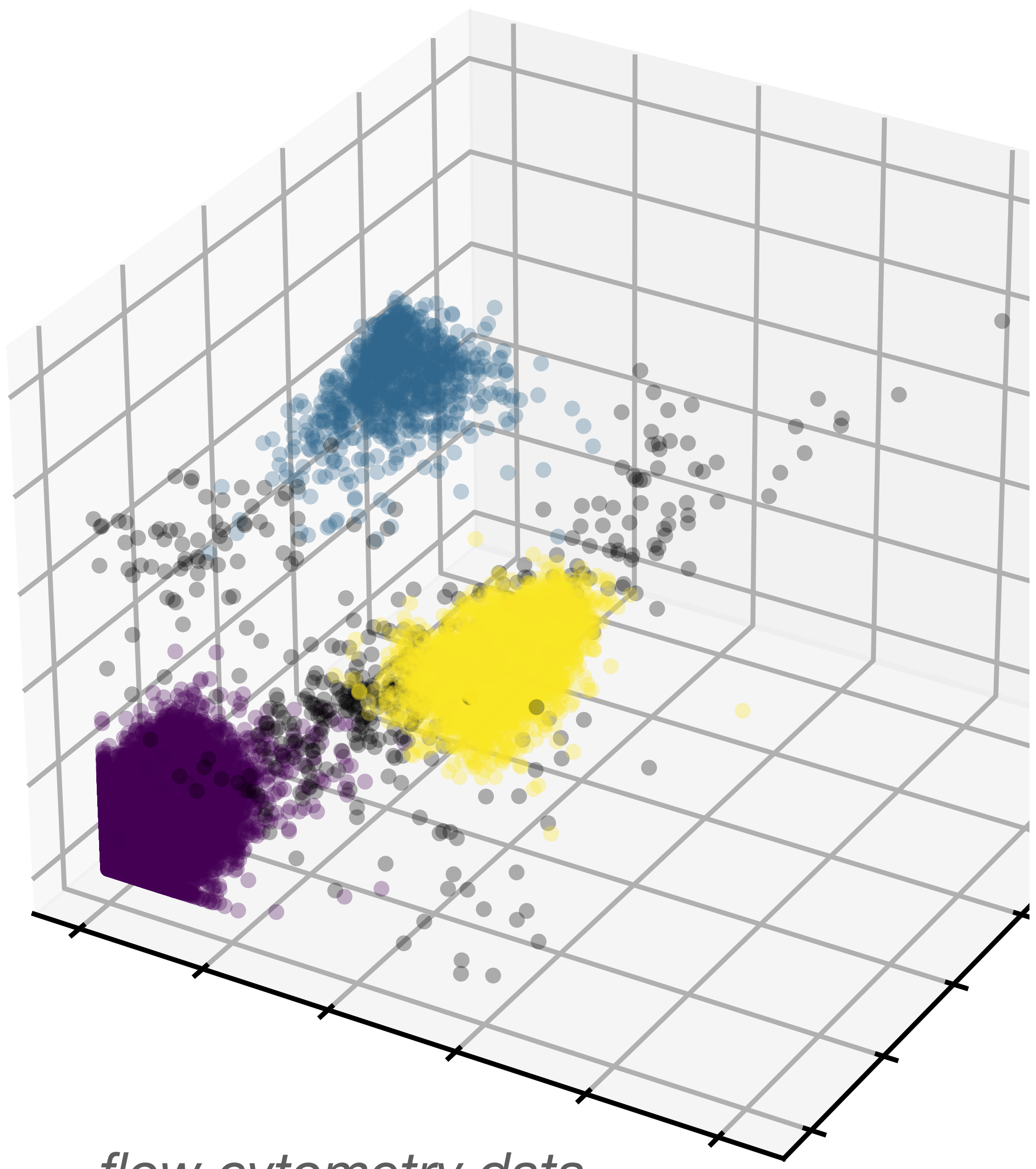
flow cytometry data

Standard approach: Gaussian mixture model with prior on # of components k



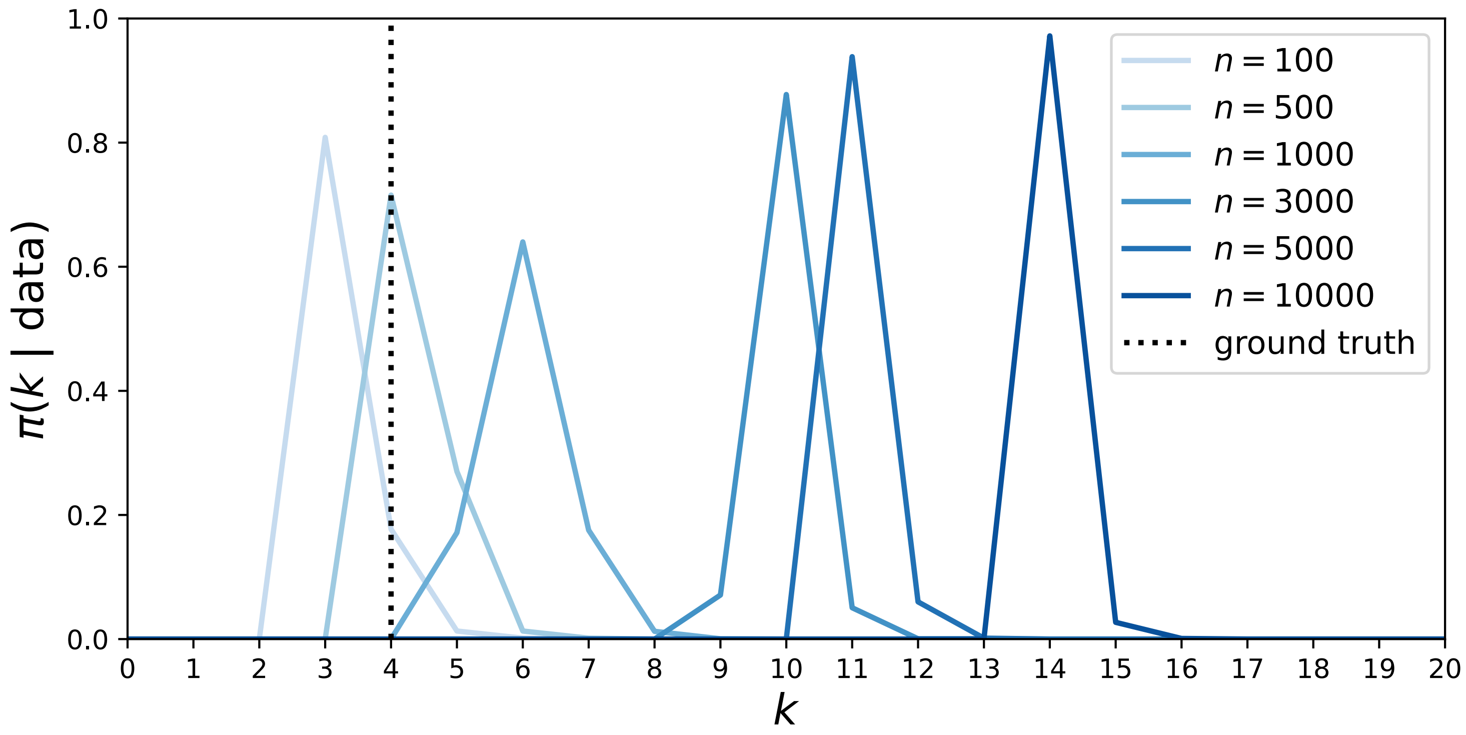
[cf. Miller & Dunson 2018, Cai, Campbell & Broderick 2021]

Clustering cells by type



flow cytometry data

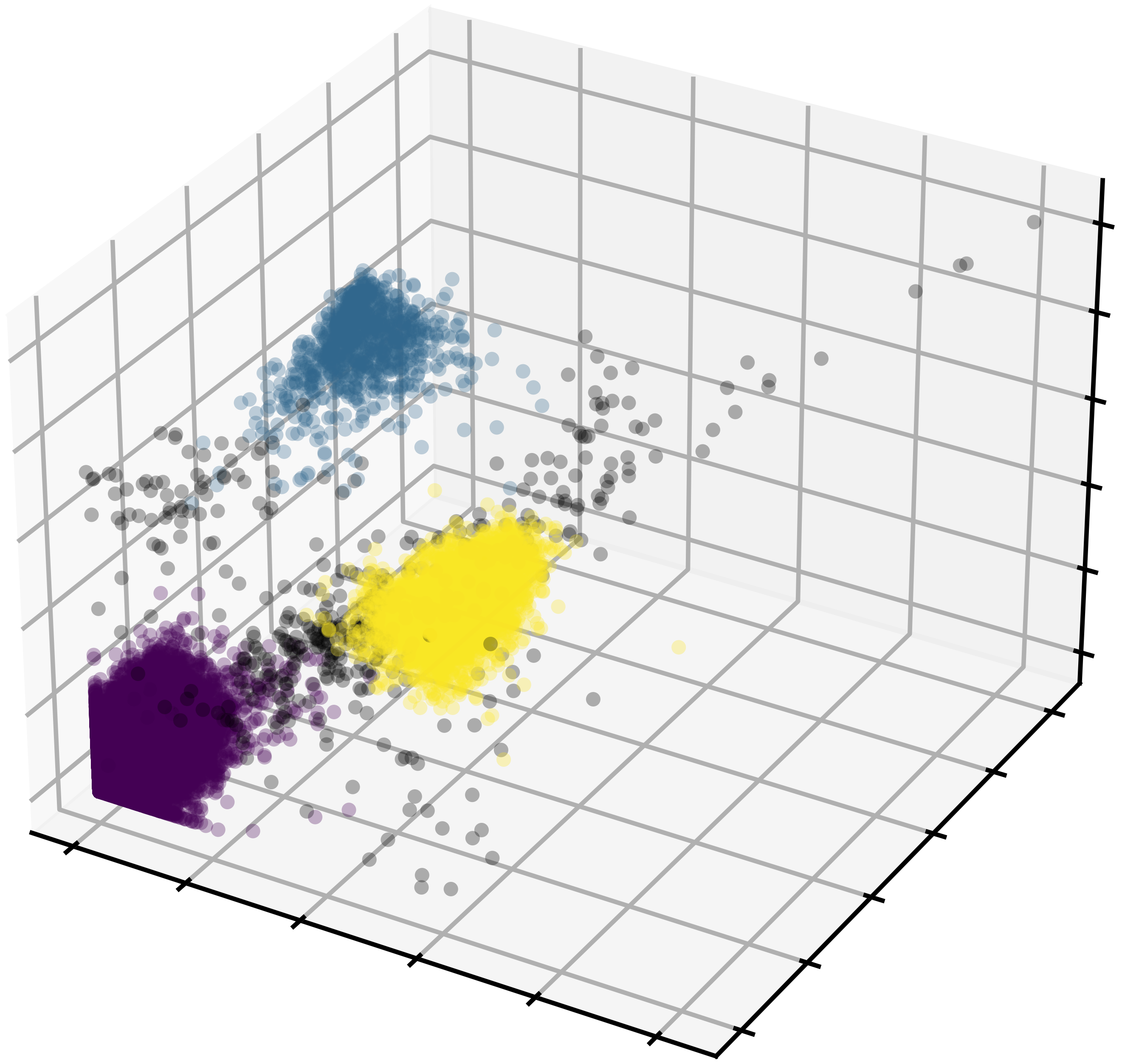
Standard approach: Gaussian mixture model with prior on # of components k



More data, more contradictions!

[cf. Miller & Dunson 2018, Cai, Campbell & Broderick 2021]

Clustering for type discovery

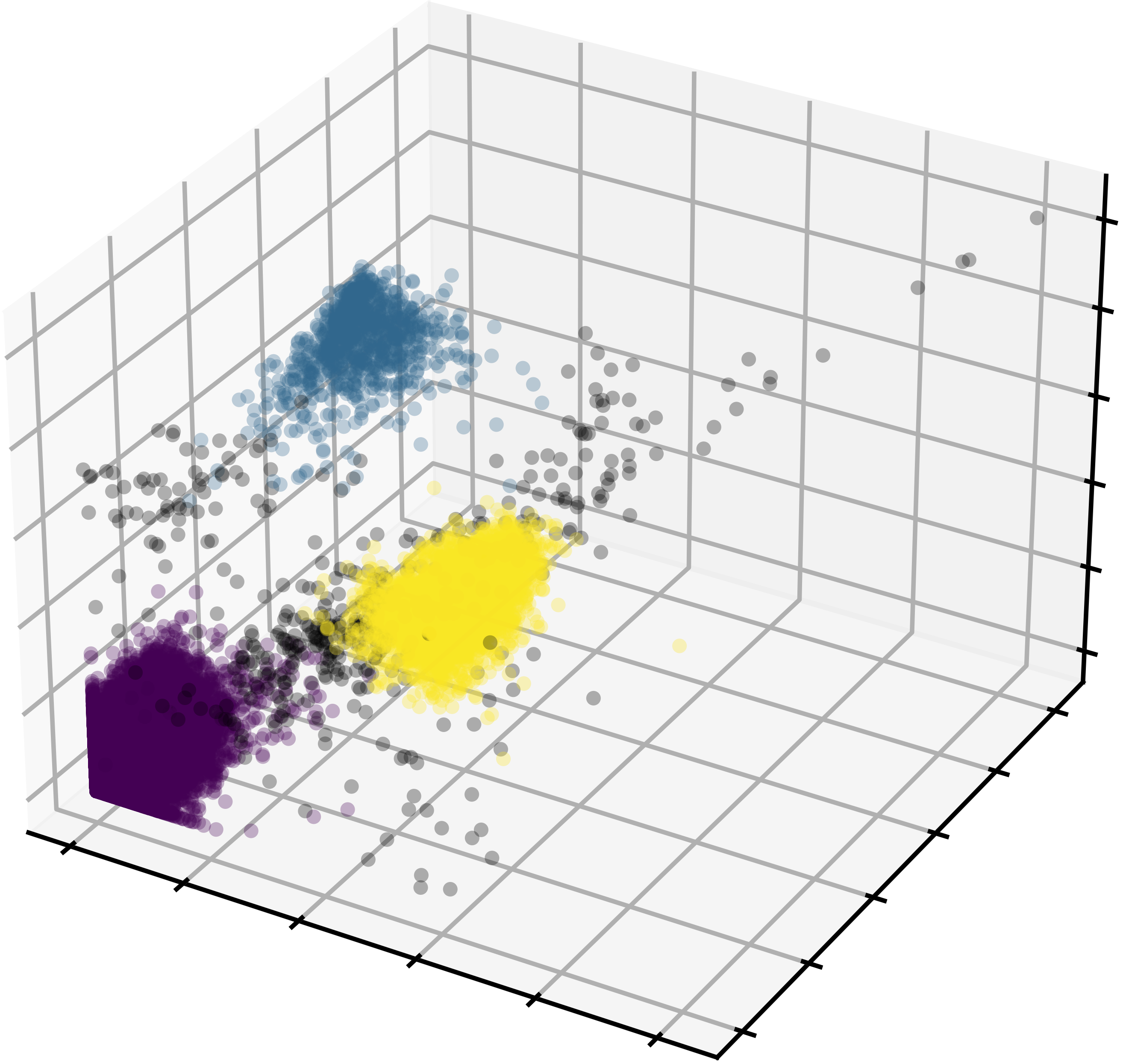


Clustering for type discovery

- Data-generating process:

$$x_n \sim P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$$

  - Each component is a meaningful ***type***



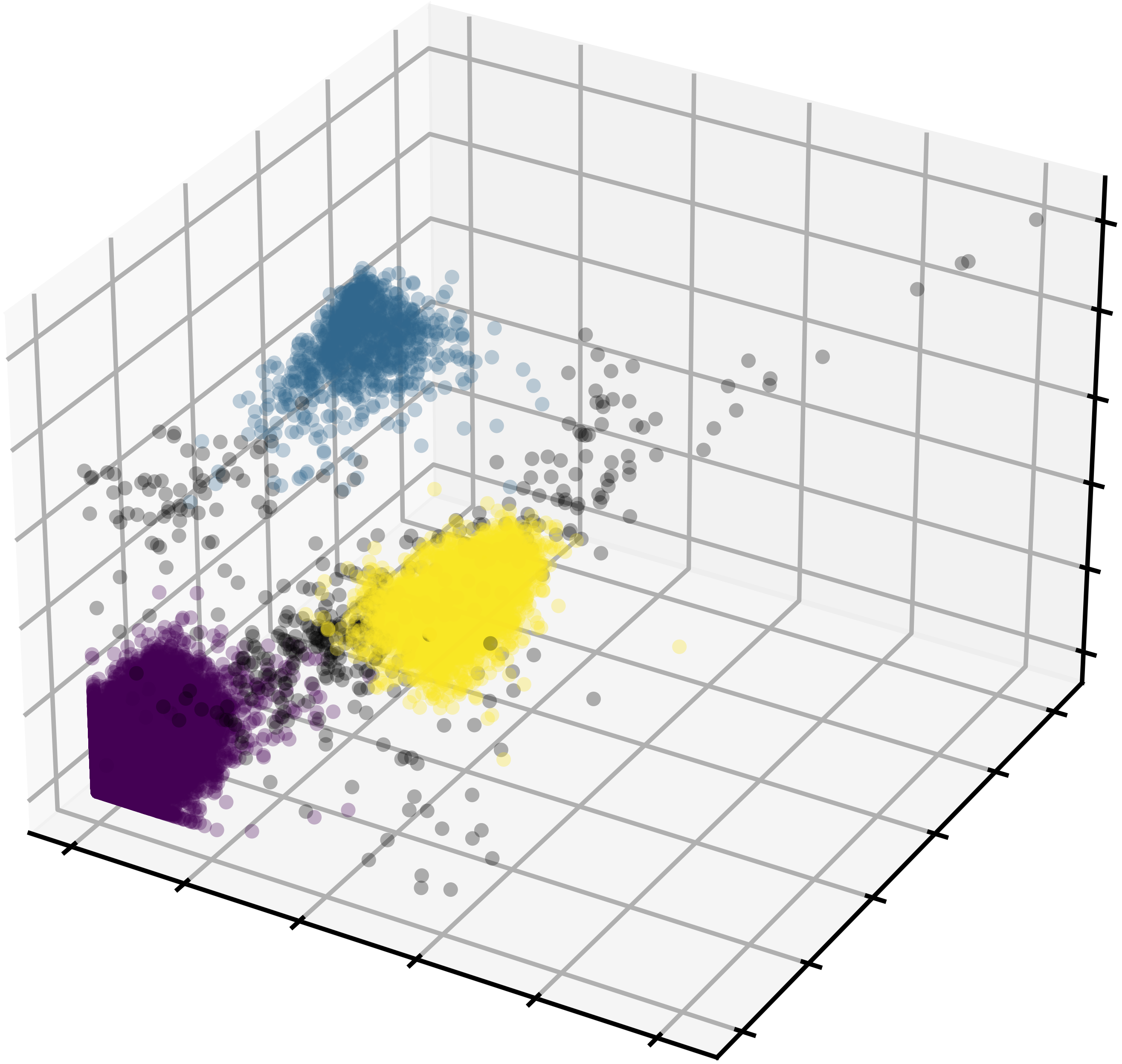
Clustering for type discovery

- Data-generating process:

$$x_n \sim P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$$

  - Each component is a meaningful ***type***

- **Goal:** discover these types
  - Determine $K_\circ$



Clustering for type discovery

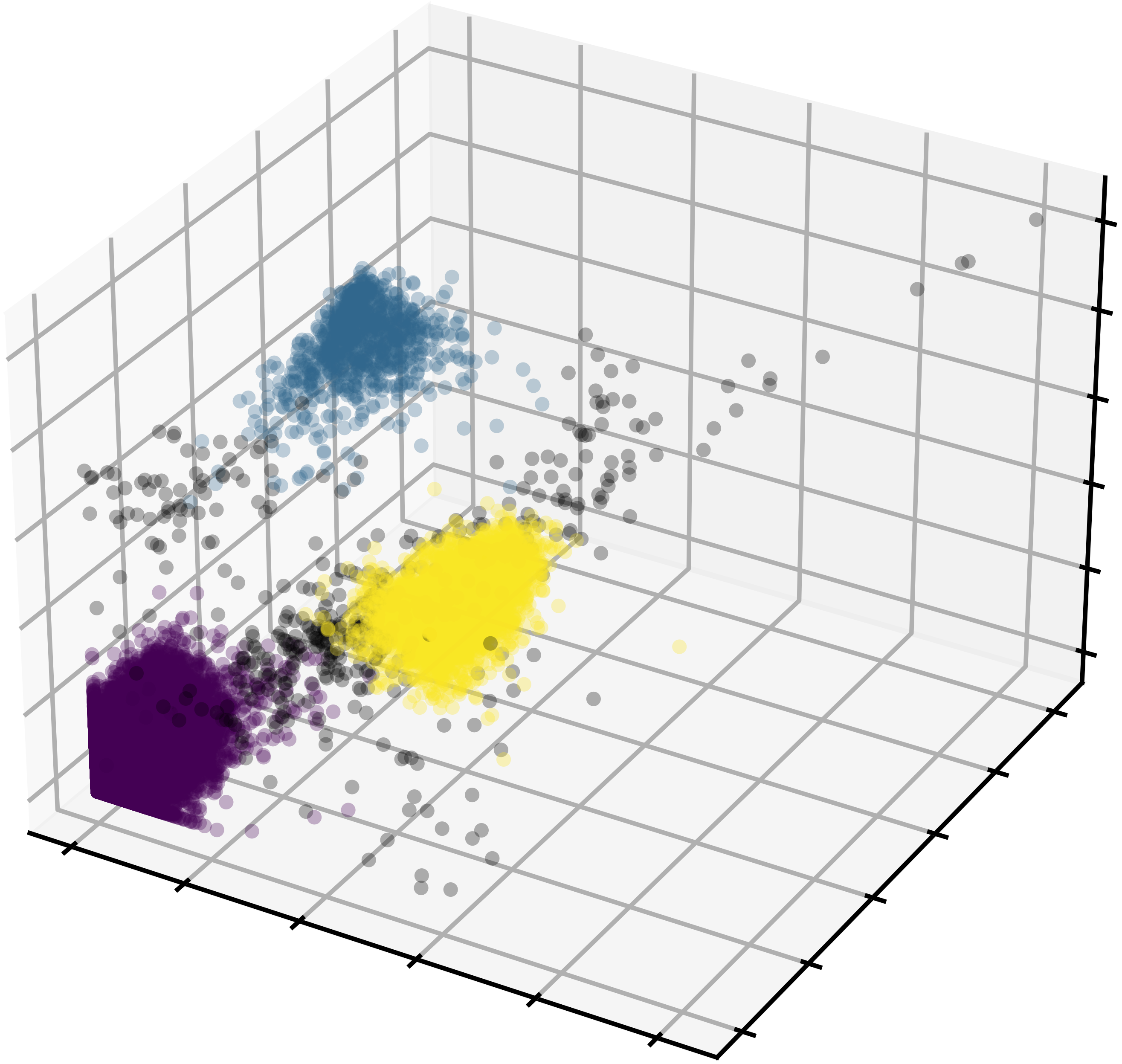
- Data-generating process:

$$x_n \sim P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$$

component distribution

  - Each component is a meaningful ***type***

- **Goal:** discover these types
  - Determine $K_\circ$



Clustering for type discovery

- Data-generating process:

$$x_n \sim P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} \boxed{P_{\circ k}}$$

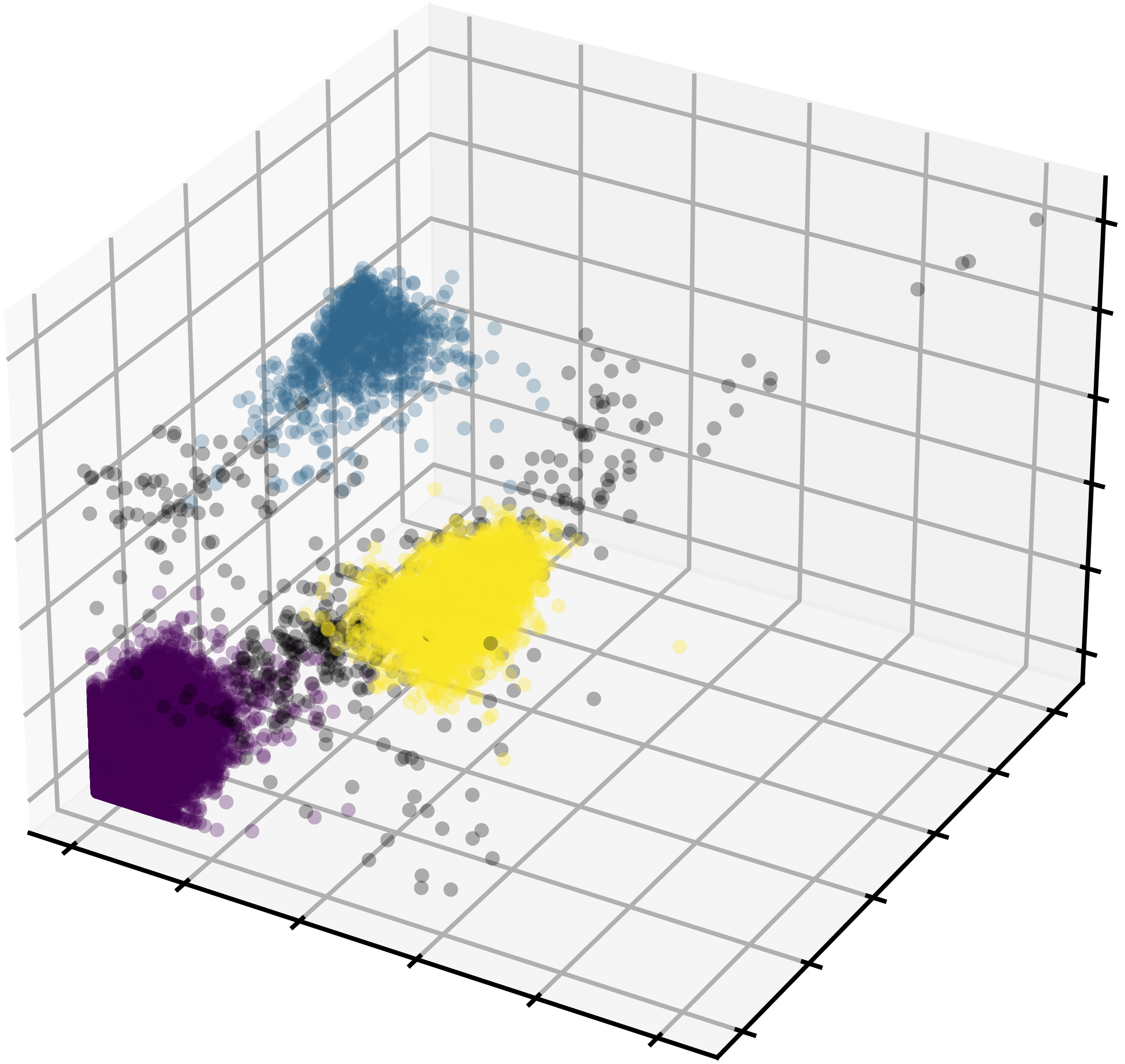
component distribution

  - Each component is a meaningful ***type***

- **Goal:** discover these types
  - Determine $K_\circ$

- Assumed mixture model:

$$x_n \sim P_\theta := \sum_{k=1}^{K} \pi_k F_{\phi_k}$$



Clustering for type discovery

- Data-generating process:

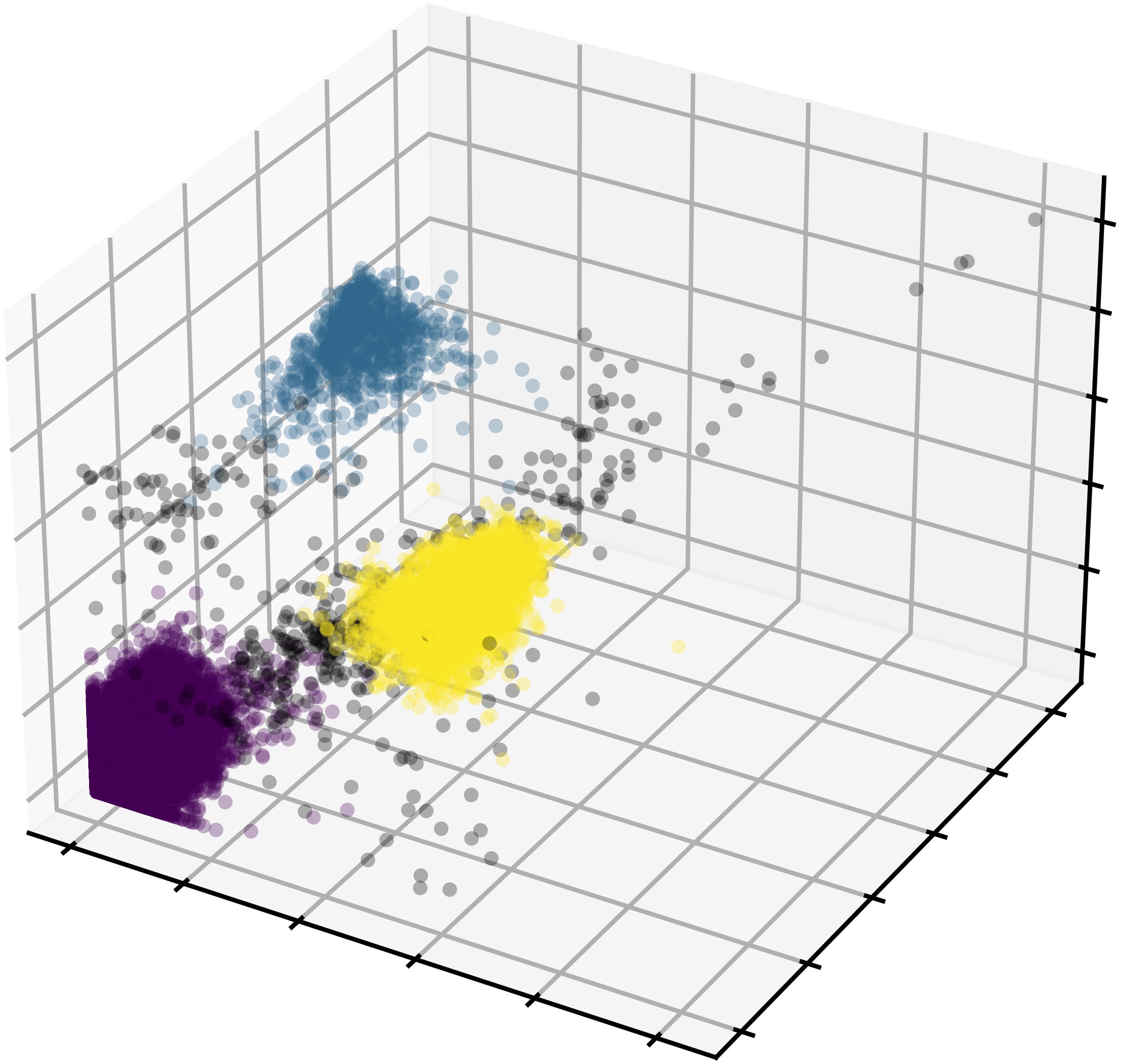
$$x_n \sim P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} \boxed{P_{\circ k}} \leftarrow \text{component distribution}$$

  - Each component is a meaningful ***type***

- **Goal:** discover these types
  - Determine $K_\circ$

- Assumed mixture model:

$$x_n \sim P_\theta := \sum_{k=1}^{K} \pi_k F_{\boxed{\phi_k}} \leftarrow \text{component parameter}$$



Clustering for type discovery

- Data-generating process:

$$x_n \sim P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} \boxed{P_{\circ k}} \longleftarrow \text{component distribution}$$

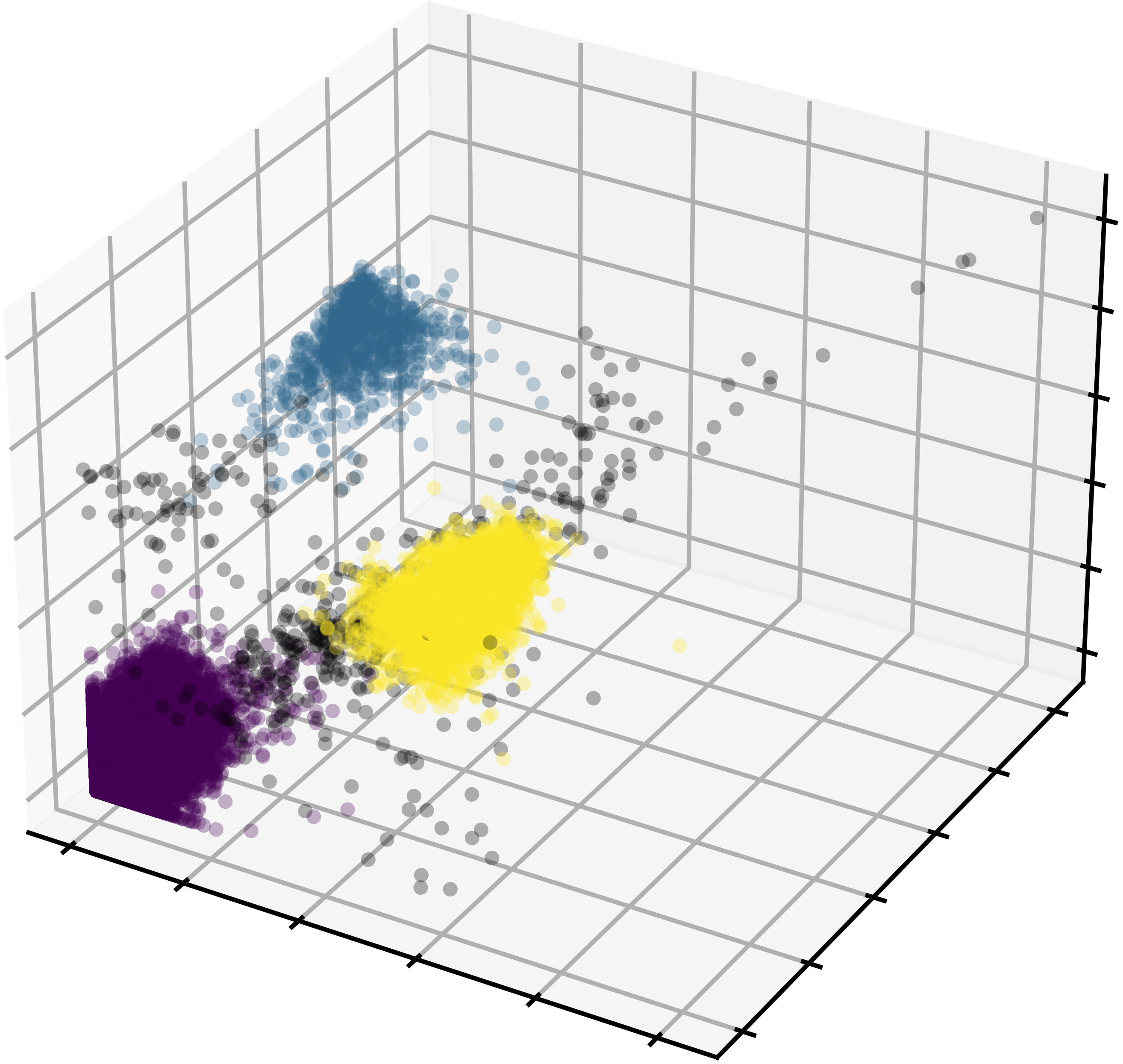
  - Each component is a meaningful ***type***

- **Goal:** discover these types
  - Determine $K_\circ$

- Assumed mixture model:

$$x_n \sim P_\theta := \sum_{k=1}^{K} \pi_k F_{\boxed{\phi_k}} \longleftarrow \text{component parameter}$$

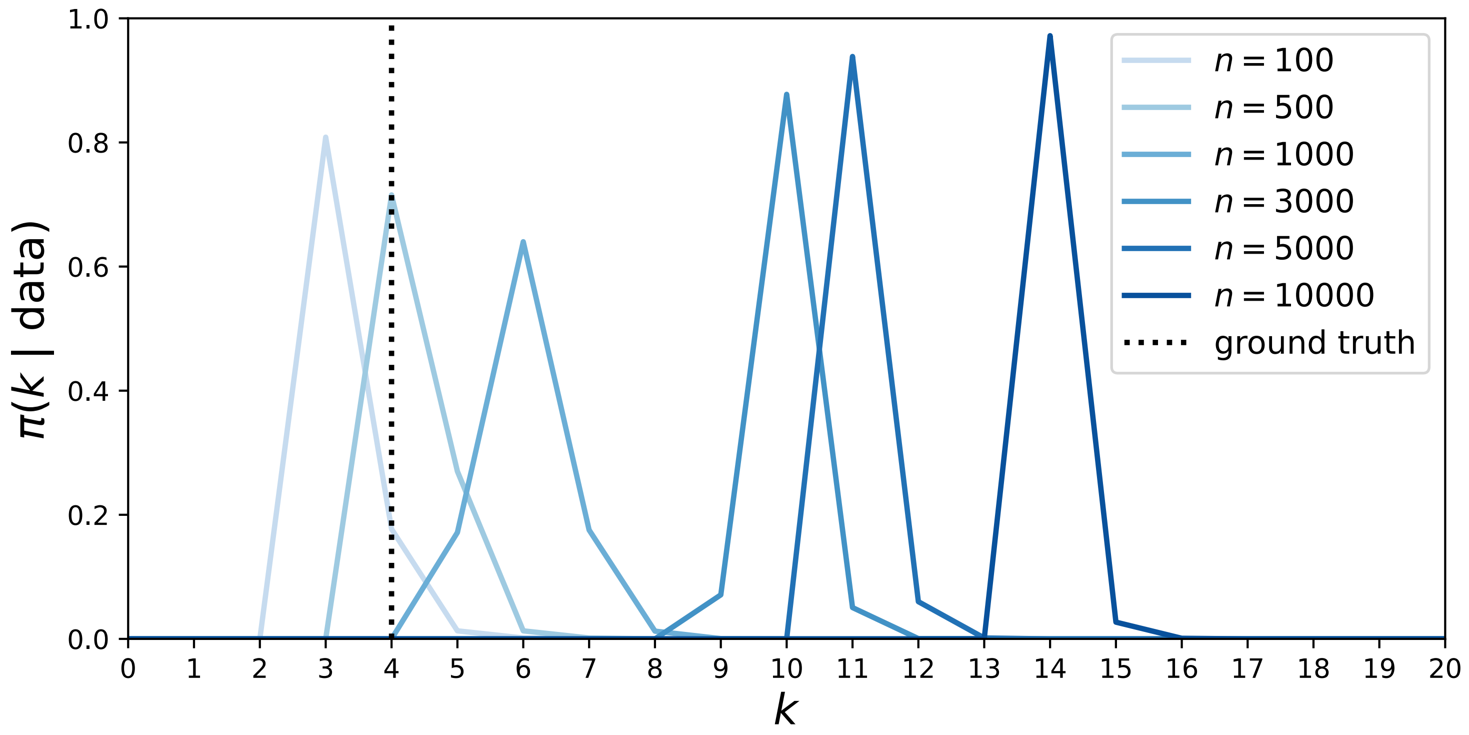
- Parameters: $\theta = \theta^{(K)} = (K, \pi, \phi_1, \ldots, \phi_K)$



Why is type discovery hard?

Model: $x_n \sim P_\theta = \sum_{k=1}^{K} \pi_k F_{\phi_k}$

DGP: $x_n \sim P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$



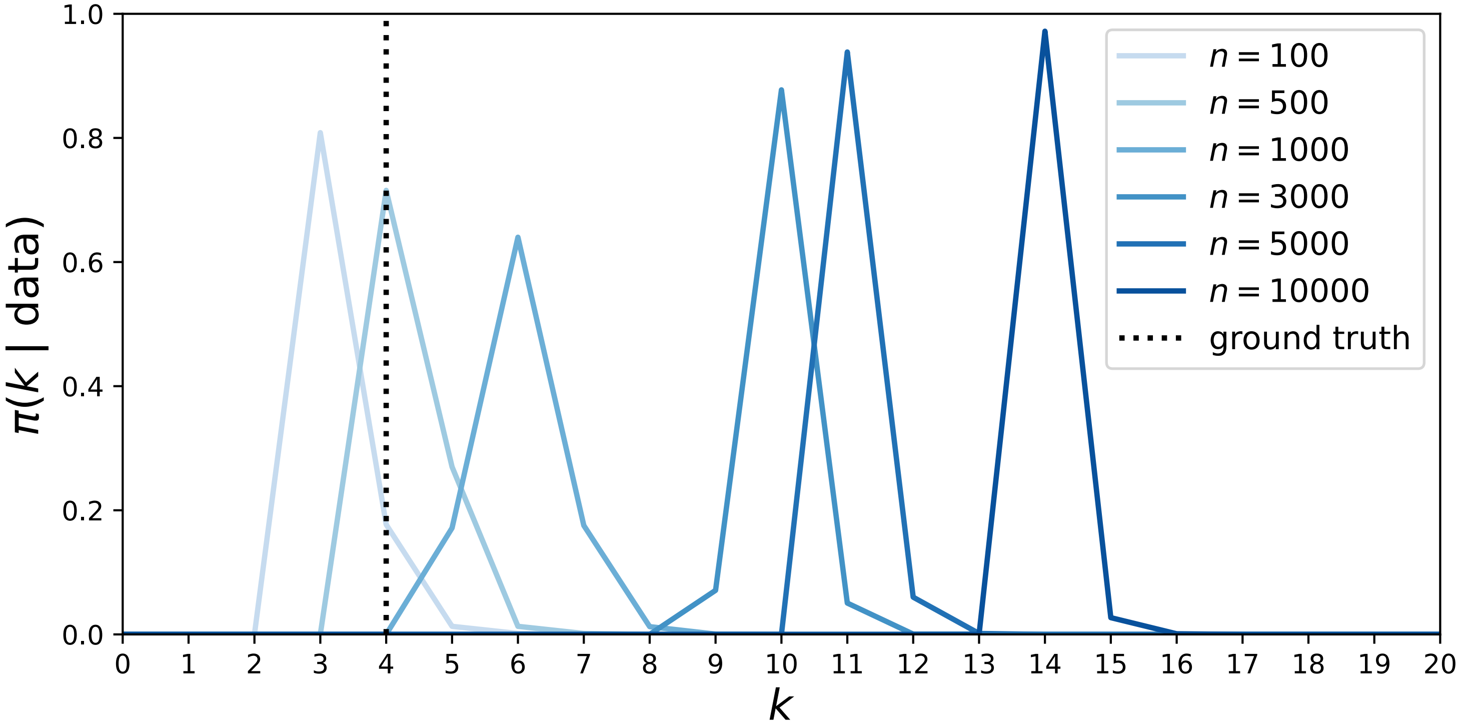
[Miller & Dunson 2018, Cai, Campbell & Broderick 2021]

Why is type discovery hard?

Model: $x_n \sim P_\theta = \sum_{k=1}^{K} \pi_k F_{\phi_k}$

DGP: $x_n \sim P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$

- Usually, $F_\phi \neq P_{\circ k}$ for all ϕ



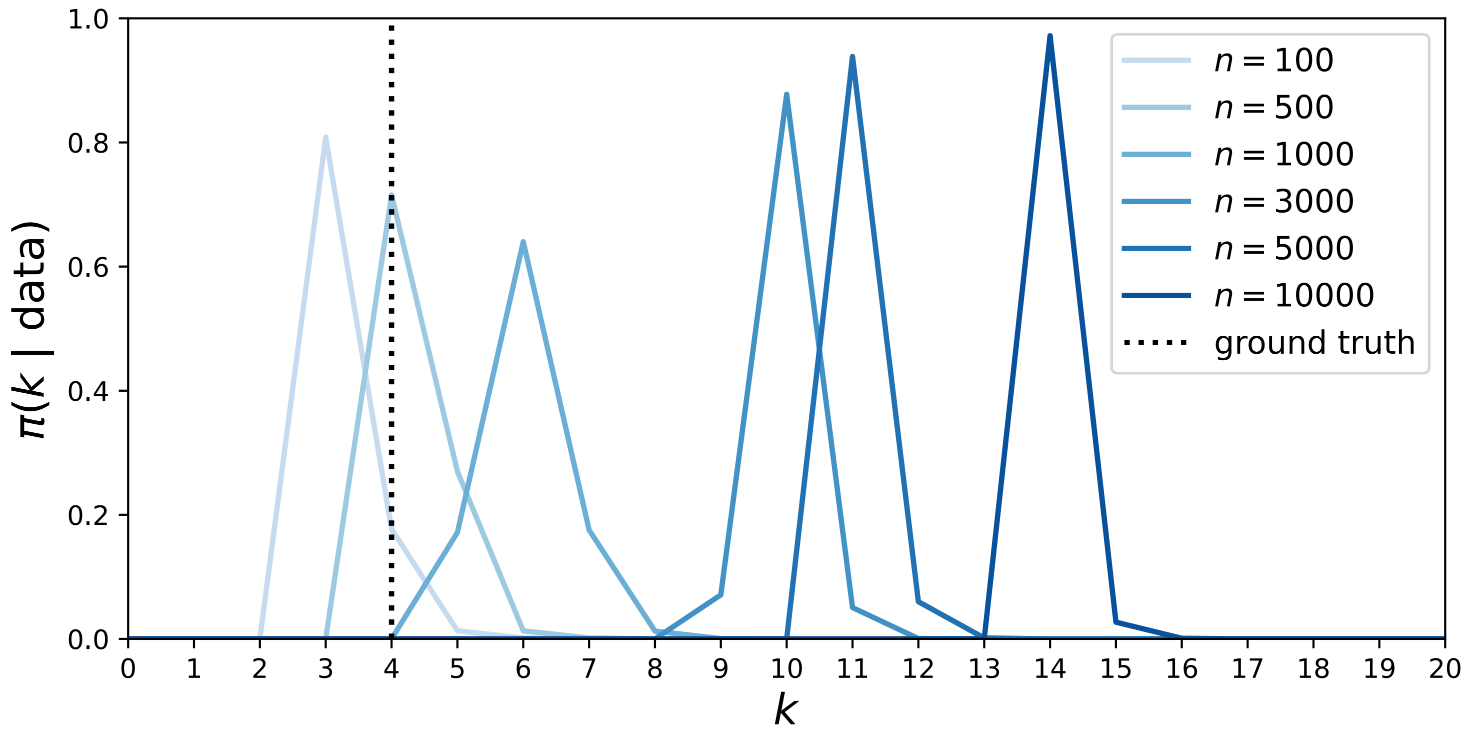
[Miller & Dunson 2018, Cai, Campbell & Broderick 2021]

Why is type discovery hard?

Model: $x_n \sim P_\theta = \sum_{k=1}^{K} \pi_k F_{\phi_k}$

DGP: $x_n \sim P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$

- Usually, $F_\phi \neq P_{\circ k}$ for all ϕ
- But standard inference methods ≡ density estimation



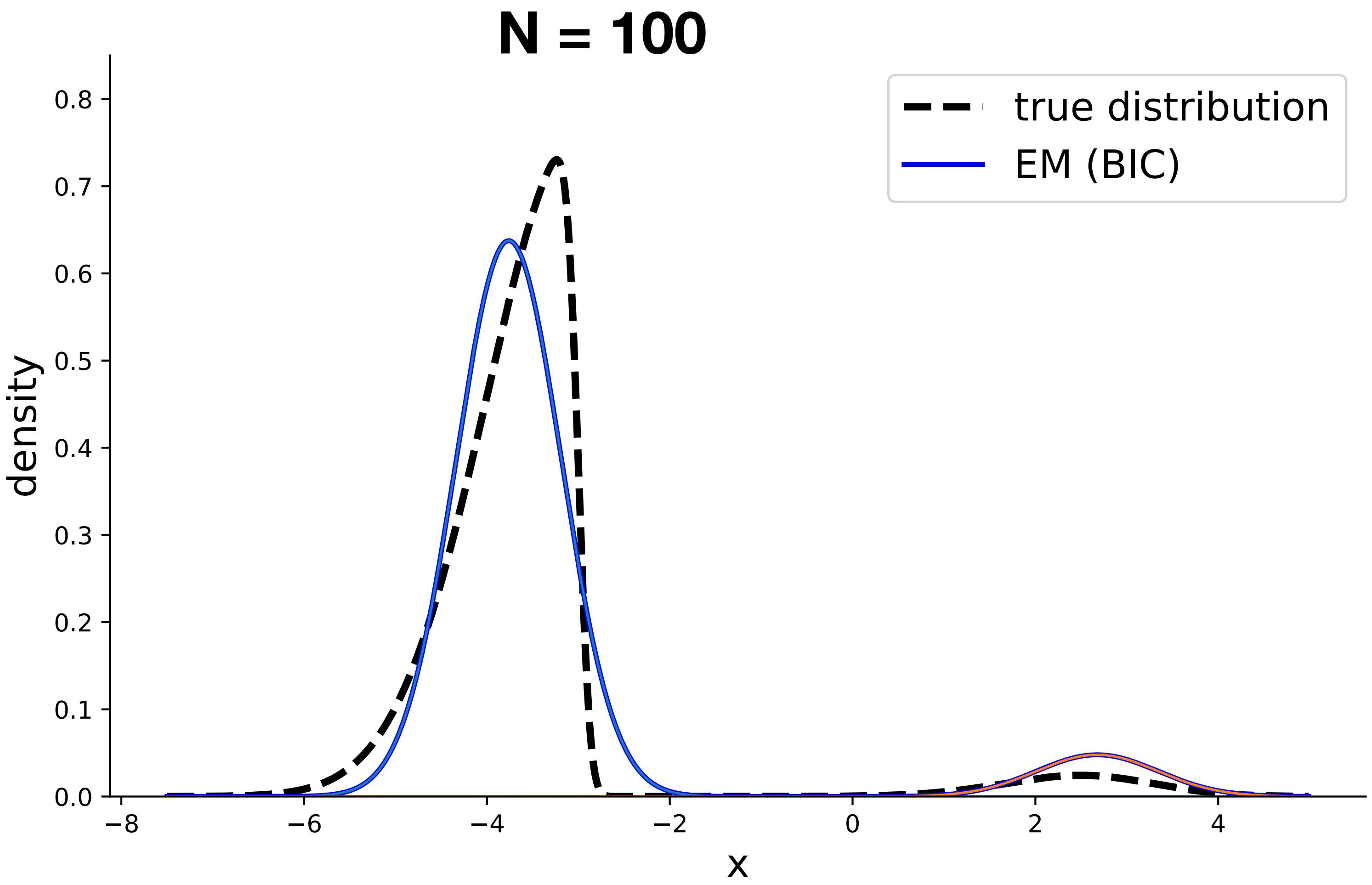
[Miller & Dunson 2018, Cai, Campbell & Broderick 2021]

Why is type discovery hard?

Model: $x_n \sim P_\theta = \sum_{k=1}^{K} \pi_k F_{\phi_k}$

DGP: $x_n \sim P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$

- Usually, $F_\phi \neq P_{\circ k}$ for all ϕ
- But standard inference methods ≡ density estimation
- Synthetic example: skew-normal mixture



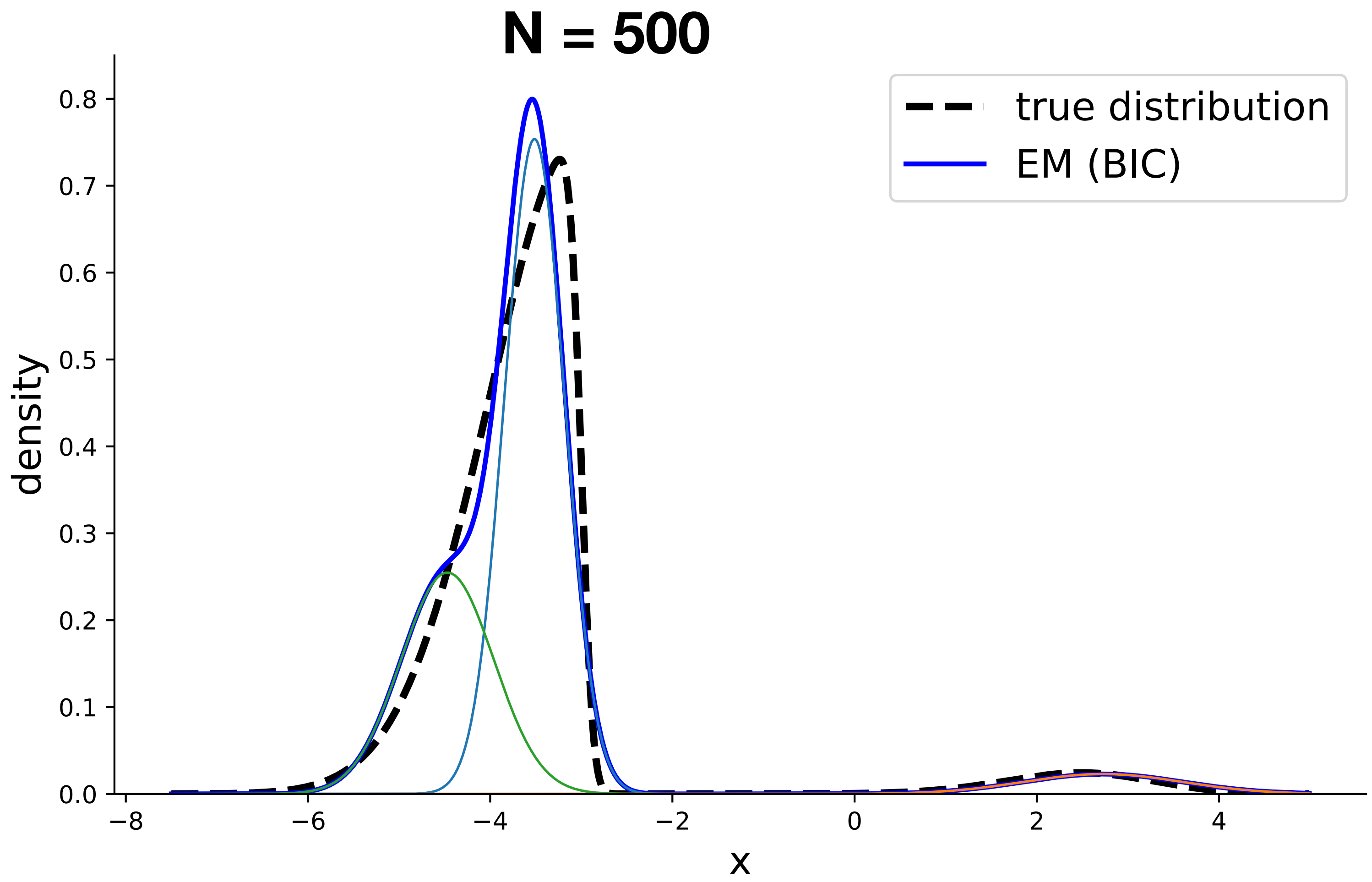
[Miller & Dunson 2018, Cai, Campbell & Broderick 2021]

Why is type discovery hard?

Model: $x_n \sim P_\theta = \sum_{k=1}^{K} \pi_k F_{\phi_k}$

DGP: $x_n \sim P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$

- Usually, $F_\phi \neq P_{\circ k}$ for all ϕ
- But standard inference methods ≡ density estimation
- Synthetic example: skew-normal mixture



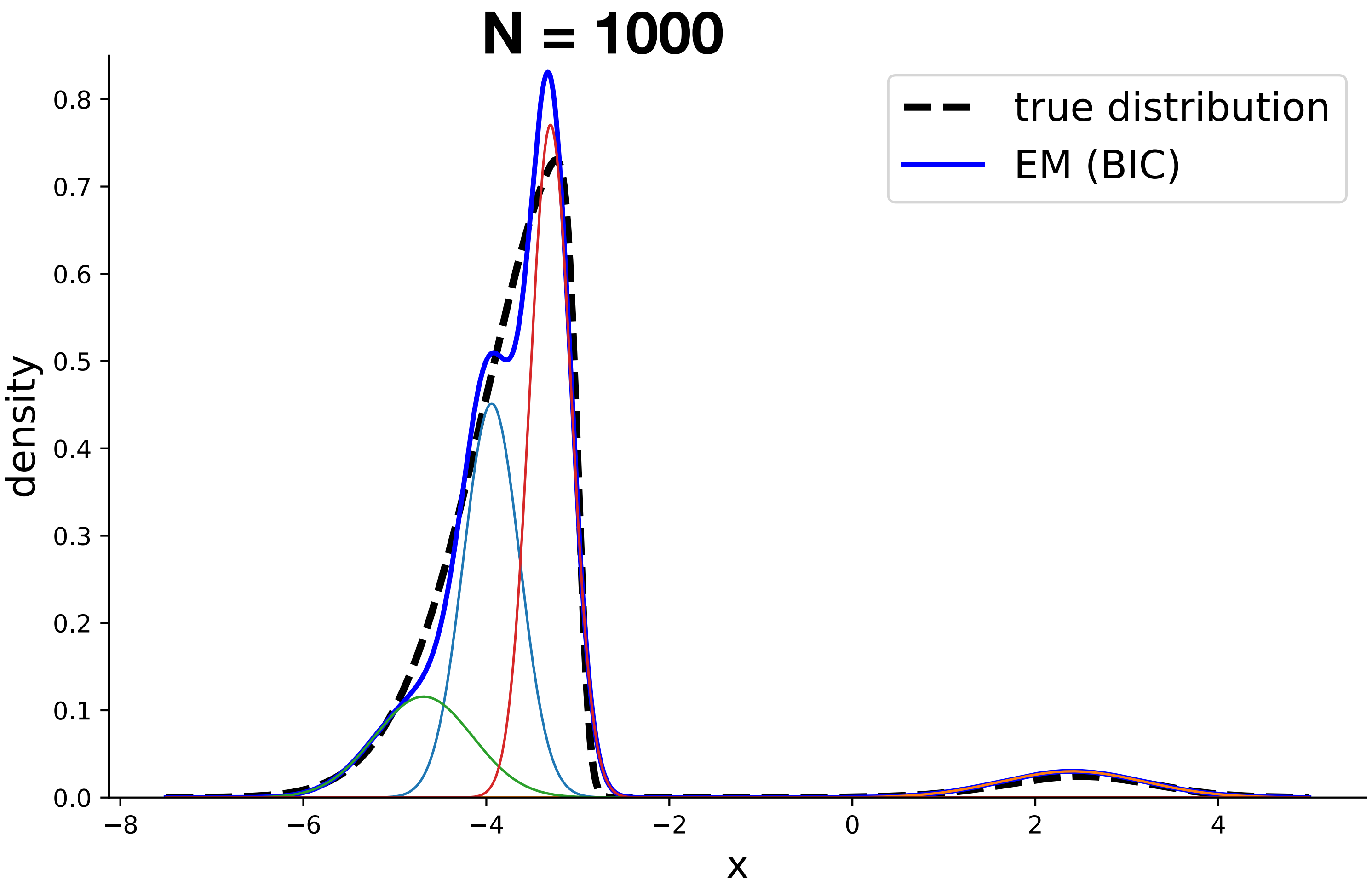
[Miller & Dunson 2018, Cai, Campbell & Broderick 2021]

Why is type discovery hard?

Model: $x_n \sim P_\theta = \sum_{k=1}^{K} \pi_k F_{\phi_k}$

DGP: $x_n \sim P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$

- Usually, $F_\phi \neq P_{\circ k}$ for all ϕ
- But standard inference methods ≡ density estimation
- Synthetic example: skew-normal mixture



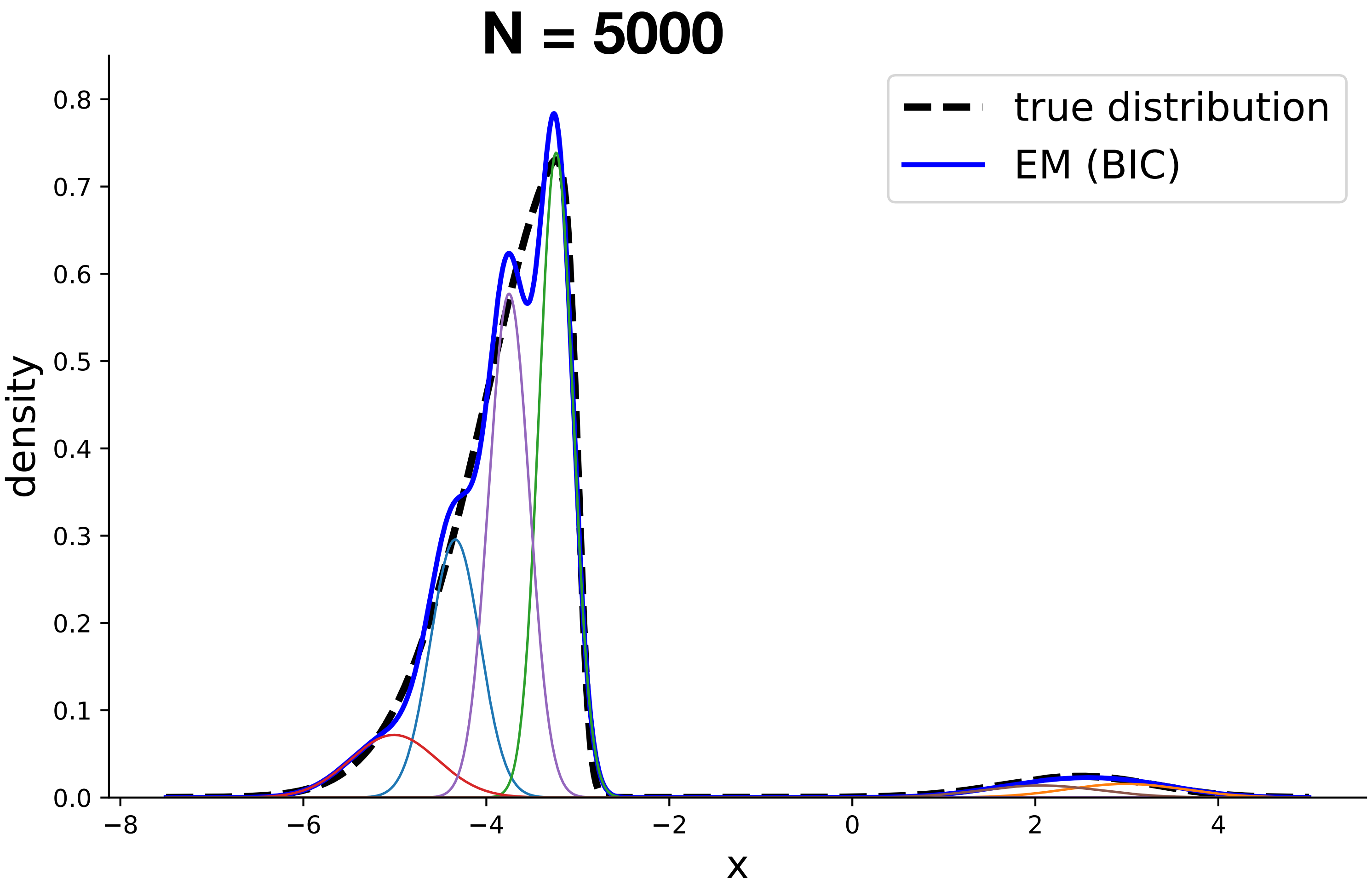
[Miller & Dunson 2018, Cai, Campbell & Broderick 2021]

Why is type discovery hard?

Model: $x_n \sim P_\theta = \sum_{k=1}^{K} \pi_k F_{\phi_k}$

DGP: $x_n \sim P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$

- Usually, $F_\phi \neq P_{\circ k}$ for all ϕ
- But standard inference methods ≡ density estimation
- Synthetic example: skew-normal mixture



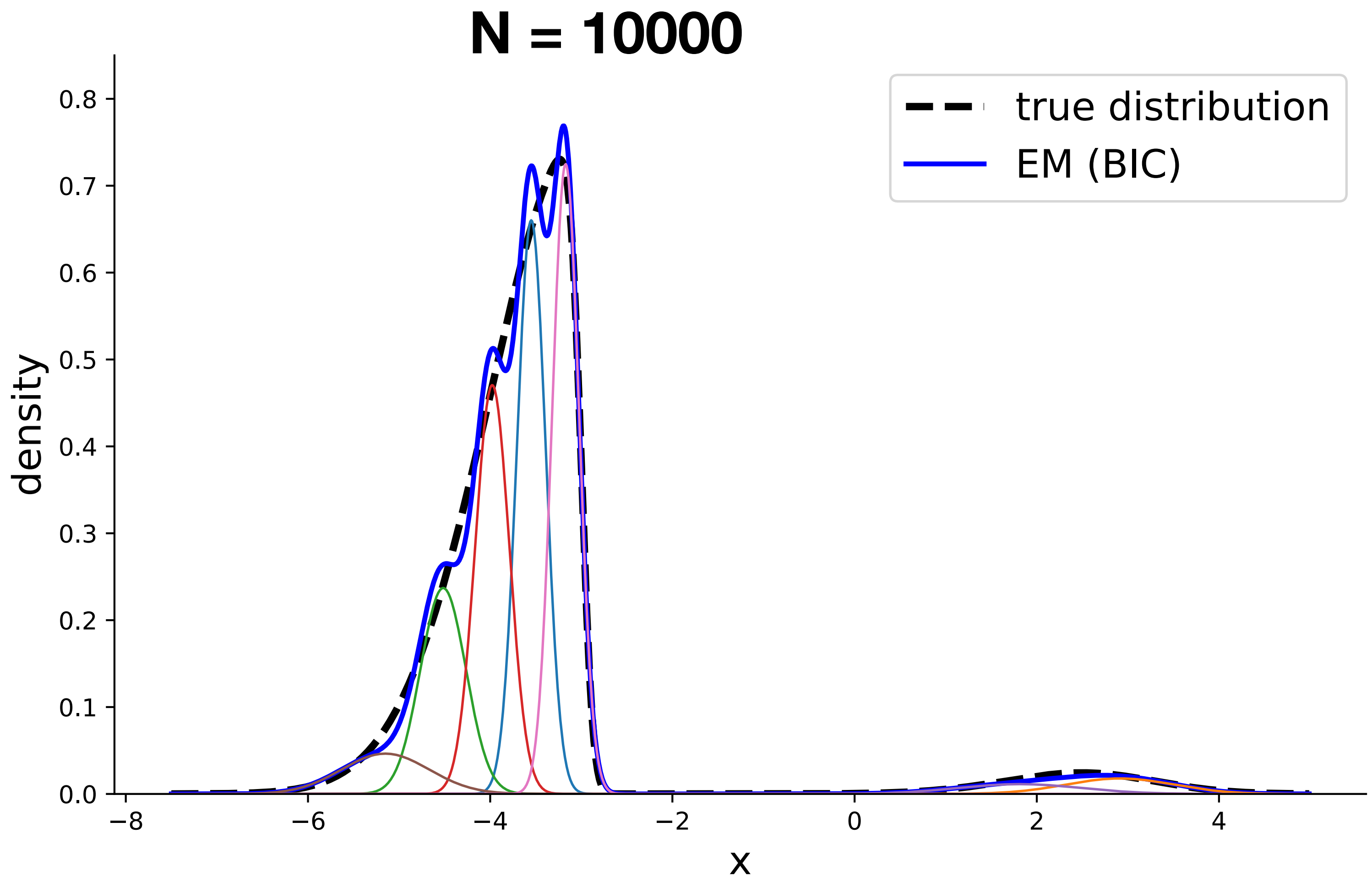
[Miller & Dunson 2018, Cai, Campbell & Broderick 2021]

Why is type discovery hard?

Model: $x_n \sim P_\theta = \sum_{k=1}^{K} \pi_k F_{\phi_k}$

DGP: $x_n \sim P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$

- Usually, $F_\phi \neq P_{\circ k}$ for all ϕ
- But standard inference methods ≡ density estimation
- Synthetic example: skew-normal mixture



[Miller & Dunson 2018, Cai, Campbell & Broderick 2021]

Skew-Gaussian mixture: coarsened posterior

[Miller & Dunson 2018, Li & **H** 2023+]

Skew-Gaussian mixture: coarsened posterior

- One solution: ***coarsened posterior*** $\pi_{\alpha}(\theta \mid X) \propto p(X \mid \theta)^{\frac{\alpha}{N+\alpha}} \pi_0(\theta)$

[Miller & Dunson 2018, Li & **H** 2023+]

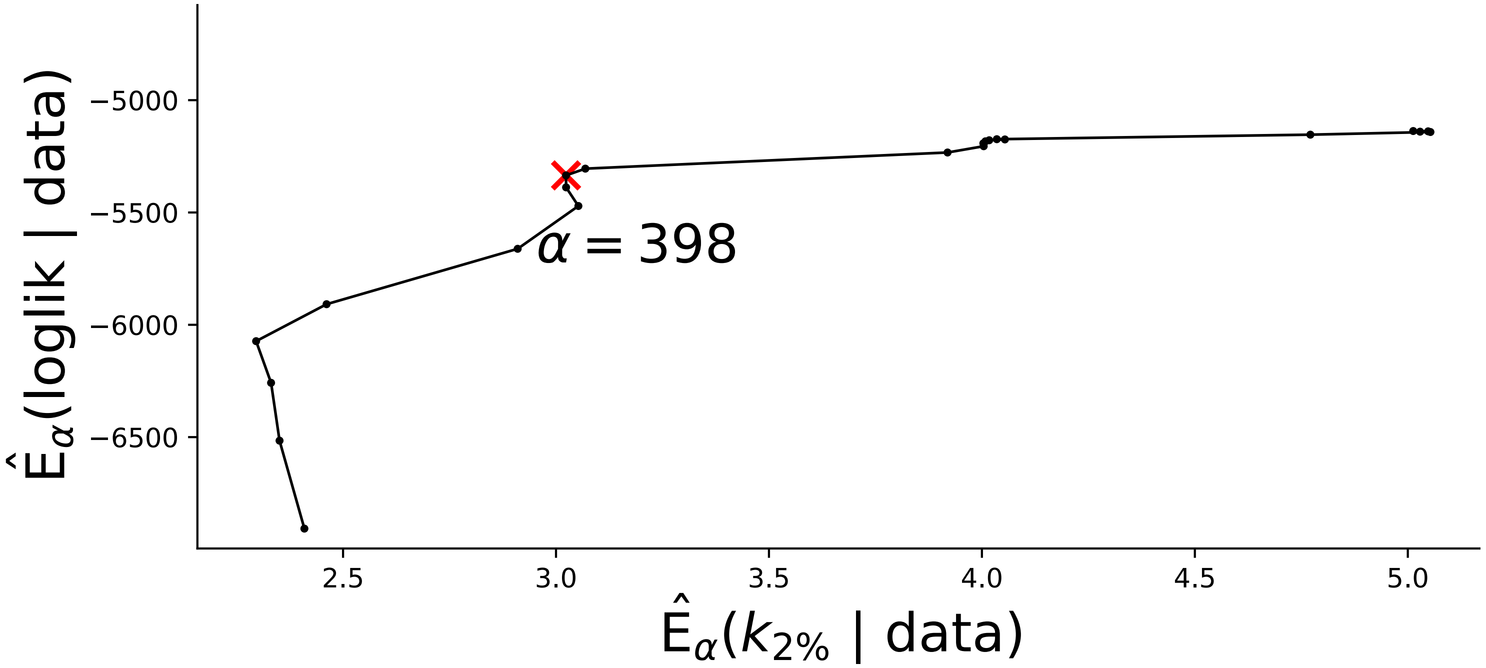
Skew-Gaussian mixture: coarsened posterior

- One solution: ***coarsened posterior*** $\pi_\alpha(\theta \mid X) \propto p(X \mid \theta)^{\frac{\alpha}{N+\alpha}} \pi_0(\theta)$
  - Roughly, $\min_\theta \mathrm{KL}(P_\circ \mid P_\theta) \lesssim \alpha^{-1}$

[Miller & Dunson 2018, Li & **H** 2023+]

Skew-Gaussian mixture: coarsened posterior

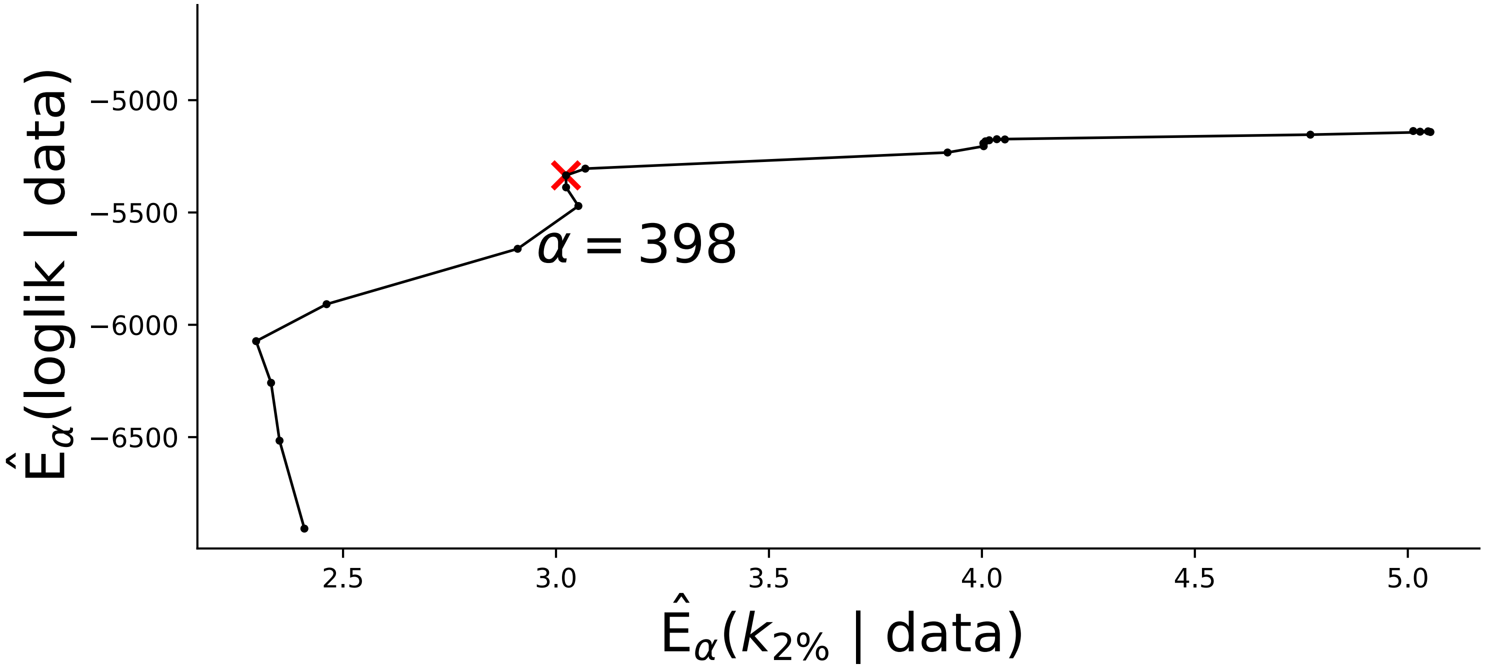
- One solution: ***coarsened posterior*** $\pi_{\alpha}(\theta \mid X) \propto p(X \mid \theta)^{\frac{\alpha}{N+\alpha}} \pi_0(\theta)$
  - Roughly, $\min_{\theta} \mathrm{KL}(P_{\circ} \mid P_{\theta}) \lesssim \alpha^{-1}$



[Miller & Dunson 2018, Li & **H** 2023+]

Skew-Gaussian mixture: coarsened posterior

- One solution: ***coarsened posterior*** $\pi_\alpha(\theta \mid X) \propto p(X \mid \theta)^{\frac{\alpha}{N+\alpha}} \pi_0(\theta)$
  - Roughly, $\min_\theta \mathrm{KL}(P_\circ \mid P_\theta) \lesssim \alpha^{-1}$

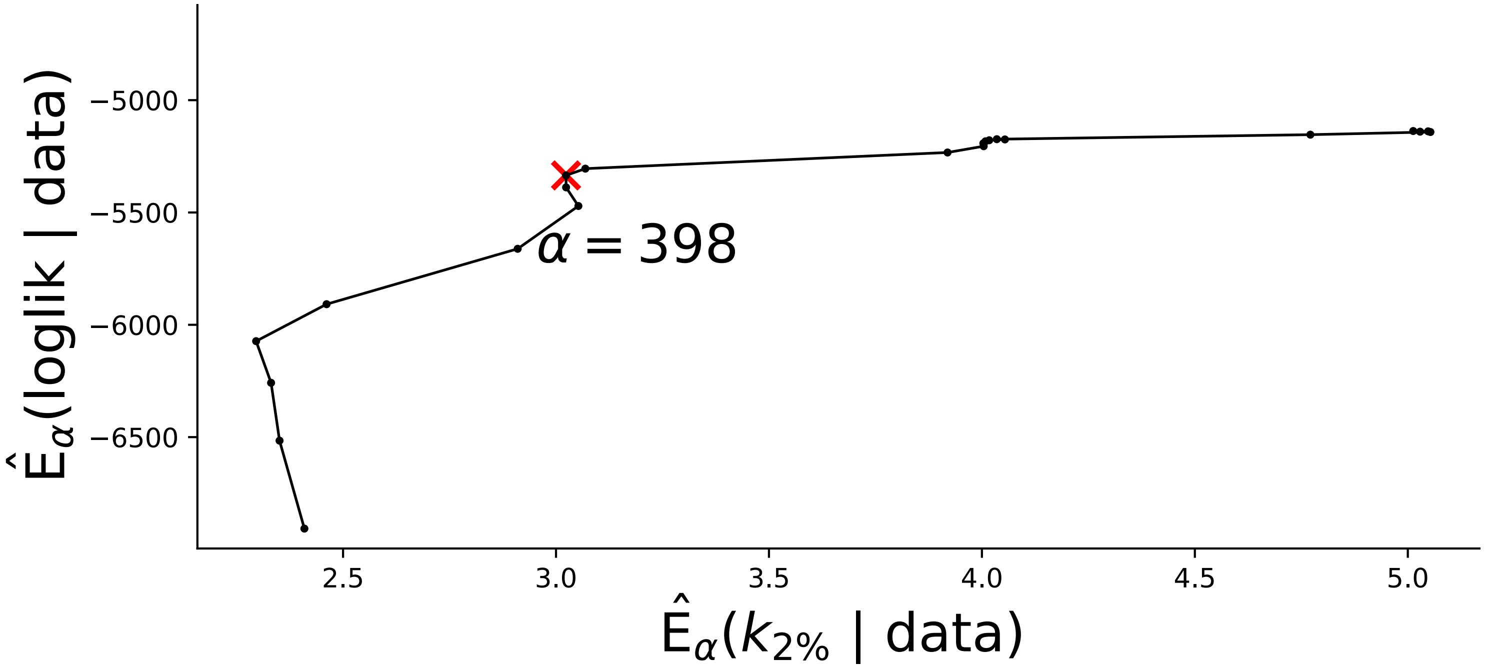


Computationally costly!

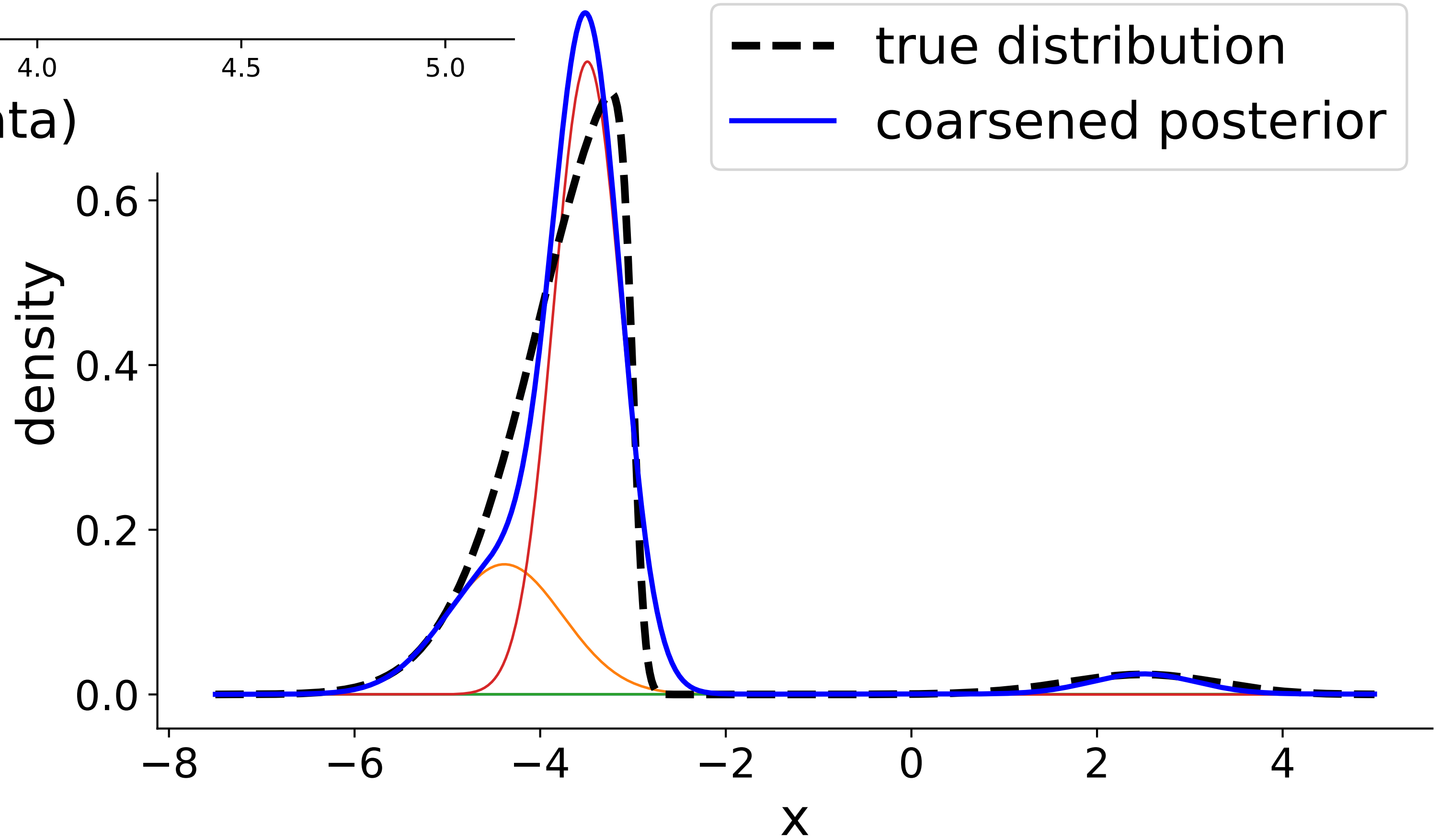
[Miller & Dunson 2018, Li & **H** 2023+]

Skew-Gaussian mixture: coarsened posterior

- One solution: ***coarsened posterior*** $\pi_\alpha(\theta \mid X) \propto p(X \mid \theta)^{\frac{\alpha}{N+\alpha}} \pi_0(\theta)$
  - Roughly, $\min_\theta \mathrm{KL}(P_\circ \mid P_\theta) \lesssim \alpha^{-1}$



Computationally costly!



[Miller & Dunson 2018, Li & **H** 2023+]

(Causal) structure matters

Data-generating process

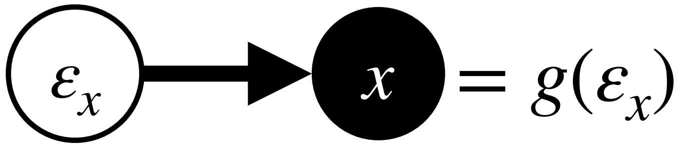
$$x_n \sim P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$$

(Causal) structure matters

Data-generating process

$$x_n \sim P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$$

Structural causal model

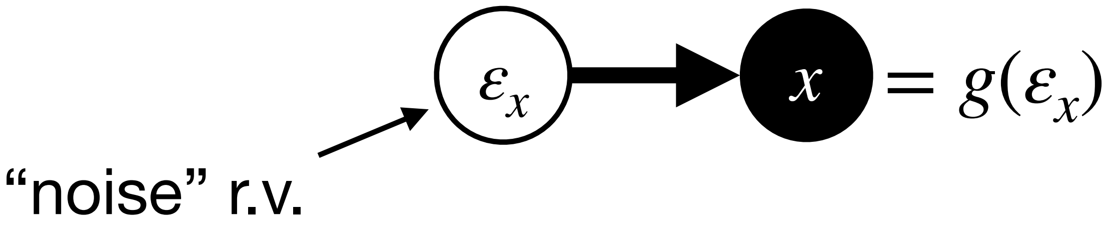


(Causal) structure matters

Data-generating process

$$x_n \sim P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$$

Structural causal model



ε_x

x $= g(\varepsilon_x)$

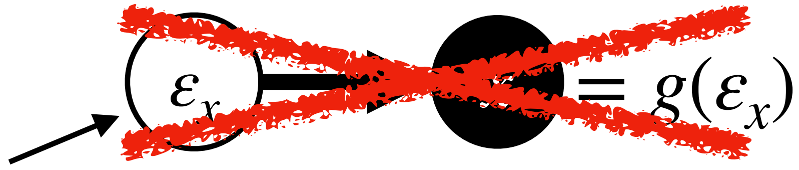
“noise” r.v.

(Causal) structure matters

Data-generating process

~~$x_n \sim P_n = \sum_{k=1}^{K} \pi_{\circ k} P_{\circ k}$~~

Structural causal model



“noise” r.v.

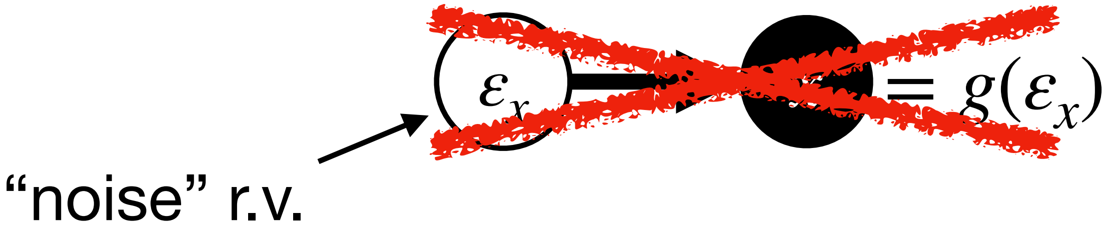
(Causal) structure matters

Data-generating process

$$\xcancel{x_n \sim P_\circ = \textstyle\sum_{k=1}^{K} \pi_{\circ k} P_{\circ k}}$$

$$z_n \sim \text{Categorical}(\pi_\circ)$$
$$x_n \mid z_n = k \sim P_{\circ k}$$

Structural causal model



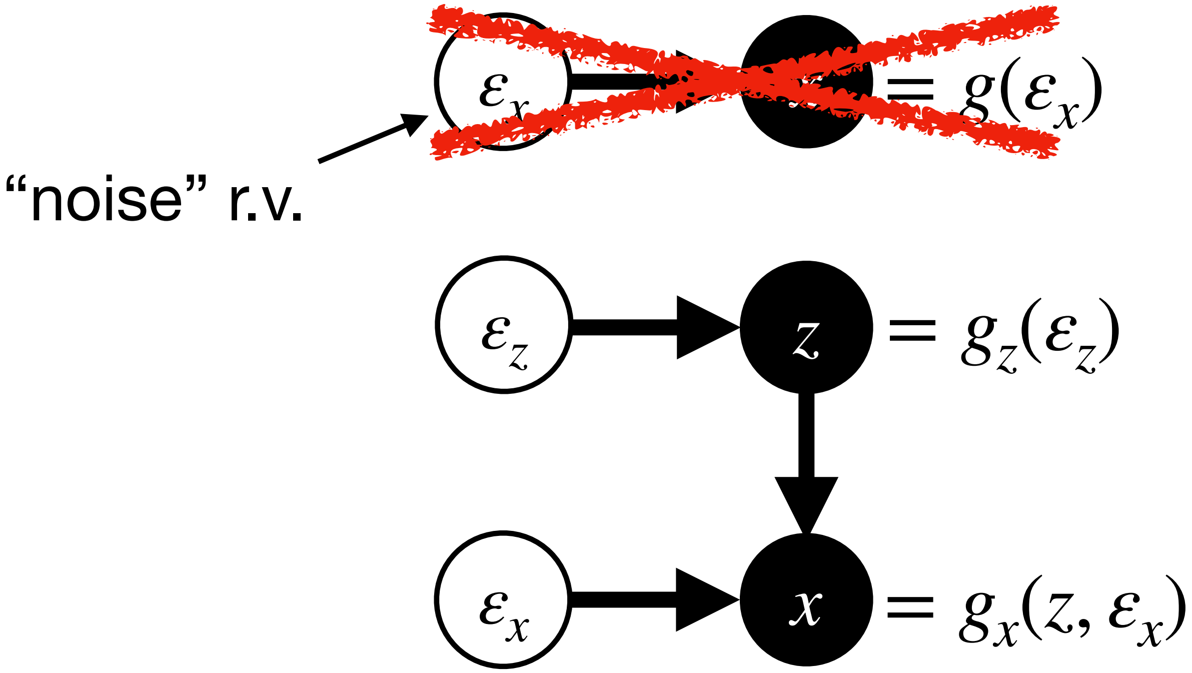
(Causal) structure matters

Data-generating process

~~$x_n \sim P_\circ = \sum_{k=1}^{K} \pi_{\circ k} P_{\circ k}$~~

$$z_n \sim \text{Categorical}(\pi_\circ)$$
$$x_n \mid z_n = k \sim P_{\circ k}$$

Structural causal model



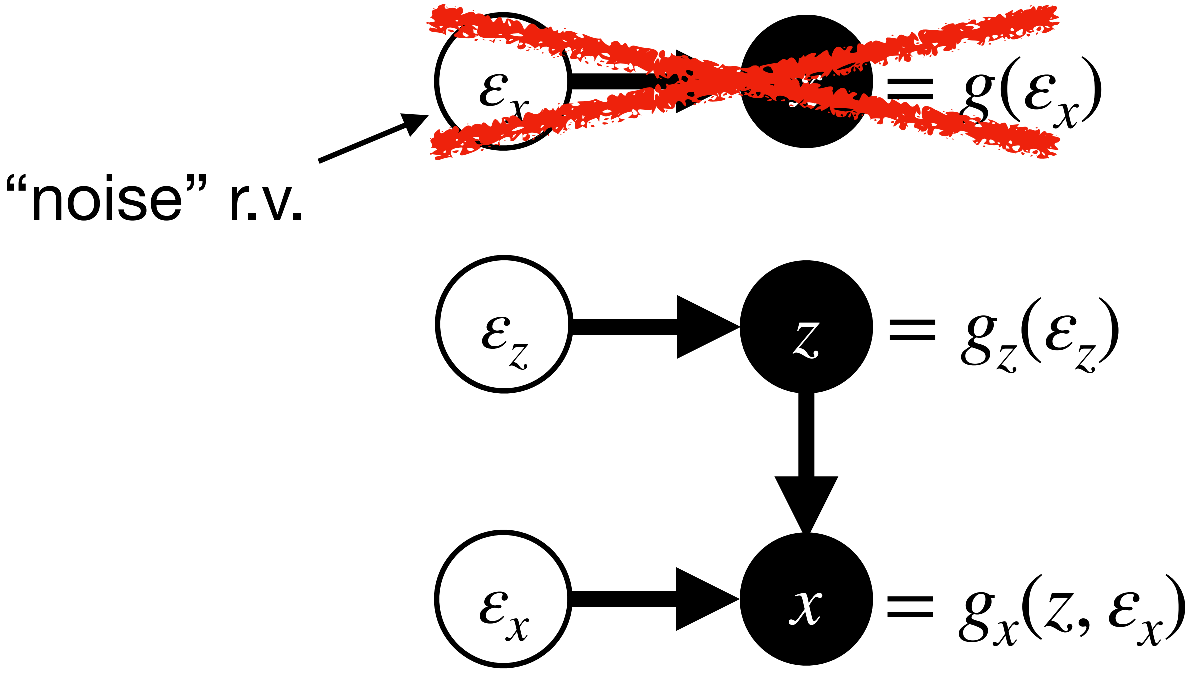
(Causal) structure matters

Data-generating process

$$\xcancel{x_n \sim P_\circ = \textstyle\sum_{k=1}^{K} \pi_{\circ k} P_{\circ k}}$$

$$z_n \sim \text{Categorical}(\pi_\circ)$$
$$x_n \mid z_n = k \sim P_{\circ k}$$

Structural causal model



Assumed model

$$z_n \sim \text{Categorical}(\pi)$$
$$x_n \mid z_n = k \sim F_{\phi_k}$$

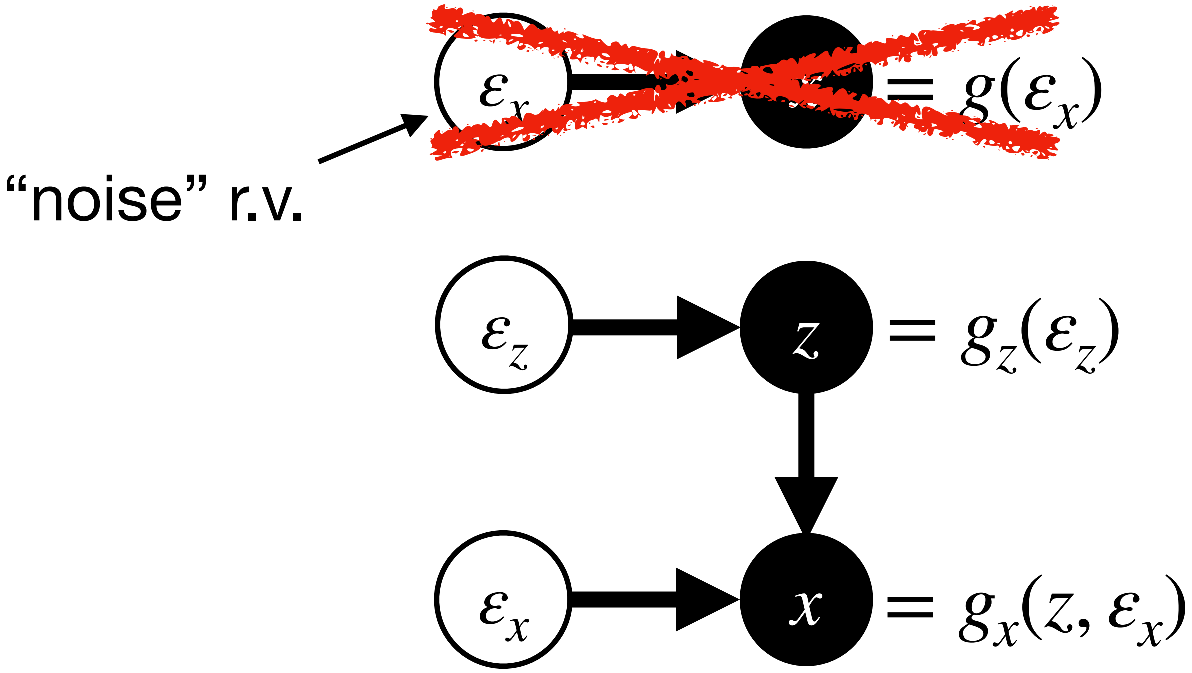
(Causal) structure matters

Data-generating process

~~$x_n \sim P_\circ = \sum_{k=1}^{K} \pi_{\circ k} P_{\circ k}$~~

$$z_n \sim \text{Categorical}(\pi_\circ)$$
$$x_n \mid z_n = k \sim P_{\circ k}$$

Structural causal model



Assumed model

$$z_n \sim \text{Categorical}(\pi)$$
$$x_n \mid z_n = k \sim F_{\phi_k}$$

✓

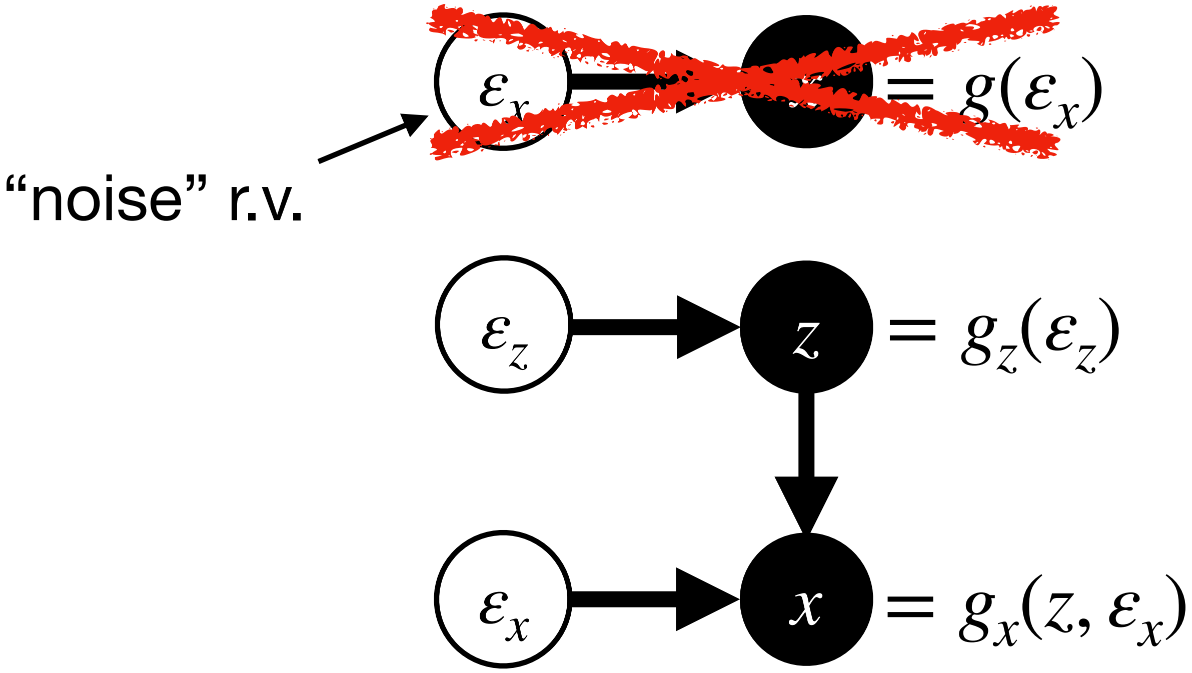
(Causal) structure matters

<u>Data-generating process</u>

$$\xcancel{x_n \sim P_\circ = \sum_{k=1}^{K} \pi_{\circ k} P_{\circ k}}$$

$$z_n \sim \text{Categorical}(\pi_\circ)$$
$$x_n \mid z_n = k \sim P_{\circ k}$$

<u>Structural causal model</u>



<u>Assumed model</u>

$z_n \sim \text{Categorical}(\pi)$ ✓

$x_n \mid z_n = k \sim F_{\phi_k}$ ✗ $F_{\phi_k} \neq P_{\circ k}$

Structurally aware model selection

$$z_n \sim \mathrm{Categorical}(\pi)$$

$$x_n \mid z_n = k \sim F_{\phi_k}$$

$$P_\theta = \sum_{k=1}^{K} \pi_k F_{\phi_k}$$

$$x_n \sim P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$$

[Li & **H** 2023+]

Structurally aware model selection

$$z_n \sim \mathrm{Categorical}(\pi)$$

$$x_n \mid z_n = k \sim F_{\phi_k}$$

$$P_\theta = \sum_{k=1}^{K} \pi_k F_{\phi_k}$$

$$x_n \sim P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$$

- **Recall:** for coarsened posterior, assume $\min_\theta \mathrm{KL}(P_\circ \mid P_\theta) \lesssim \alpha^{-1}$

[Li & **H** 2023+]

Structurally aware model selection

$$z_n \sim \text{Categorical}(\pi)$$

$$x_n \mid z_n = k \sim F_{\phi_k}$$

$$P_\theta = \textstyle\sum_{k=1}^{K} \pi_k F_{\phi_k}$$

$$x_n \sim P_\circ = \textstyle\sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$$

- **Recall:** for coarsened posterior, assume $\min_\theta \mathrm{KL}(P_\circ \mid P_\theta) \lesssim \alpha^{-1}$
- But want to coarsen at **component level**

[Li & **H** 2023+]

Structurally aware model selection

$$z_n \sim \text{Categorical}(\pi)$$

$$x_n \mid z_n = k \sim F_{\phi_k}$$

$$P_\theta = \sum_{k=1}^{K} \pi_k F_{\phi_k} \qquad x_n \sim P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$$

- **Recall:** for coarsened posterior, assume $\min_\theta \mathrm{KL}(P_\circ \mid P_\theta) \lesssim \alpha^{-1}$
- But want to coarsen at **component level**
- Our approach:

[Li & **H** 2023+]

Structurally aware model selection

$$z_n \sim \text{Categorical}(\pi) \qquad P_\theta = \textstyle\sum_{k=1}^{K} \pi_k F_{\phi_k} \qquad x_n \sim P_\circ = \textstyle\sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$$

$$x_n \mid z_n = k \sim F_{\phi_k}$$

- **Recall:** for coarsened posterior, assume $\min_\theta \mathrm{KL}(P_\circ \mid P_\theta) \lesssim \alpha^{-1}$
- But want to coarsen at **component level**
- Our approach:
  - Choose divergence $\mathscr{D}(P \mid Q)$

[Li & **H** 2023+]

Structurally aware model selection

$$z_n \sim \text{Categorical}(\pi)$$
$$x_n \mid z_n = k \sim F_{\phi_k}$$

$$P_\theta = \sum_{k=1}^{K} \pi_k F_{\phi_k}$$

$$x_n \sim P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$$

- **Recall:** for coarsened posterior, assume $\min_\theta \mathrm{KL}(P_\circ \mid P_\theta) \lesssim \alpha^{-1}$
- But want to coarsen at **<u>component level</u>**
- Our approach:
  - Choose divergence $\mathscr{D}(P \mid Q)$
  - **Assumption:** $\min_\phi \mathscr{D}(P_{\circ k} \mid F_\phi) \lesssim \rho$

[Li & **H** 2023+]

Structurally aware model selection

$z_n \sim \text{Categorical}(\pi)$

$x_n \mid z_n = k \sim F_{\phi_k}$

$P_\theta = \sum_{k=1}^{K} \pi_k F_{\phi_k}$

$x_n \sim P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$

- **Recall:** for coarsened posterior, assume $\min_\theta \mathrm{KL}(P_\circ \mid P_\theta) \lesssim \alpha^{-1}$
- But want to coarsen at **<u>component level</u>**
- Our approach:
    - Choose divergence $\mathscr{D}(P \mid Q)$
    - **Assumption:** $\min_\phi \mathscr{D}(P_{\circ k} \mid F_\phi) \lesssim \rho$
    - Need (consistent) estimator $\widehat{\mathscr{D}}(x_1, \ldots, x_n \mid Q)$ for $\mathscr{D}(P \mid Q)$

[Li & H 2023+]

Structurally aware model selection

$$z_n \sim \text{Categorical}(\pi) \qquad P_\theta = \textstyle\sum_{k=1}^{K} \pi_k F_{\phi_k} \qquad x_n \sim P_\circ = \textstyle\sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$$

$$x_n \mid z_n = k \sim F_{\phi_k}$$

- **Recall:** for coarsened posterior, assume $\min_\theta \mathrm{KL}(P_\circ \mid P_\theta) \lesssim \alpha^{-1}$
- But want to coarsen at **component level**
- Our approach:
  - Choose divergence $\mathscr{D}(P \mid Q)$
  - **Assumption:** $\min_\phi \mathscr{D}(P_{\circ k} \mid F_\phi) \lesssim \rho$
  - Need (consistent) estimator $\widehat{\mathscr{D}}(x_1, \ldots, x_n \mid Q)$ for $\mathscr{D}(P \mid Q)$
  - $X_k(z) := \{x_n : z_n = k\}$ [observations assigned to kth component]

[Li & H 2023+]

Structurally aware model selection

$$z_n \sim \text{Categorical}(\pi) \qquad x_n \mid z_n = k \sim F_{\phi_k} \qquad P_\theta = \textstyle\sum_{k=1}^{K} \pi_k F_{\phi_k} \qquad x_n \sim P_\circ = \textstyle\sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$$

- **Recall:** for coarsened posterior, assume $\min_\theta \mathrm{KL}(P_\circ \mid P_\theta) \lesssim \alpha^{-1}$
- But want to coarsen at **<u>component level</u>**
- Our approach:
  - Choose divergence $\mathscr{D}(P \mid Q)$
  - **Assumption:** $\min_\phi \mathscr{D}(P_{\circ k} \mid F_\phi) \lesssim \rho$
  - Need (consistent) estimator $\widehat{\mathscr{D}}(x_1, \ldots, x_n \mid Q)$ for $\mathscr{D}(P \mid Q)$
  - $X_k(z) := \{x_n : z_n = k\}$ [observations assigned to kth component]
  - ***Structurally aware loss***:

$$\mathscr{L}_\rho(\theta \mid z) = \textstyle\sum_{k=1}^{K} |X_k(z)| \max\left\{0, \widehat{\mathscr{D}}(X_k(z) \mid F_{\phi_k}) - \rho\right\}$$

[Li & **H** 2023+]

Structurally aware model selection

$$z_n \sim \mathrm{Categorical}(\pi)$$
$$x_n \mid z_n = k \sim F_{\phi_k}$$

$$P_\theta = \textstyle\sum_{k=1}^{K} \pi_k F_{\phi_k} \qquad x_n \sim P_\circ = \textstyle\sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$$

- **Recall:** for coarsened posterior, assume $\min_\theta \mathrm{KL}(P_\circ \mid P_\theta) \lesssim \alpha^{-1}$
- But want to coarsen at <u>**component level**</u>
- Our approach:
  - Choose divergence $\mathscr{D}(P \mid Q)$
  - **Assumption:** $\min_\phi \mathscr{D}(P_{\circ k} \mid F_\phi) \lesssim \rho$
  - Need (consistent) estimator $\widehat{\mathscr{D}}(x_1, \dots, x_n \mid Q)$ for $\mathscr{D}(P \mid Q)$
  - $X_k(z) := \{x_n : z_n = k\}$ [observations assigned to kth component]
  - ***Structurally aware loss***:

$$\mathscr{L}_\rho(\theta \mid z) = \textstyle\sum_{k=1}^{K} |X_k(z)| \max\left\{0, \widehat{\mathscr{D}}(X_k(z) \mid F_{\phi_k}) - \rho\right\} + \overset{\textit{AIC-like penalty}}{\lambda K}$$

[Li & **H** 2023+]

STructurally Aware Robust Estimation (STARE)

[Li & H 2023+]

STructurally Aware Robust Estimation (STARE)

1. Estimate model parameters $\hat{\theta}^{(K)}$ for each $K \in \{1,\ldots,K_{\max}\}$

[Li & **H** 2023+]

STructurally Aware Robust Estimation (STARE)

1. Estimate model parameters $\hat{\theta}^{(K)}$ for each $K \in \{1, \ldots, K_{\max}\}$

2. Select ρ

[Li & **H** 2023+]

STructurally Aware Robust Estimation (STARE)

1. Estimate model parameters $\hat{\theta}^{(K)}$ for each $K \in \{1, \ldots, K_{\max}\}$

2. Select ρ

3. Set $\hat{K}_\rho = \arg\min_K \mathscr{L}_\rho(\hat{\theta}^{(K)} \mid z^{(K)})$

[Li & **H** 2023+]

STructurally Aware Robust Estimation (STARE)

1. Estimate model parameters $\hat{\theta}^{(K)}$ for each $K \in \{1, \ldots, K_{\max}\}$
2. Select ρ
3. Set $\hat{K}_\rho = \arg\min_K \mathscr{L}_\rho(\hat{\theta}^{(K)} \mid z^{(K)})$

- Open question: how to choose ρ

[Li & **H** 2023+]

Consistency of STARE

[Li & **H** 2023+]

Consistency of STARE

Definition: $\pi_{\star}^{(K)}, \phi_{\star 1}^{(K)}, \ldots, \phi_{\star K}^{(K)}$ = asymptotically inferred parameters *given K*

[Li & **H** 2023+]

Consistency of STARE

Definition: $\pi_{\star}^{(K)}, \phi_{\star 1}^{(K)}, \dots, \phi_{\star K}^{(K)}$ = asymptotically inferred parameters *given K*

Key Assumptions: Exists decomposition of interest $P_{\circ} = \sum_{k=1}^{K_{\circ}} \pi_{\circ k} P_{\circ k}$ s.t.

[Li & **H** 2023+]

Consistency of STARE

Definition: $\pi_\star^{(K)}, \phi_{\star 1}^{(K)}, \ldots, \phi_{\star K}^{(K)}$ = asymptotically inferred parameters *given K*

Key Assumptions: Exists decomposition of interest $P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$ s.t.

(i) *model components close when K correct:*
$\mathrm{KL}(P_{\circ k} \,||\, F_{\phi_{\star k}^{(K)}}) < \rho$

[Li & **H** 2023+]

Consistency of STARE

Definition: $\pi_\star^{(K)}, \phi_{\star 1}^{(K)}, \ldots, \phi_{\star K}^{(K)}$ = asymptotically inferred parameters *given K*

Key Assumptions: Exists decomposition of interest $P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$ s.t.

(i) *model components close when K correct:*
$\mathrm{KL}(P_{\circ k} \,||\, F_{\phi_{\star k}^{(K)}}) < \rho$

(ii) *smaller mixtures are a poor fit:* for $\pi^{\mathrm{err}} := \|\pi_\circ - \pi_\star^{(K_\circ)}\|_1$,
if $K < K_\circ$, then $d_{\mathrm{BL}}\big(\sum_{k=1}^{K} \pi_{\star k}^{(K)} F_{\phi_{\star k}^{(K)}}, P_\circ\big) > (1 - \pi^{\mathrm{err}})\sqrt{\rho/2} + \pi^{\mathrm{err}}$

[Li & **H** 2023+]

Consistency of STARE

Definition: $\pi_\star^{(K)}, \phi_{\star 1}^{(K)}, \ldots, \phi_{\star K}^{(K)}$ = asymptotically inferred parameters *given K*

Key Assumptions: Exists decomposition of interest $P_\circ = \sum_{k=1}^{K_\circ} \pi_{\circ k} P_{\circ k}$ s.t.

(i) *model components close when K correct:*
$\mathrm{KL}(P_{\circ k} \,||\, F_{\phi_{\star k}^{(K)}}) < \rho$

(ii) *smaller mixtures are a poor fit:* for $\pi^{\mathrm{err}} := \|\pi_\circ - \pi_\star^{(K_\circ)}\|_1$,
if $K < K_\circ$, then $d_{\mathrm{BL}}\left(\sum_{k=1}^{K} \pi_{\star k}^{(K)} F_{\phi_{\star k}^{(K)}}, P_\circ\right) > (1 - \pi^{\mathrm{err}})\sqrt{\rho/2} + \pi^{\mathrm{err}}$

Theorem **[Li & H 2023+]:** As $N \to \infty$, $\Pr(\hat{K}_\rho = K_\circ) \to 1$.

Consistency of STARE

Definition: $\pi_{\star}^{(K)}, \phi_{\star 1}^{(K)}, \ldots, \phi_{\star K}^{(K)}$ = asymptotically inferred parameters *given K*

Key Assumptions: Exists decomposition of interest $P_{\circ} = \sum_{k=1}^{K_{\circ}} \pi_{\circ k} P_{\circ k}$ s.t.

(i) *model components close when K correct:*
$\mathrm{KL}(P_{\circ k} \,||\, F_{\phi_{\star k}^{(K)}}) < \rho$

(ii) *smaller mixtures are a poor fit:* for $\pi^{\mathrm{err}} := \|\pi_{\circ} - \pi_{\star}^{(K_{\circ})}\|_1$,
if $K < K_{\circ}$, then $d_{\mathrm{BL}}\big(\sum_{k=1}^{K} \pi_{\star k}^{(K)} F_{\phi_{\star k}^{(K)}}, P_{\circ}\big) > (1 - \pi^{\mathrm{err}})\sqrt{\rho/2} + \pi^{\mathrm{err}}$

Theorem [Li & H 2023+]: As $N \to \infty$, $\Pr(\hat{K}_{\rho} = K_{\circ}) \to 1$.

- Special case of our general consistency result

Skew-Gaussian mixture: STARE

$$\mathscr{L}_{\rho}(\hat{\theta} \mid z) = \sum_{k=1}^{K} |X_k(z)| \max\left\{0, \widehat{\mathrm{KL}}(X_k(z) \mid F_{\hat{\phi}_k}) - \rho\right\} + \lambda K$$

[Li & **H** 2023+]

Skew-Gaussian mixture: STARE

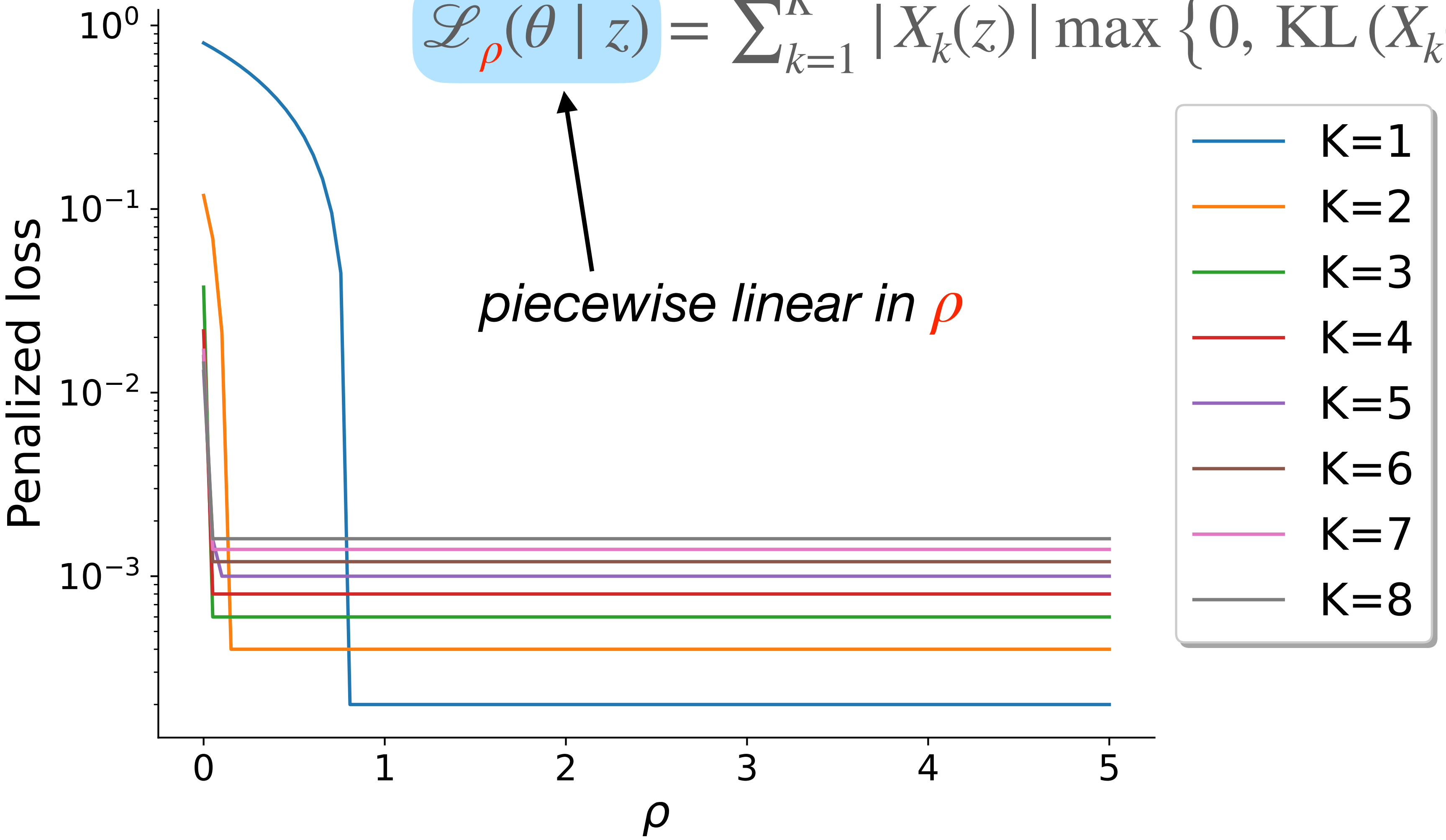
$$\mathscr{L}_{\rho}(\hat{\theta} \mid z) = \sum_{k=1}^{K} |X_k(z)| \max\left\{0, \widehat{\mathrm{KL}}(X_k(z) \mid F_{\hat{\phi}_k}) - \rho\right\} + \lambda K$$

piecewise linear in ρ

[Li & **H** 2023+]

Skew-Gaussian mixture: STARE

$$\mathscr{L}_{\rho}(\hat{\theta} \mid z) = \sum_{k=1}^{K} |X_k(z)| \max\left\{0, \widehat{\mathrm{KL}}(X_k(z) \mid F_{\hat{\phi}_k}) - \rho\right\} + \lambda K$$

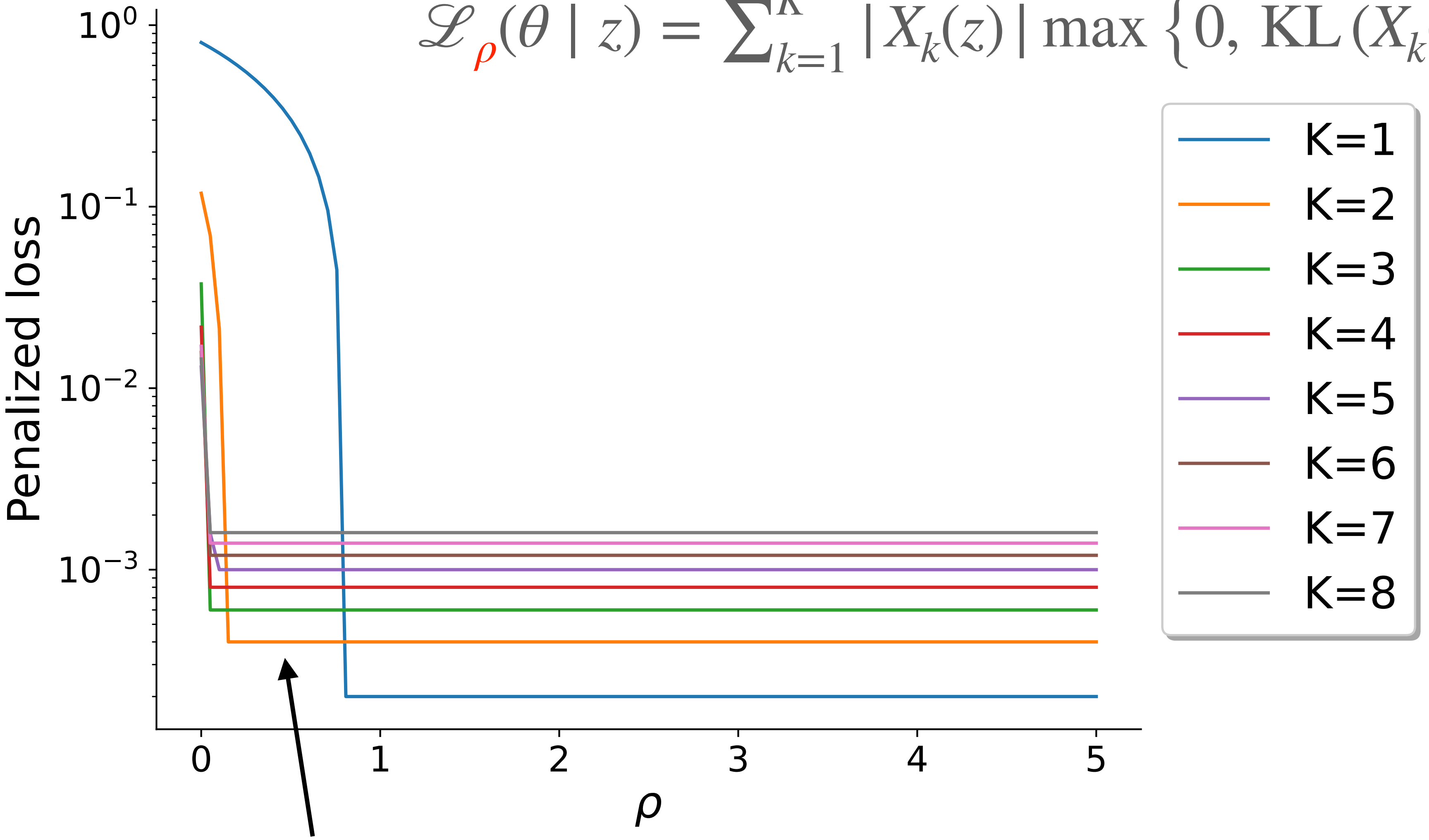


piecewise linear in ρ

[Li & **H** 2023+]

Skew-Gaussian mixture: STARE

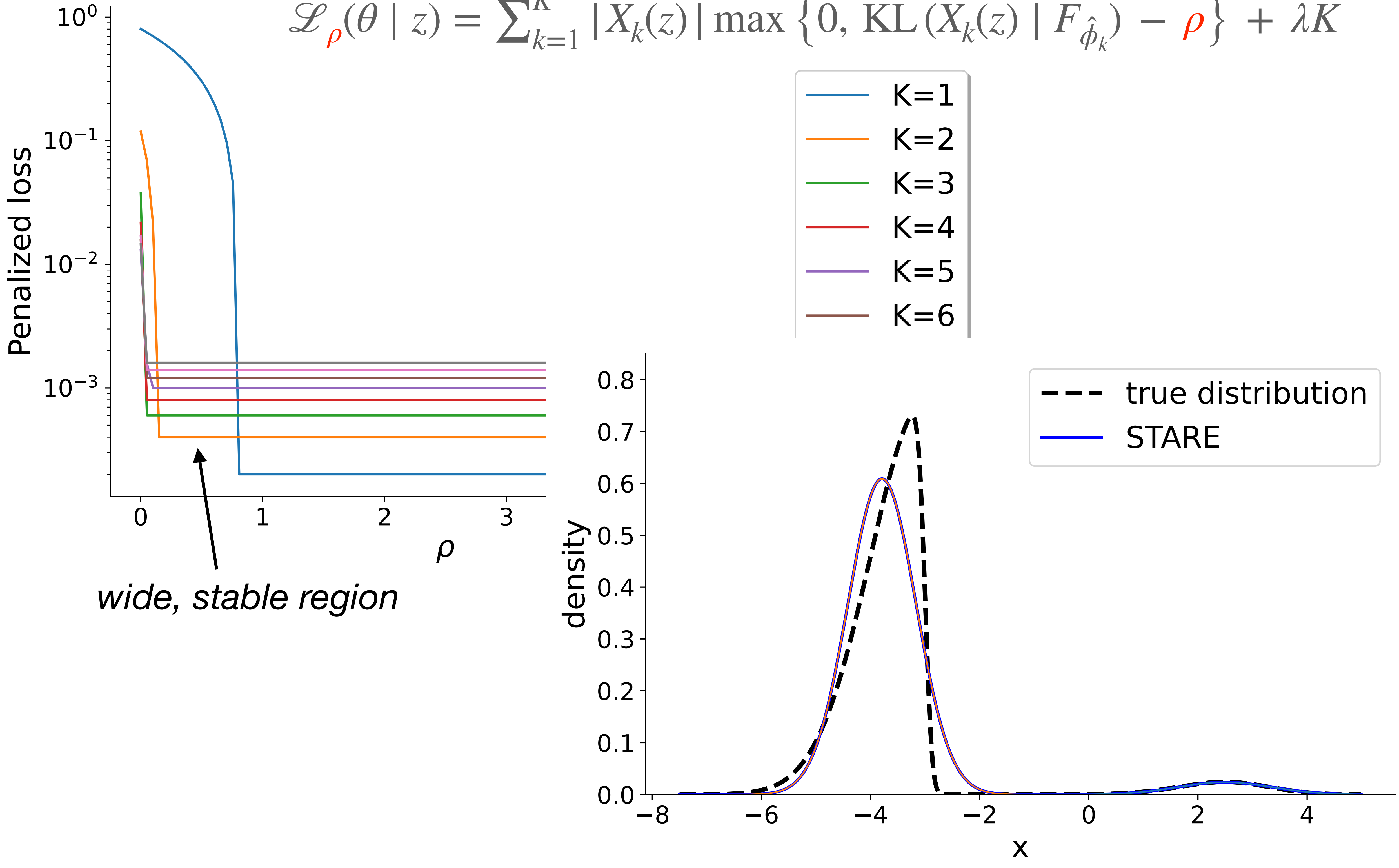
$$\mathscr{L}_{\rho}(\hat{\theta} \mid z) = \sum_{k=1}^{K} |X_k(z)| \max\left\{0, \widehat{\mathrm{KL}}\left(X_k(z) \mid F_{\hat{\phi}_k}\right) - \rho\right\} + \lambda K$$



wide, stable region

Skew-Gaussian mixture: STARE

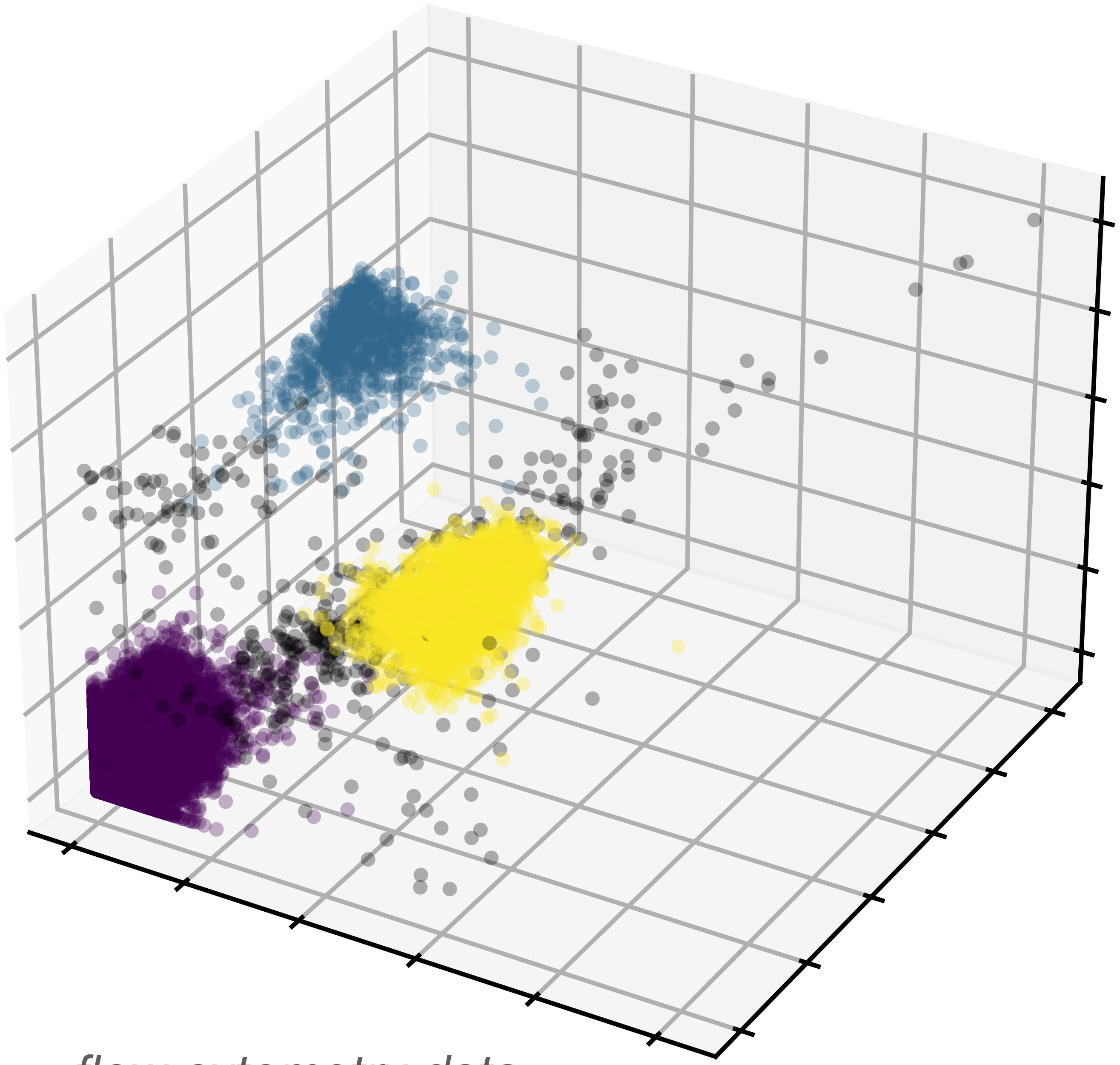
$$\mathscr{L}_{\rho}(\hat{\theta} \mid z) = \sum_{k=1}^{K} |X_k(z)| \max\left\{0, \widehat{\mathrm{KL}}(X_k(z) \mid F_{\hat{\phi}_k}) - \rho\right\} + \lambda K$$



wide, stable region

[Li & **H** 2023+]

Clustering cells by type: calibrating α

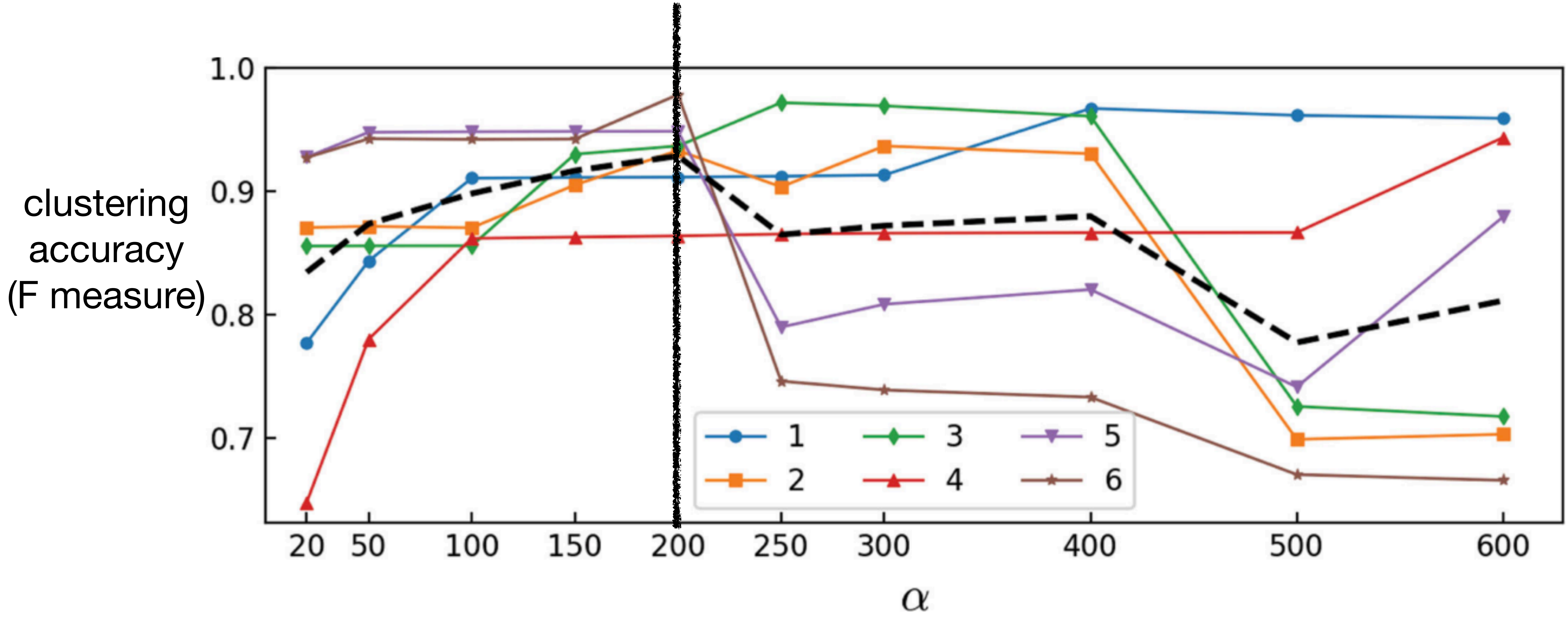


flow cytometry data

[Miller & Dunson 2018]

Clustering cells by type: calibrating α

$\alpha = 200$

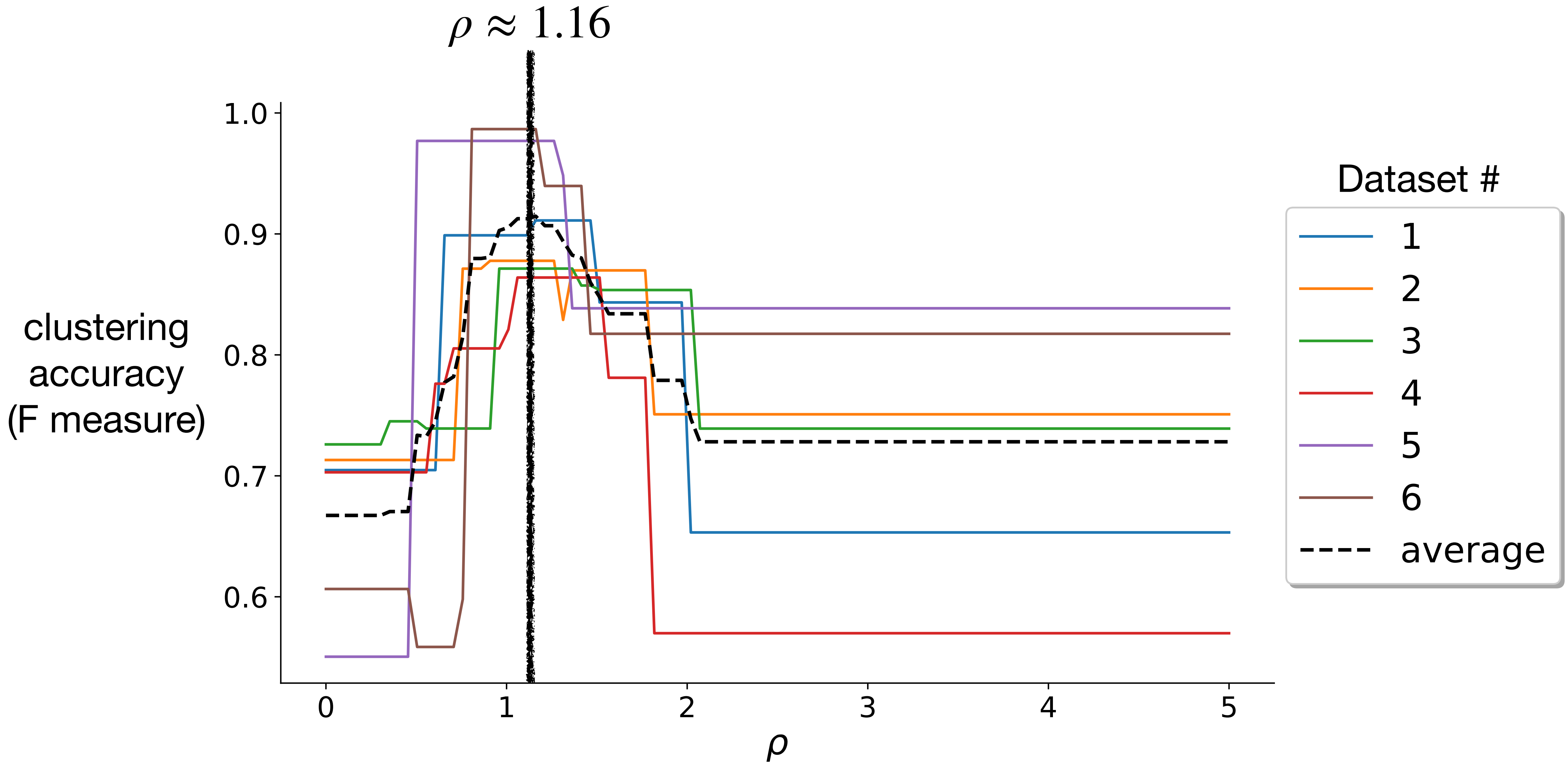


[Miller & Dunson 2018]

Clustering cells by type: calibrating ρ

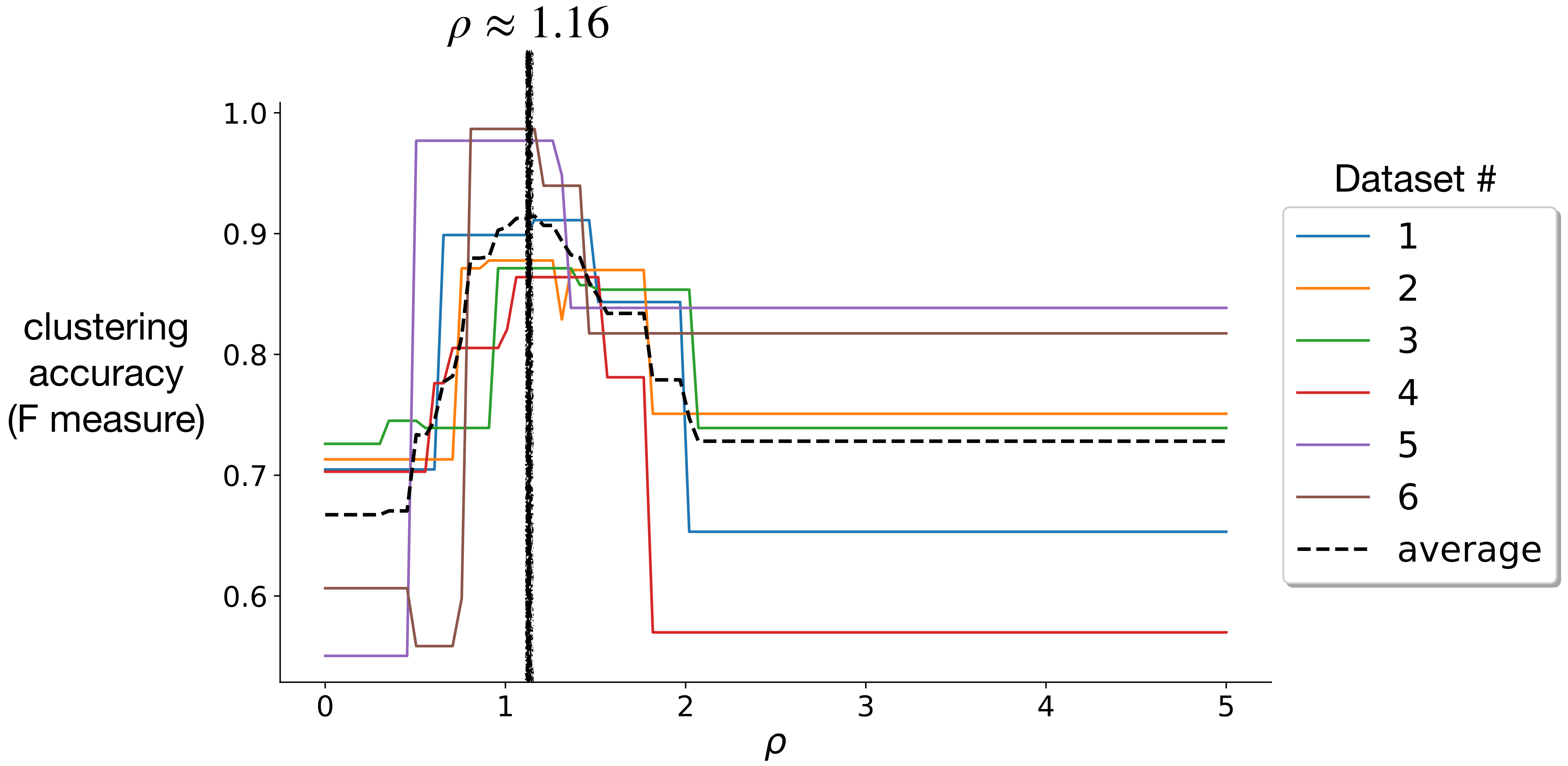
[Li & **H** 2023+]

Clustering cells by type: calibrating ρ



[Li & **H** 2023+]

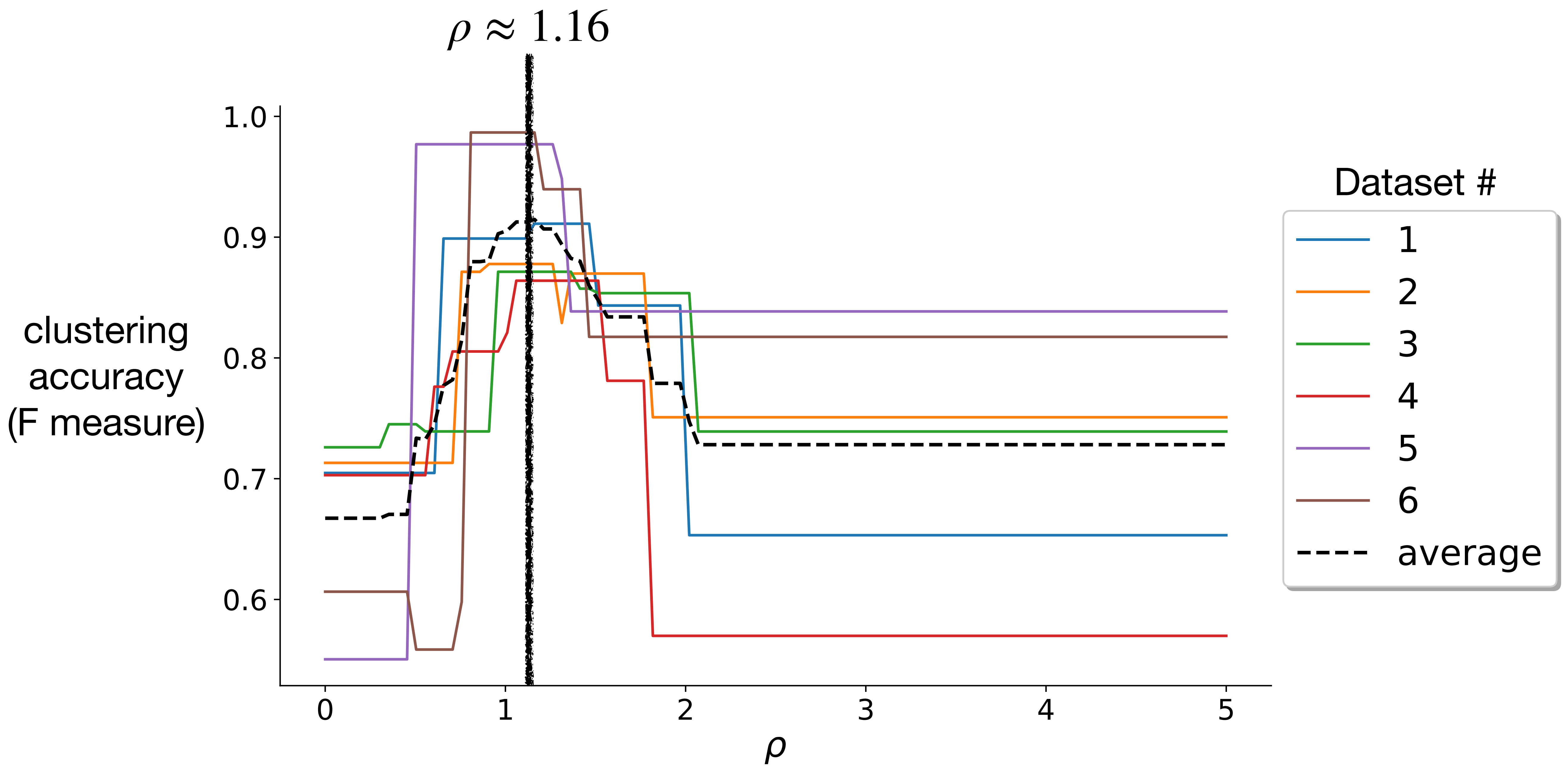
Clustering cells by type: calibrating ρ



- STARE runtime: 30 min [Python]

[Li & **H** 2023+]

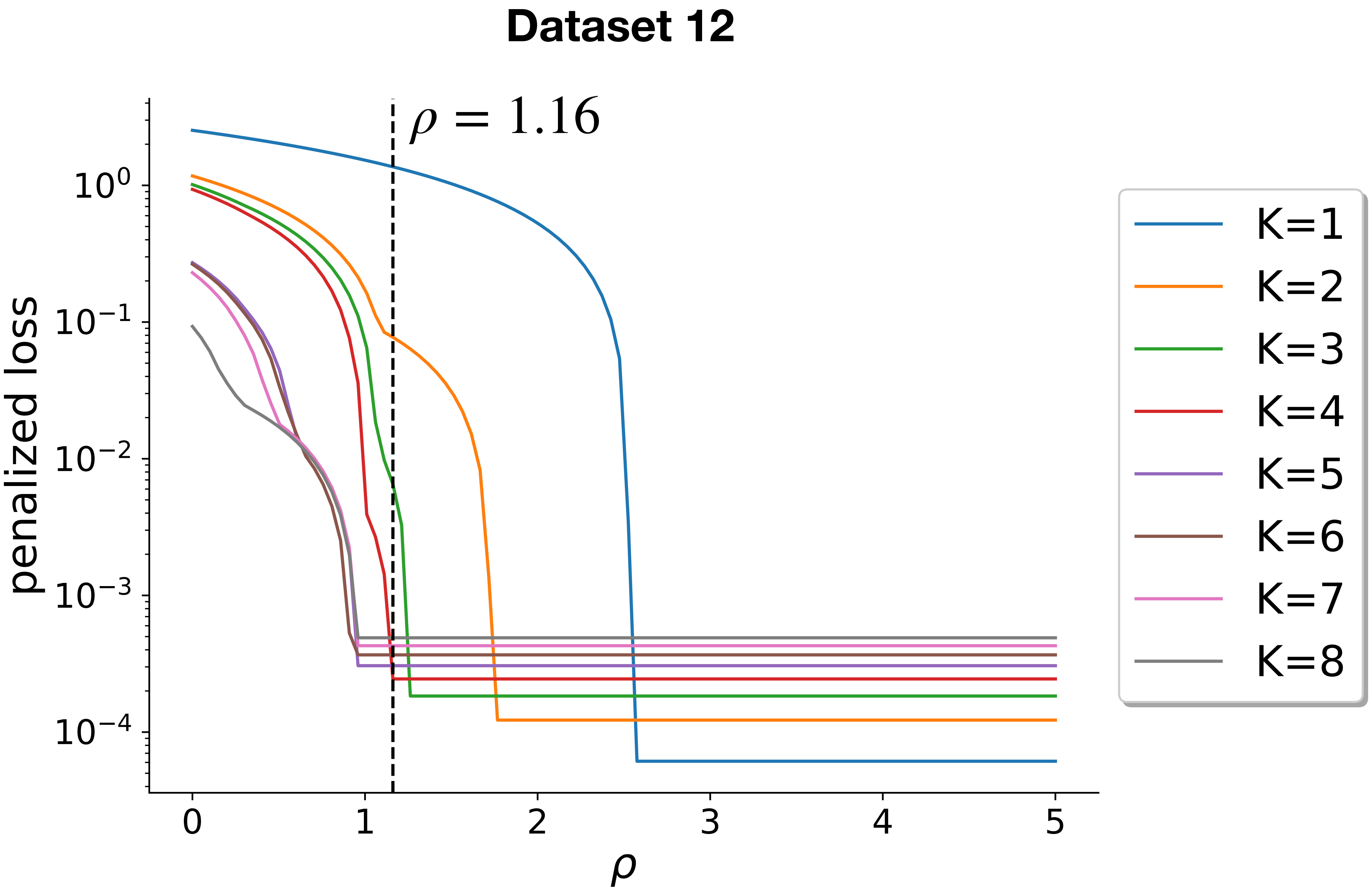
Clustering cells by type: calibrating ρ



- STARE runtime: 30 min [Python]
- Coarsened posterior runtime: 2.5 hours [Julia]

[Li & **H** 2023+]

A borderline case



[Li & **H** 2023+]

Clustering cells by type: STARE is fast and accurate

Clustering accuracy (F-measure)

Dataset	7	8	9	10	11	12
STARE	0.63	**0.92**	**0.94**	**0.99**	**0.99**	**0.98**
Coarsened posterior	**0.67**	0.88	**0.93**	**0.99**	**0.99**	**0.99**

[Miller & Dunson 2018, Li & **H** 2023+]

Closing thoughts

Closing thoughts

- Model misspecification can dramatically affect stability/reproducibility of inferences
    - Ideally, want to default to **stable methods** that don't degrade **statistical efficiency**
    - Examples: **Bagged posterior** and **STARE**

Closing thoughts

- Model misspecification can dramatically affect stability/reproducibility of inferences
  - Ideally, want to default to **stable methods** that don't degrade **statistical efficiency**
  - Examples: **Bagged posterior** and **STARE**

References

J. H. Huggins & J. W. Miller (2023). Reproducible Model Selection Using Bagged Posteriors. *Bayesian Analysis* 18(1): 79–104

J. H. Huggins & J. W. Miller (2019). Robust Inference and Model Criticism Using Bagged Posteriors. arXiv:1912.07104 [stat.ME].

J. H. Huggins & J. W. Miller (2023+). Reproducible Parameter Inference Using Bagged Posteriors.

J. Li & J. H. Huggins (2023+). Robust, Structurally-Aware Inference with Mixture Models.