How Should We Construct Prediction Sets?
Insights from Conformal Prediction

March 6, 2025

Tiffany Ding
Department of Statistics
UC Berkeley

What this talk will be

Philosophical musings about the *role of prediction sets in decision-making*

+

An overview of *"Class-Conditional Conformal Prediction with Many Classes"* (NeurIPS 2023)

Part 1

Part 2

Key themes

The goal of statistics should be to **inform decision-making**

Thinking "statistically" allows us to **naturally arrive at new (and hopefully *useful*) methodologies.**

"Usefulness" is nuanced but important to think about as we design statistical methods

I. Prediction Sets

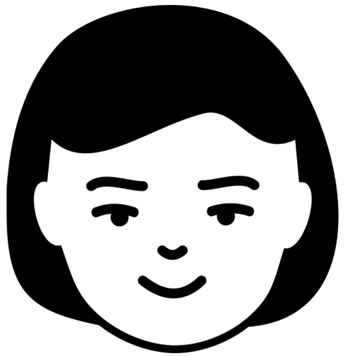
Motivating prediction sets from a
***practical**, rather than statistical, perspective*

Why prediction sets?

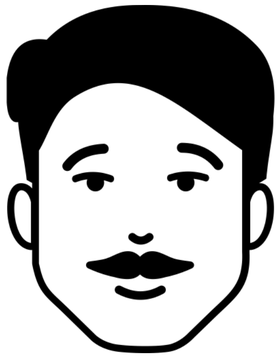
A motivating example

🌱 Suppose you want to identify plants around town,
but you currently know nothing about plants.

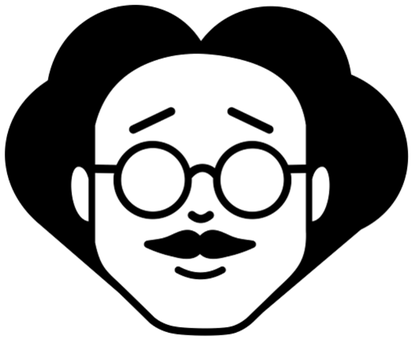
Luckily, you are friends with some machine learning researchers who offer to help you.



Alice



Bob



Cameron

Alice’s algorithm gives you **a single best guess**

Prickly acanthus

❌

Echium virescens

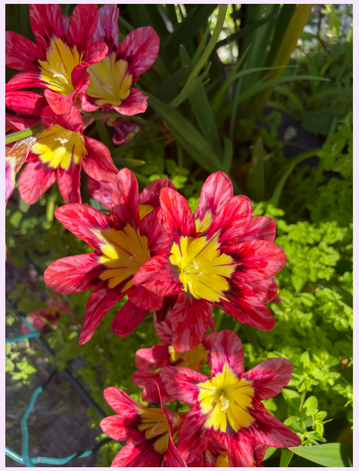
✅

Frangipani

❌

Saint Martin’s lily

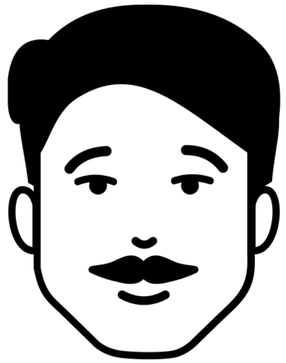
❌



Three-color harlequin flower

✅

Bob’s algorithm gives you **several guesses**



Spiny bear’s-breeches OR prickly acanthus

✅

Echium virescens

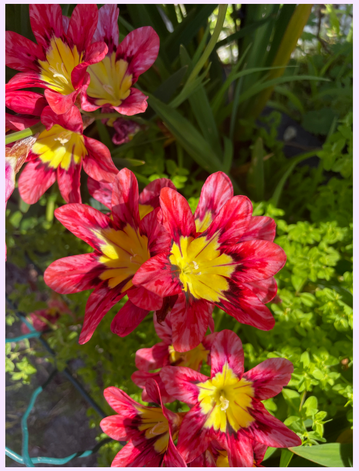
✅

Frangipani OR miconia

✅

Saint Martin’s lily OR blackeyed Susan

❌

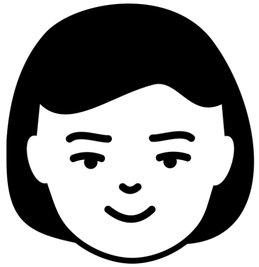


Three-color harlequin-flower OR Painted-tongue OR Eyed tulip

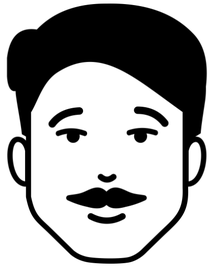
✅

Bob, I like how your prediction sets contain the correct species most of the time.

You

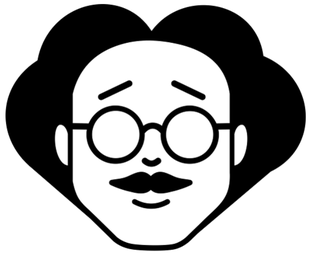


Alice



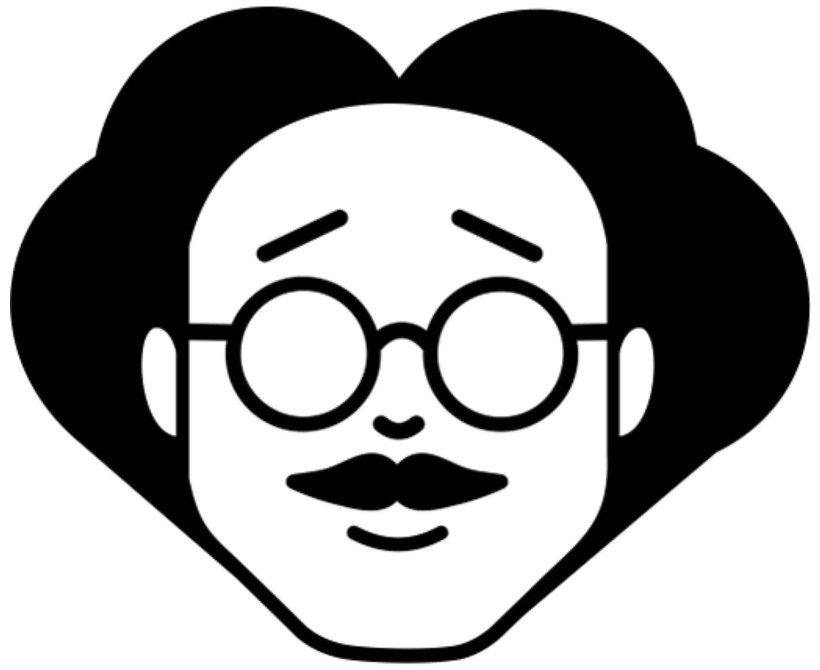
Bob

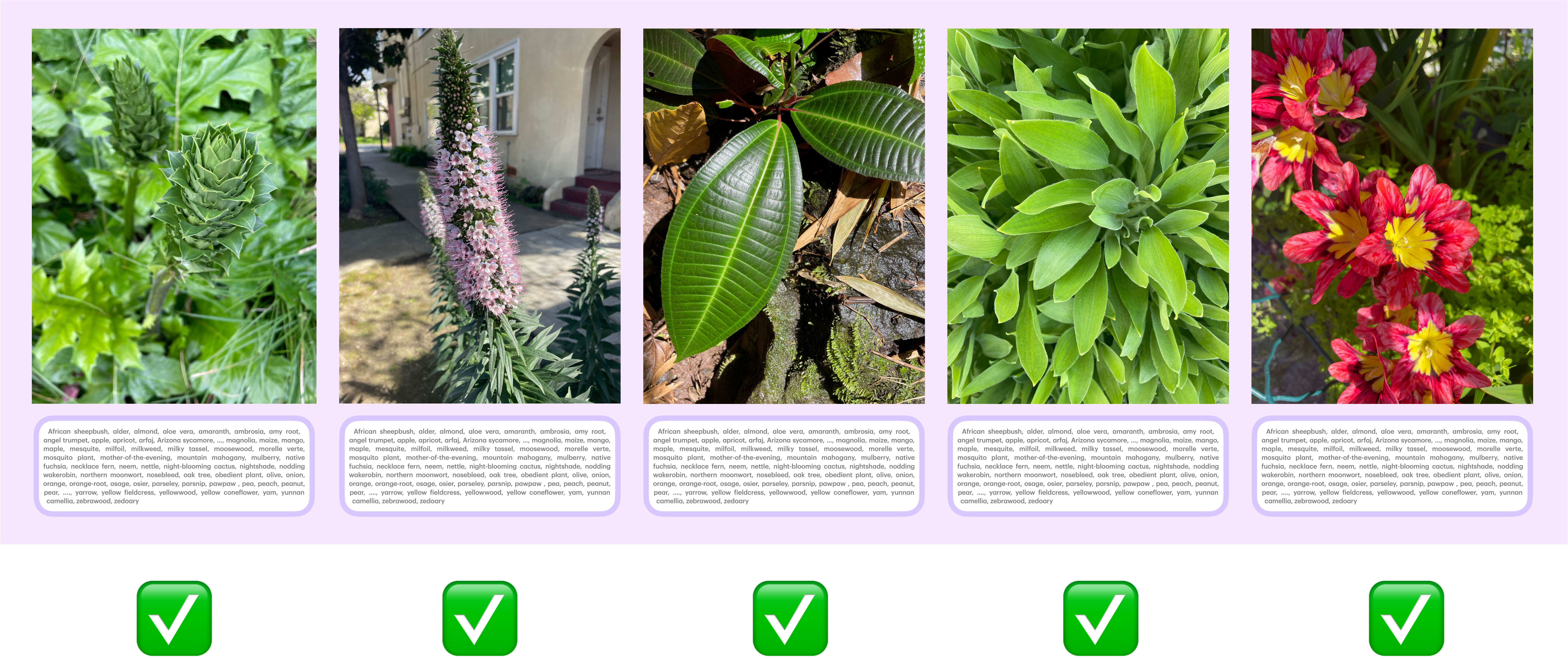
I can do even better! My sets contain the correct species **100%** of the time.



Cameron

Cameron’s algorithm gives you **a list of candidates that *always* contains the correct label**





but these sets aren’t useful at all...

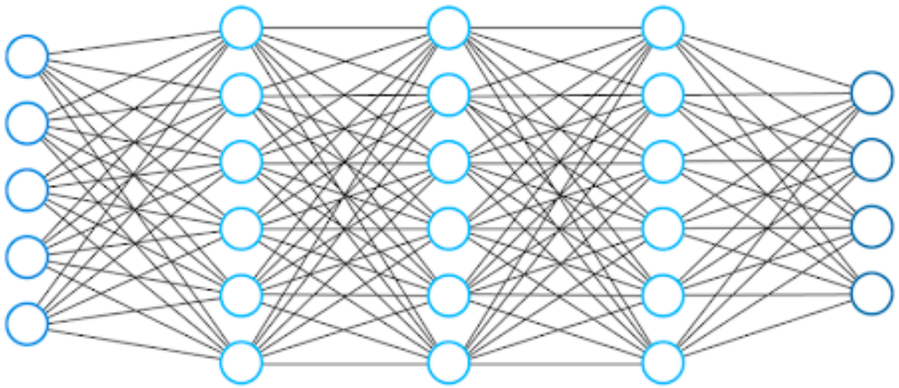
What can we learn from this example?

- In many settings,

prediction sets > point predictions

- How you construct the prediction set matters (and higher coverage does not necessarily mean better)

Now let's think more carefully about constructing prediction sets



→ [0.03, 0.23, 0.02, ..., 0.04, 0.13]

softmax vector

How can we go from the softmax vector to a prediction set?

"top-k prediction set"

A reasonable approach: select the classes with the k largest softmax scores (for some $k > 0$)

top-k prediction sets

Suppose we have four classes: 1,2,3,4

True label	softmax vectors	top-3 pred. set
1	[**.98, .01, .006**, .004]	[1, 2, 3]
1	[**.3, .4, .2,** .1]	[1, 2, 3]
3	[.03, **.4, .4, .17**]	[2, 3, 4]

By definition, these top-k prediction sets are **always the same size**.
It feels like in the first example, the set should be smaller for the first example because the model is more "confident."

*What if instead of always taking the top k classes, we **include classes that have softmax scores above a certain threshold**?*

softmax-thresholded sets

True label	softmax vectors	top-3	0.9-thresholded	0.01-thresholded	0.3-thresholded
1	[.98, .01, .006, .004]	[1, 2, 3]	[1]	[1, 2]	[1]
1	[.3, .4, .2, .1]	[1, 2, 3]	[ ]	[1, 2, 3, 4]	[1, 2, 3]
3	[.03, .4, .4, .17]	[2, 3, 4]	[ ]	[1, 2, 3, 4]	[2, 3]
			too small	too large	seems good

How should we choose the threshold?

Conformal prediction sets

Standard CP

selecting the softmax threshold in a **principled, data-dependent way** = conformal prediction

In words

1. Choose a miscoverage level $\alpha > 0$.
2. Get some *labeled* calibration data.
3. Choose the threshold $\hat{q}$ such that ~ $(1-\alpha) \times 100\,\%$ of $\hat{q}$-thresholded sets constructed on the calibration data contain the true label
4. At test time, return $\hat{q}$-thresholded set based on softmax scores for that input

In math

Given:

α — Miscoverage level

$s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ — Conformal score function*

$(X_1, Y_1), \ldots, (X_n, Y_n)$ — Calibration data

1. Compute conformal scores $s_i := s(X_i, Y_i)$ for $i = 1, \ldots, n$
2. Let $\hat{q} = \lceil (1-\alpha)(n+1) \rceil$ largest s_i
3. At test time, construct prediction set as

$$C(X_{\text{test}}) = \{y : s(X_{\text{test}}, y) \leq \hat{q}\}$$

negatively oriented*, i.e., **higher score = **more disagreement** between prediction and label

The statistical motivation for conformal prediction

If the calibration data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and test point $(X_{\text{test}}, Y_{\text{test}})$ are **iid** (or, more generally, the data are ***exchangeable***),

then the prediction sets generated by the procedure from the last side are guaranteed to satisfy

$$\mathbb{P}(Y \in C(X)) \geq 1 - \alpha$$

"Marginal coverage"

where the probability is over the randomness in the calibration data and the test point.

Interlude: what makes a prediction set "useful"?

To answer this, we must first answer

"What are some ways a prediction set may be used?"

Regression Examples

1. The decision is exactly **equal to the upper/lower end of the prediction interval** (e.g., ordering inventory)
2. The decision is determined by **whether the prediction interval crosses a certain threshold** (e.g., If the 90% prediction interval crosses below 0 Celsius, move your plants indoors)
3. The prediction interval/region is **used as a constraint set** in an optimization problem that is solved to obtain a decision (e.g., robot motion planning; see Lars Lindemann's work)

Classification Examples

1. **Human verifies whether each label in the prediction set is correct** (e.g., a setting in which search is harder than verification)
2. Same as 1, but **only using the set for cases where they cannot figure it out themselves**
3. **"Sanity check"**: If the human's initial guess is in the prediction set, that is their final guess
4. **(Automated) action is fully determined by the classes included in the prediction set** (e.g., movie recommendation system)
5. The decision of **whether to collect more info/expend more effort make a final prediction is made based on the prediction set** (e.g., medical diagnosis)

Note: 1-3 are all ways that people reported using prediction sets for classifying ImageNet images (Zhang, Chatzimparmpas, Kamali, & Hullman, 2024)

Interlude: what makes a prediction set "useful"?

Q: Can "usefulness" be measured?

"Usefulness" can refer to:

1. **A reduction in effort** to make the decision
   [Note: only relevant for human (not automated) decision-making]
2. **An improvement in decision quality**
3. Some **combination** of 1 and 2

→ Often measured via time taken

→ Often measured via prediction accuracy — but this is not obviously the right choice in all settings. Should depend on the type of decision.

Once you define the appropriate metric(s) for your decision-making setting, you can run RCTs or A/B tests to compare the performance of different prediction (set) generating procedures.

A failure mode of the aforementioned procedures

classes that consistently have small softmax scores may not appear in any prediction sets

This is true for top-k, softmax-thresholding, and standard conformal prediction

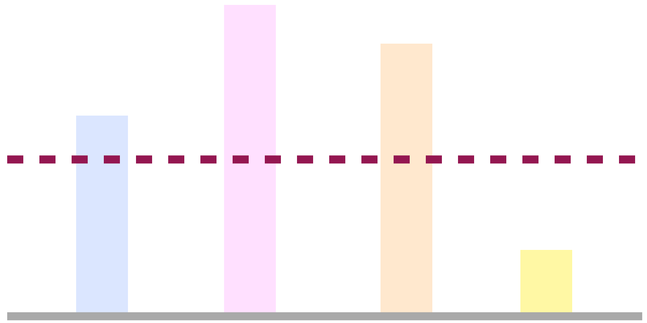
Example of when this may occur:

- Class 1 and Class 2 look very similar but Class 1 is very common whereas Class 2 is very rare.
- An ML model may very reasonably only ever assign high softmax scores to Class 1 and low softmax scores to Class 2

When this could be a problem:

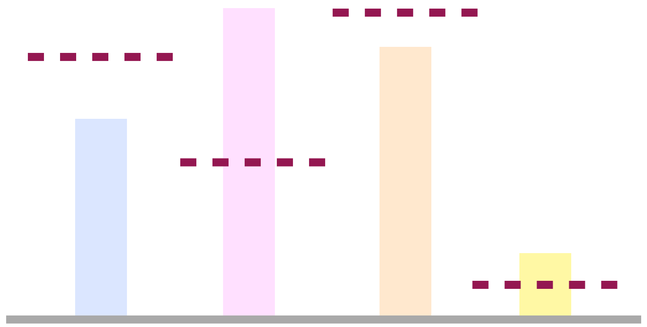
- Ex 1: Doctor wants to catch less common diseases that are very harmful if not treated
- Ex 2: ML researchers wants to identify more training examples of less common classes in order to train their model to do better on those classes

A solution to this failure mode



Same threshold for all classes

(e.g., top-k, softmax-thresholding, standard conformal prediction)



A solution:

Set different thresholds for each class

How should we choose the threshold for each class?

Classwise conformal prediction

Intuitively,

class tends to get high softmax scores (when it is the true label) → set high threshold

Essentially, we want to estimate the $1 - \alpha$ *quantile of the score distribution* ***of each class***

Formalizing this intuition,

CLASSWISE CP

1. Split calibration data by class.
2. Estimate separate $\hat{q}_y$ for each class.
3. Construct prediction sets as $C_{\text{CLASSWISE}}(X_{\text{test}}) = \{y : s(X_{\text{test}}, y) \leq \hat{q}_y\}$

Statistical guarantee: $\mathbb{P}(Y \in C_{\text{CLASSWISE}}(X) \mid Y = y) \geq 1 - \alpha$ for all classes $y \in \mathcal{Y}$

"Class-conditional coverage"

Q: Why is class-conditional conformal prediction hard?

A: We often have **limited data *per class***

e.g., ImageNet has 1000 classes, so if you have 10,000 calibration images, that's only 10 per class.

II. Class-Conditional Conformal Prediction with Many Classes

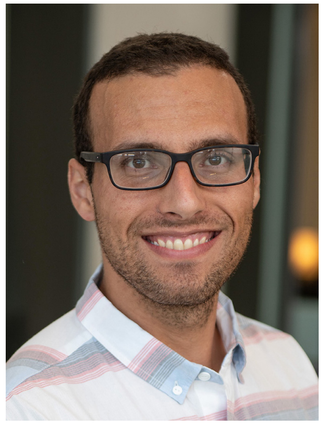
(NeurIPS 2023)

Joint work with

Anastasios Angelopoulos

Stephen Bates

Michael Jordan



Ryan Tibshirani

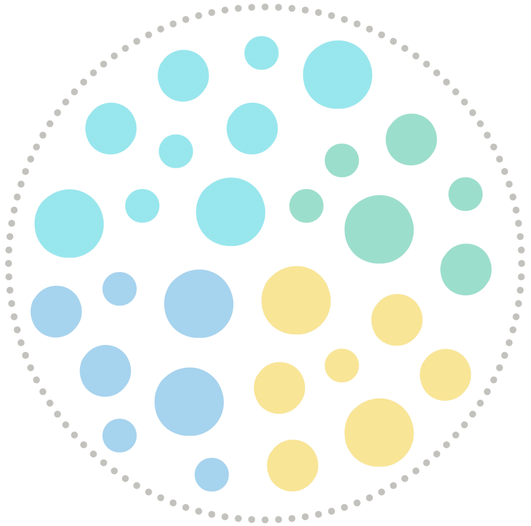
Goal: class-conditional coverage

Given features $X \in \mathcal{X}$ and unknown label $Y \in \mathcal{Y}$, we want a prediction set $C(X)$ with class-conditional coverage for some small $\alpha > 0$:

$$\mathbb{P}(Y \in C(X) \mid Y = y) \geq 1 - \alpha$$

$$\text{for all classes } y \in \mathcal{Y}$$

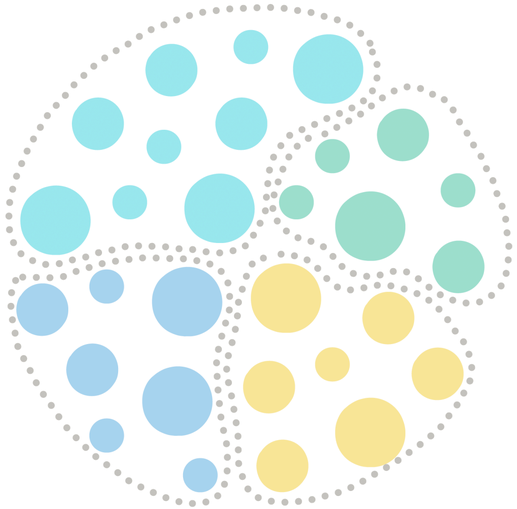
The idea behind our method



Standard CP

✅ Low variance

☹️ No class-conditional coverage guarantee



Our method: Clustered CP

🔑 **Key idea:**

Combine data from classes that are "similar"

Classwise CP

☹️ High variance

✅ Class-conditional coverage guarantee

Clustered CP

(in one line)

$$C_{\text{CLUSTERED}}(X_{\text{test}}) = \{y : s(X_{\text{test}}, y) \leq \hat{q}(\hat{h}(y))\}$$

where

- $\hat{h} : \mathcal{Y} \to \{1, \ldots, M\}$ is a clustering function
- $\hat{q}(m)$ is the conformal quantile computed using the calibration data in cluster m

How should we design our clustering function $\widehat{h}$?

For any $\widehat{h}$, we get cluster-conditional coverage:

$$\mathbb{P}(Y_{\text{test}} \in C_{\text{CLUSTERED}}(X_{\text{test}}) \mid \widehat{h}(Y_{\text{test}}) = m) \geq 1 - \alpha$$

$$\text{for all clusters } m = 1, \ \ldots, \ M$$

?

But our goal is to get class-conditional coverage:

$$\mathbb{P}(Y_{\text{test}} \in C_{\text{CLUSTERED}}(X_{\text{test}}) \mid Y_{\text{test}} = y) \geq 1 - \alpha$$

$$\text{for all classes } y \in \mathcal{Y}$$

When does cluster-conditional coverage imply class-conditional coverage?

Proposition 1 (informally):

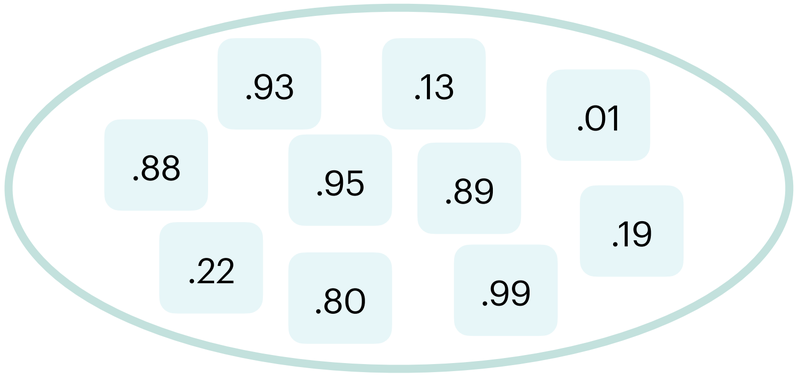
Let h^* be a clustering function such that **all classes assigned to the same cluster have conformal scores that are exchangeable**. Then, cluster-conditional coverage will imply class-conditional coverage.

In other words, we should group classes that have *similar score distributions*.

Designing clusters with exchangeable scores

Quantile-based clustering

Step 1: Create an embedding for the empirical score distribution of each class by creating a **vector of quantiles**.



Scores for class y

Take
[0.5, 0.6, 0.7, 0.8, 0.9]-quantiles

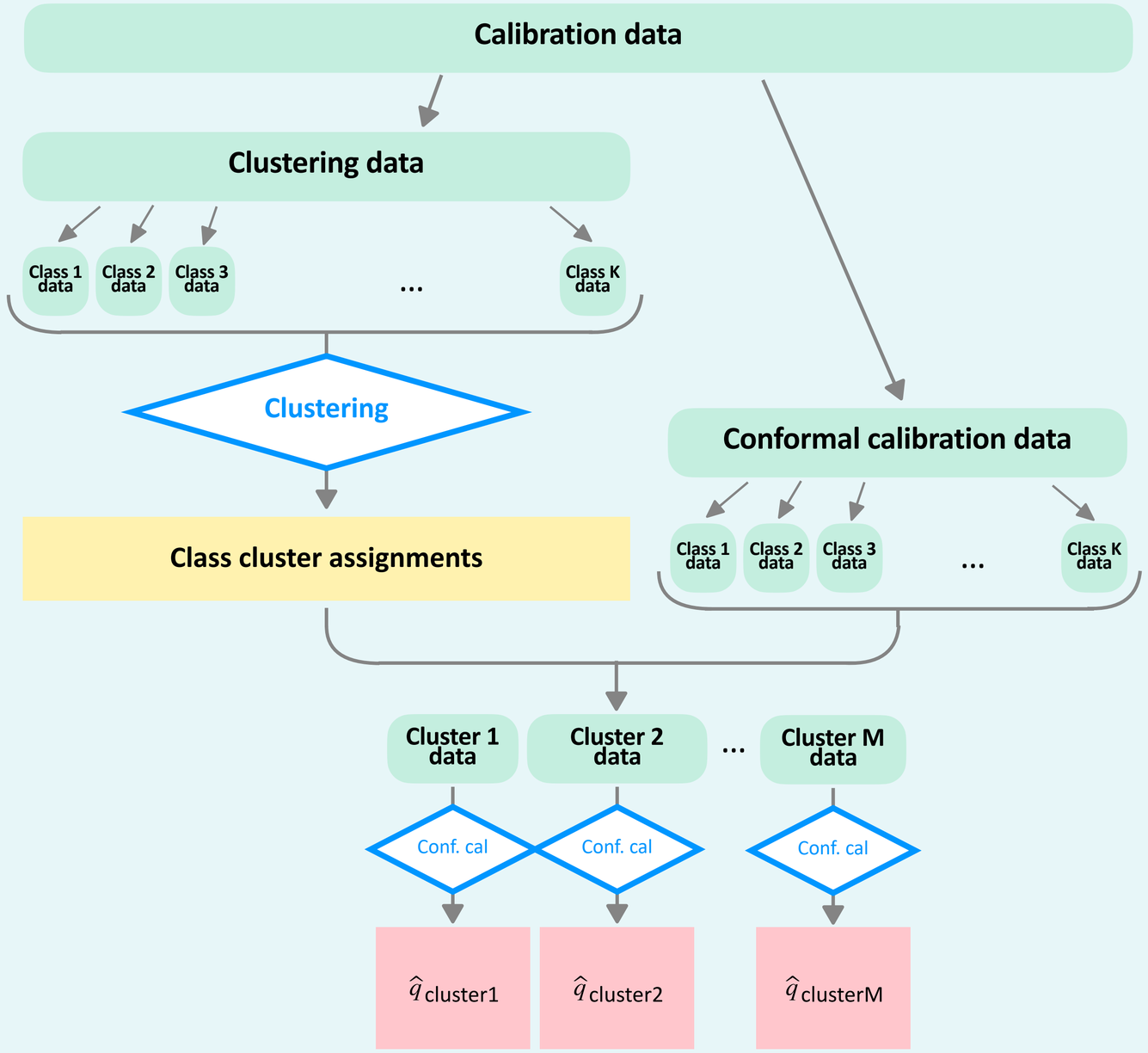
[0.88, 0.89, 0.93, 0.95, 0.99]

vector embedding of class y

Step 2: Apply **k-means** to these embeddings.

Clustered CP

(as a diagram)



What if we don't have perfect exchangeability within clusters?

Proposition 2: Let S^y denote a random variable sampled from the score distribution for class y. If the clusters given by $\widehat{h}$ satisfy

$$D_{\text{KS}}(S^y, S^{y'}) \leq \epsilon \qquad \text{for all } y, y' \text{ s.t. } \widehat{h}(y) = \widehat{h}(y'),$$

then $C_{\text{CLUSTERED}}$ will satisfy

$$P(Y_{\text{test}} \in C(X_{\text{test}}) \mid Y_{\text{test}} = y) \geq 1 - \alpha - \epsilon, \; \forall y \in \mathcal{Y} \; .$$

Note: The Kolmogorov-Smirnov distance of r.v.s X and Y is defined as

$$D_{\text{KS}}(X, Y) = \sup_{\lambda \in \mathbb{R}} |P(X \leq \lambda) - P(Y \leq \lambda)|$$

Experiments

Data sets and score functions

Data

Data set	ImageNet (Russakovsky et al., 2015)	CIFAR-100 (Krizhevsky, 2009)	Places365 (Zhou et al., 2018)	iNaturalist (Van Horn et al., 2018)
Number of classes	1000	100	365	663*
Class balance **	0.79	0.90	0.77	0.12
Example classes	mitten	orchid	beach	salamander
	triceratops	forest	sushi bar	legume
	guacamole	bicycle	catacomb	common fern

*The number of classes in the iNaturalist data set can be adjusted by selecting which taxonomy level (e.g., species, genus, family) to use as the class labels. We use the species family as our label and then filter out any classes with < 250 examples in order to have sufficient examples to properly perform evaluation.

**The class balance metric is defined as the number of examples in the rarest 5% of classes divided by the expected number of examples if the class distribution were perfectly uniform. This metric is bounded between 0 and 1, with lower values denoting more class imbalance

Conformal score functions

softmax: 1 - (softmax score of base classifier)

APS (Romano et al., 2020) : designed to achieve better X-conditional coverage

RAPS (Angelopoulos et al., 2022): regularized version of APS that often produces smaller sets

APS and RAPS in more detail

APS (Adaptive Prediction Sets, from Romano et. al, 2020):

- *Roughly*, "sum the probability mass until you reach the true label"
- *More specifically,* sort softmax scores and sum scores up until just before the score of the true label, then add Unif([0,1]) * score of true label

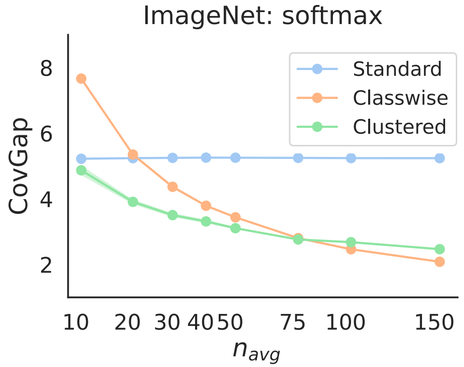
RAPS (Regularized APS, from Angelopoulos et. al, 2022):

$$\underbrace{\rho_x(y) + \hat{\pi}_x(y) \cdot u}_{\text{APS}} + \underbrace{\lambda \cdot (o_x(y) - k_{reg})^+}_{\text{regularization}}$$

ranking of y

A closer look at iNaturalist

Challenges: **many classes** and **extreme class imbalance** (the most common class has **275x** more images than the least common class)

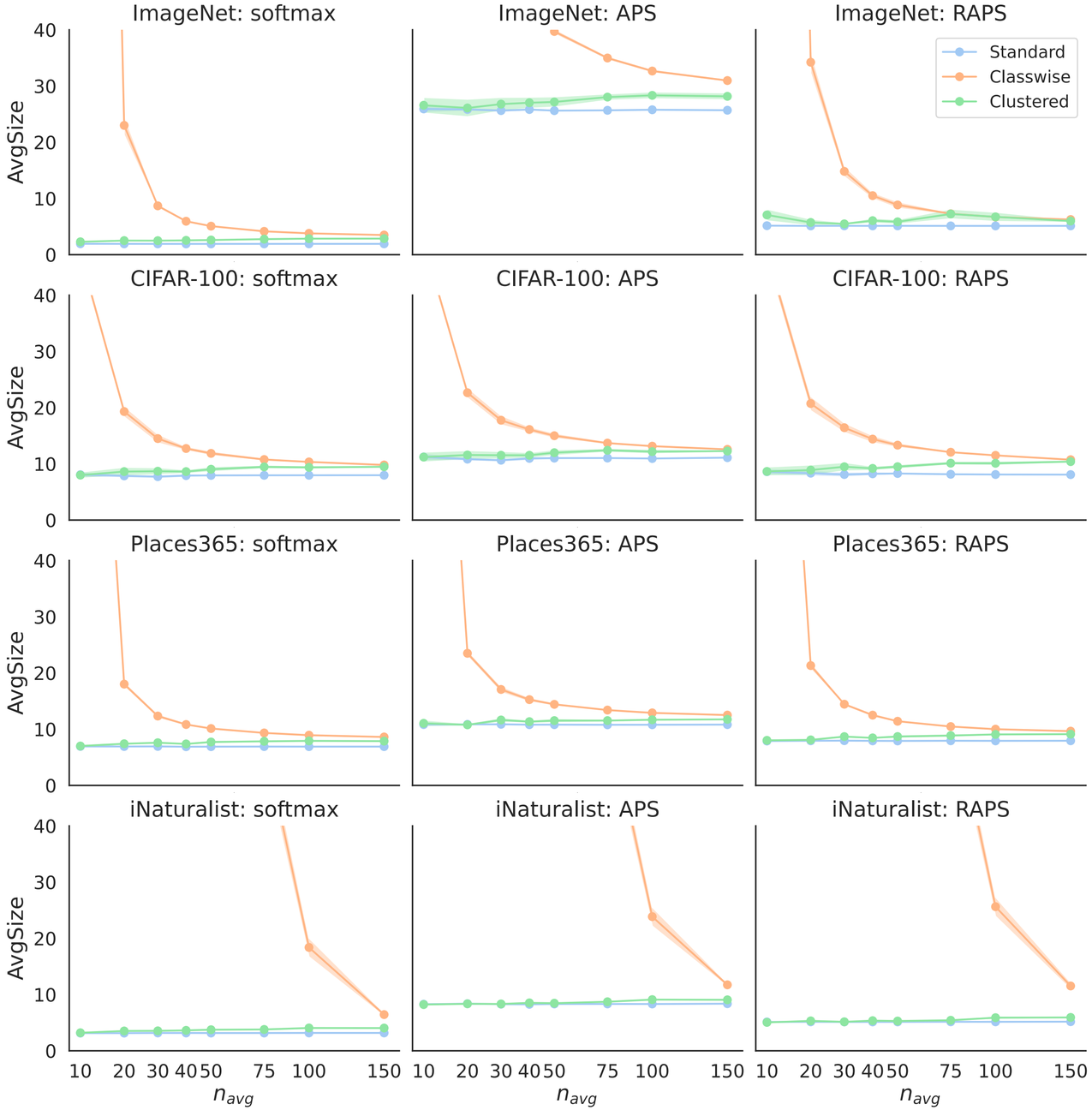


CovGap: *how far is the class-conditional coverage from our desired coverage level of* $(1-\alpha)$*?*

$$\text{CovGap} = 100 \times \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} |\hat{c}_y - (1-\alpha)|$$

where $\hat{c}_y$ is the coverage of class y, as computed on our validation dataset.

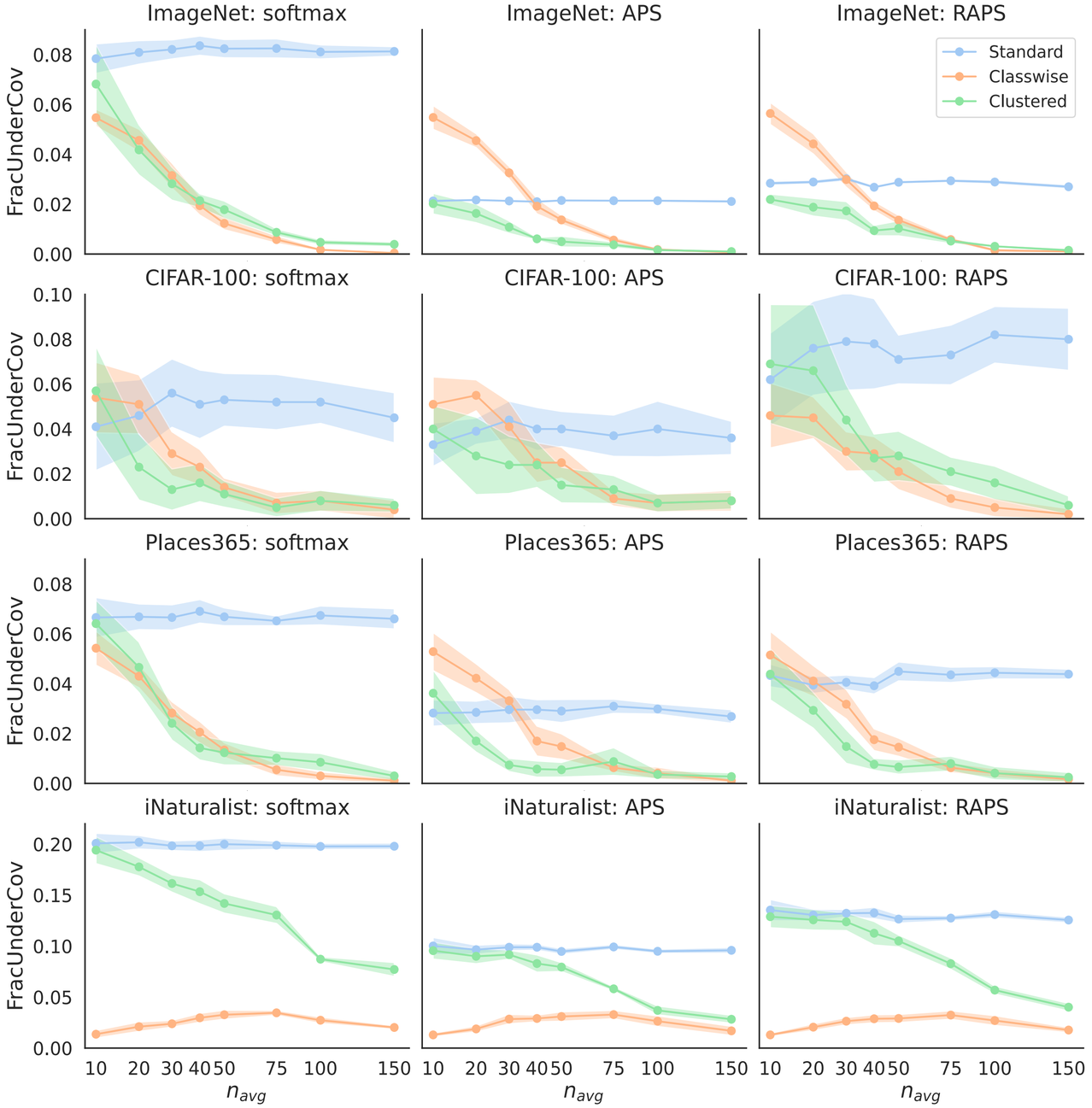
AvgSize: *what is the average size of the sets?*



FracUndercov: *what fraction of classes are severely* under-covered?*

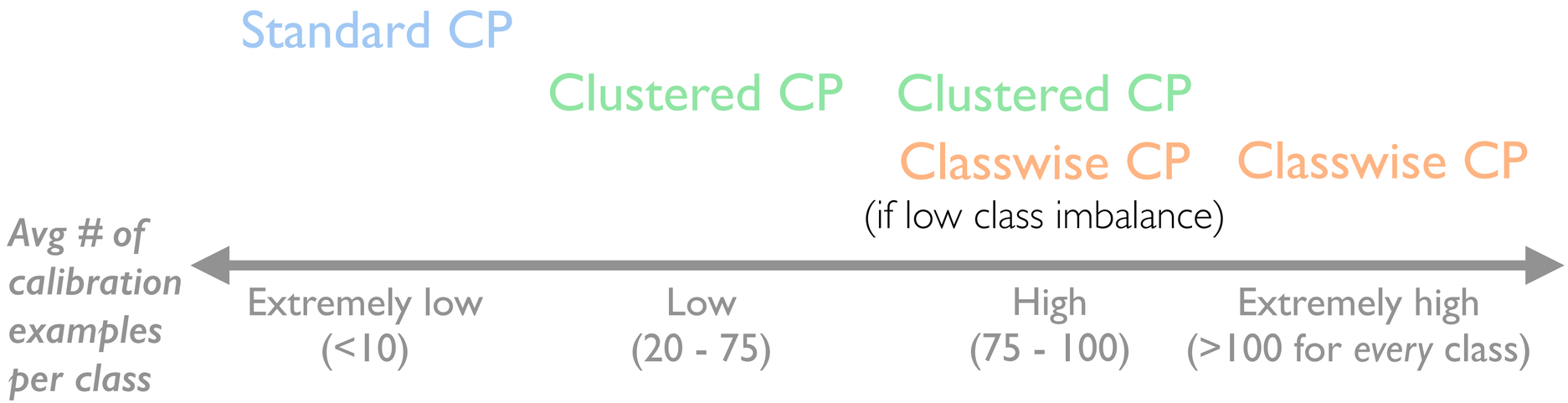
$$\text{FracUnderCov} = \frac{1}{|\mathcal{Y}|}\sum_{y=1}^{|\mathcal{Y}|} \mathbf{1}\{\hat{c}_y \leq 1 - \alpha - 0.1\}$$

* having a class-conditional coverage more than 10% below the desired coverage level



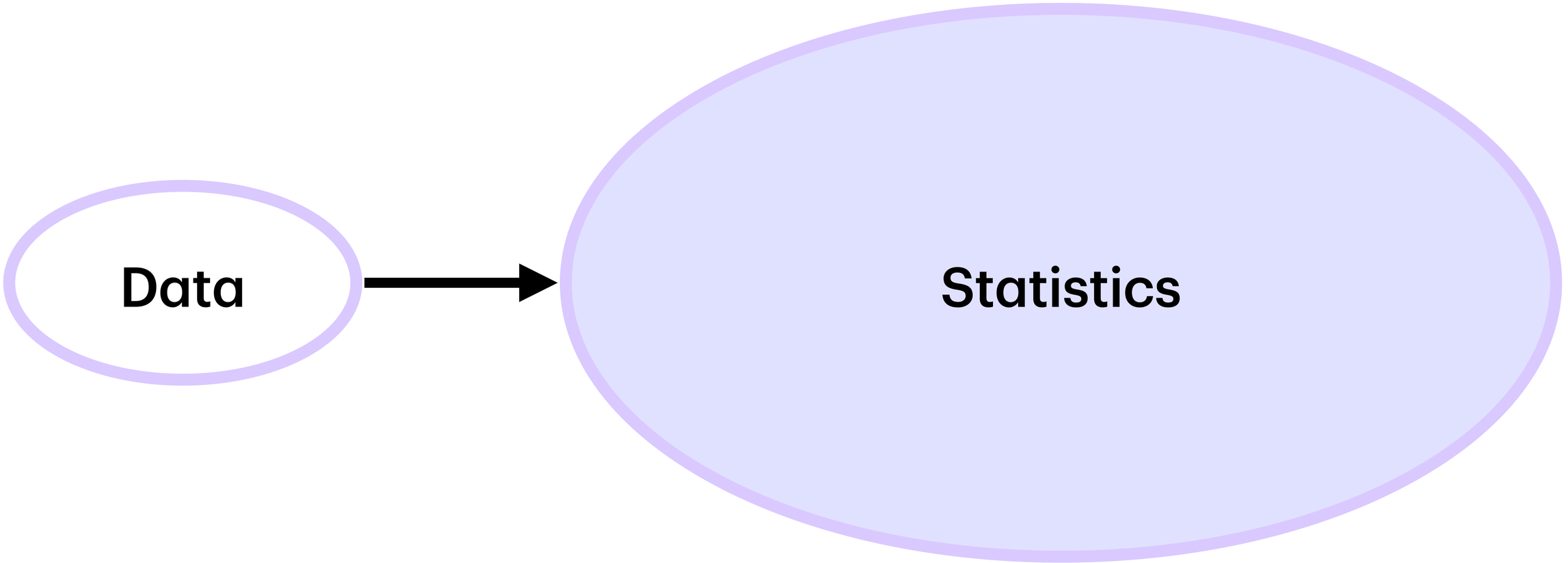
Recommendations for practitioners

For a given problem setting, what is the best way to produce prediction sets that have *good class-conditional coverage* but are *not too large to be useful*?



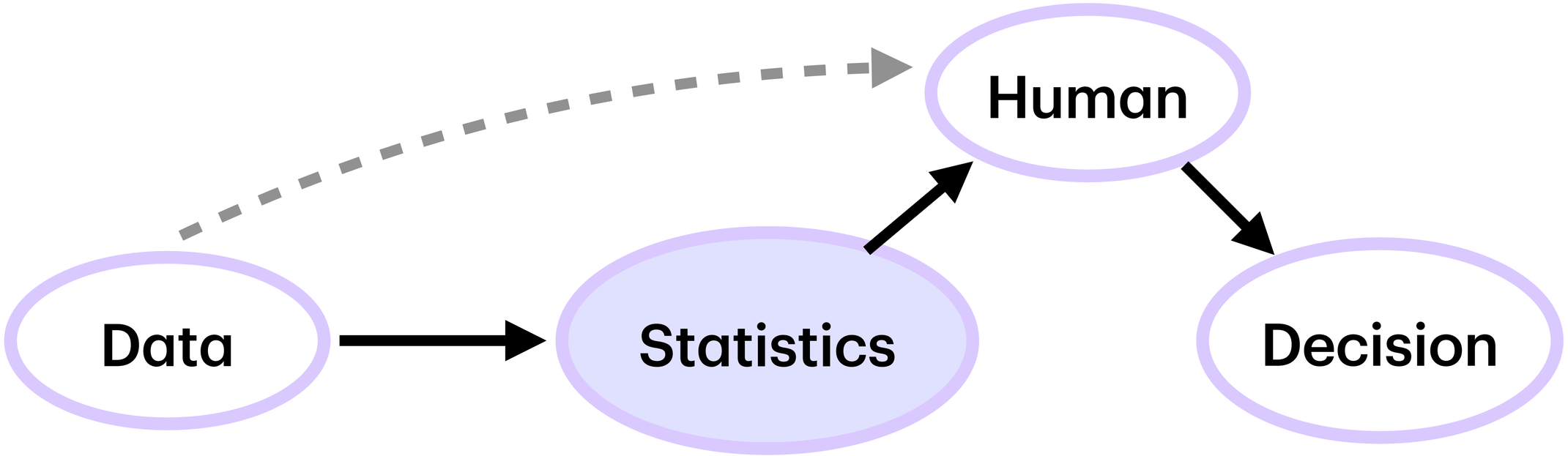
Concluding Thoughts

Statisticians often view the world like this:



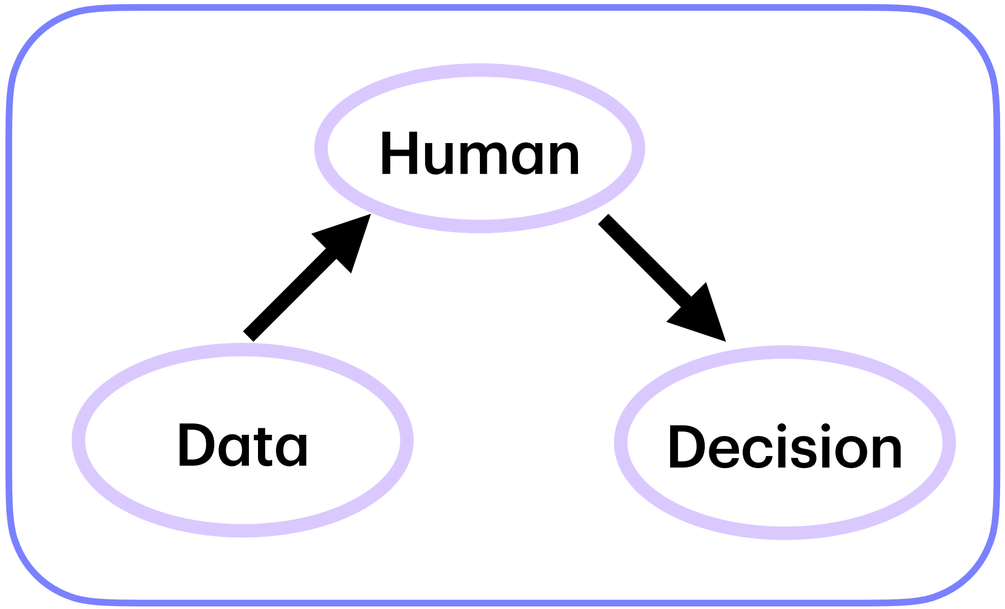
But really the world is like this:

or this:

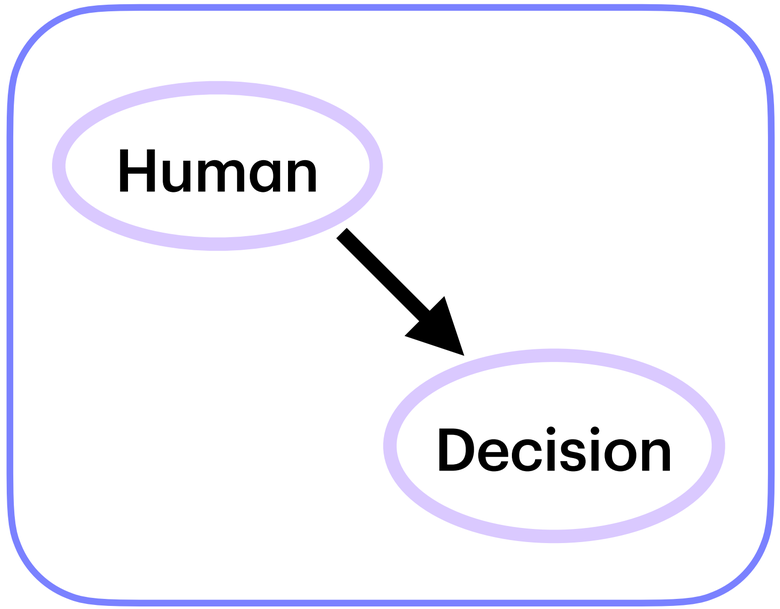


or even,

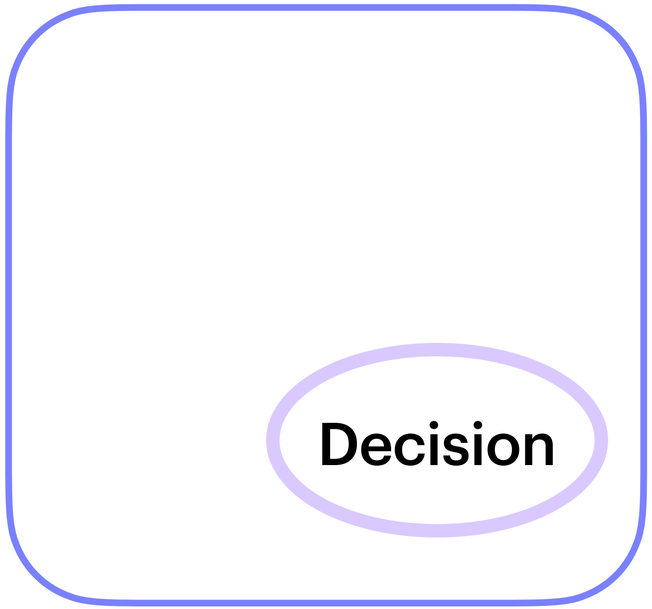
this:



or this:



or this:



As statisticians, we should think more about the full decision-making pipeline

This means:

1. Designing methods with the goal of improving decision-making (either human or automated)

    In conformal, this means not pursuing coverage not for the sake of coverage but *for the sake of useful prediction sets*

2. Demonstrating the usefulness of statistical methods in real world settings and that the pipelines with statistics perform better

Some takeaways

- You don’t have to “believe in statistics” to believe that conformal prediction is useful
- Clustered conformal prediction is one way in which statistical reasoning can be used to motivate a new (and hopefully useful) methodology
- But there is much more to do explore in the area of *prediction sets* + *decision-making* and, more generally, *statistics* + *decision-making*

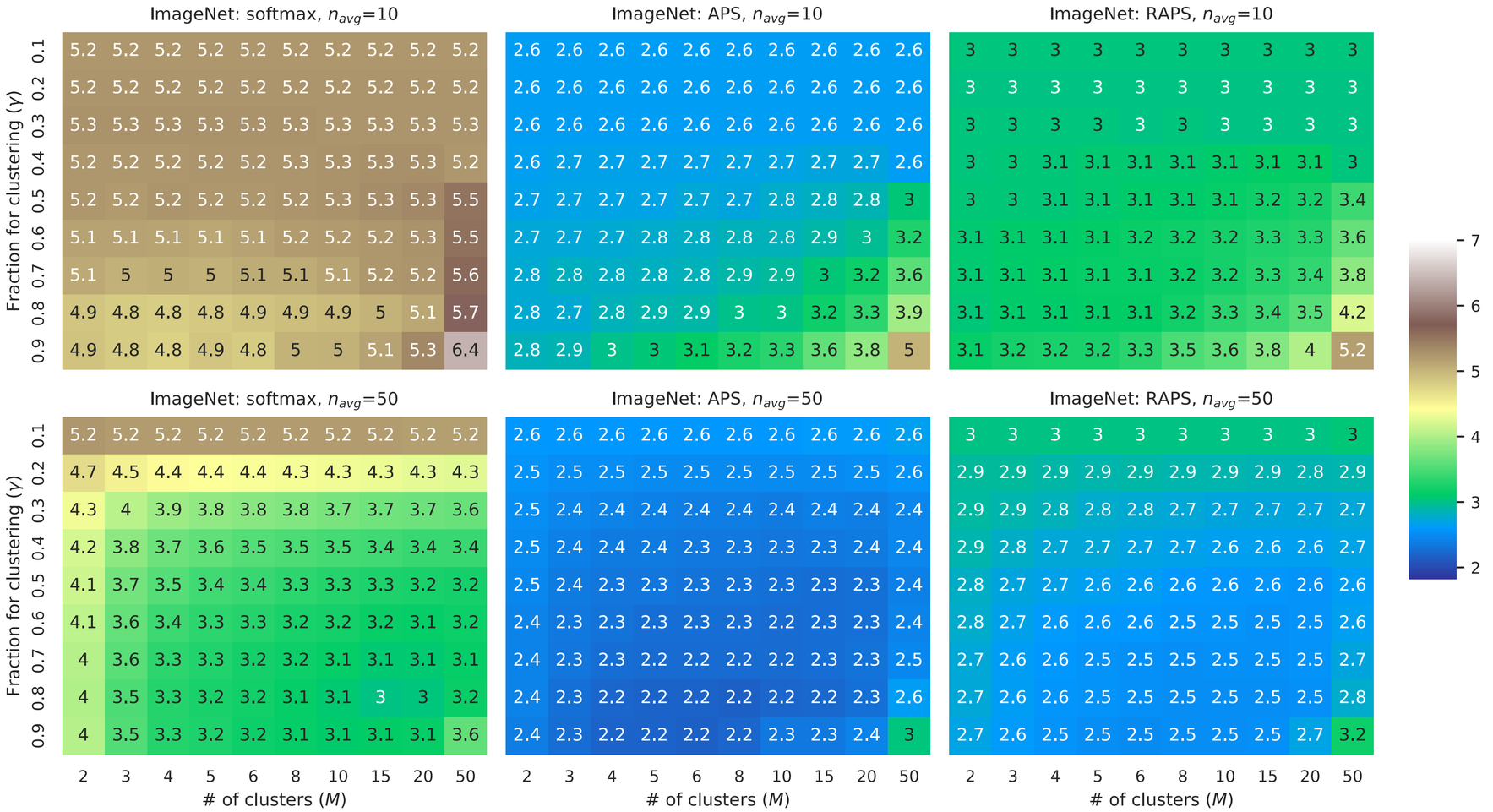
Thank you!

📧 tiffany_ding@berkeley.edu

Extra slides

Sensitivity analysis for Clustered CP parameters

Average class coverage gap on ImageNet



A flurry of recent work in prediction sets + decision making

(Just a few example; not an exhaustive list!)

Theory/methodology:

- *Towards Human-AI Complementarity with Prediction Sets* (De Toni, Okati, Thejaswi, Straitouri, Gomex-Rodriguez, 2024)
- *Designing Decision Support Systems Using Counterfactual Prediction Sets* (Straitouri & Gomez Rodriguez, 2024)
- *Decision-Focused Uncertainty Quantification* (Cortes-Gomez, Patiño, Byun, Wu, Horvitz, Wilder, 2024)
- *Decision Theoretic Foundations for Conformal Prediction: Optimal Uncertainty Quantification for Risk-Averse Agents* (Kiyani, Pappas, Roth, Hassani, 2025)

Empirical studies:

- *Conformal Prediction Sets Improve Human Decision Making* (Cresswell, Kumar, Vouitsis, 2024)
- *Conformal Prediction Sets Can Cause Disparate Impact* (Cresswell, Kumar, Sui, Belbahri, 2024)
- *Evaluating the Utility of Conformal Prediction Sets for AI-Advised Image Labeling* (Zhang, Chatsimparmpas, Kamali, Hullman, 2024)