Class-conditional conformal prediction with many classes

Tiffany Ding

February 6, 2024

Joint work with

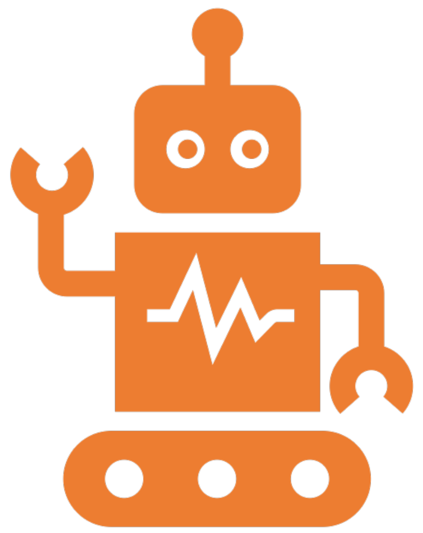
Anastasios Angelopoulos

Stephen Bates

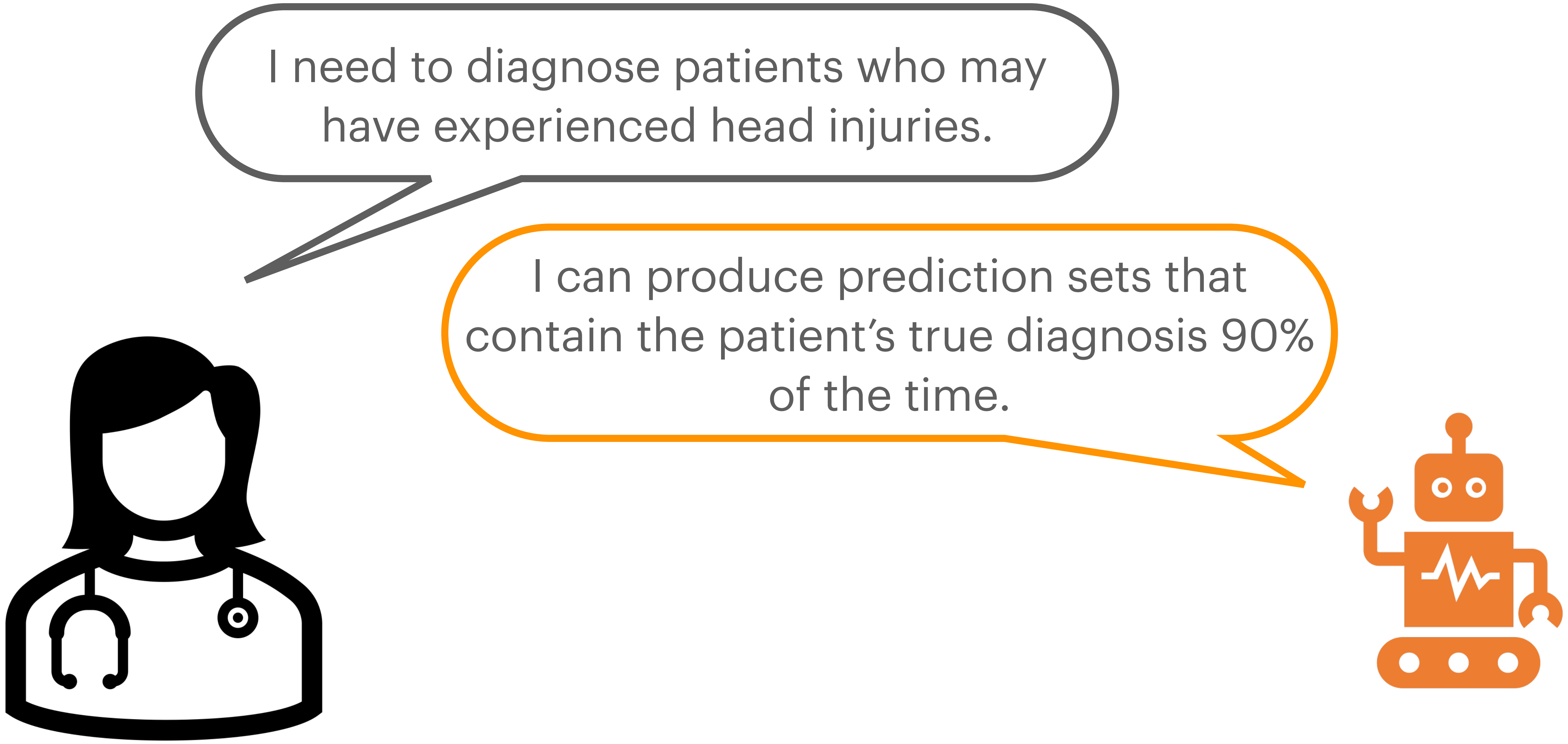
Michael Jordan

Ryan Tibshirani

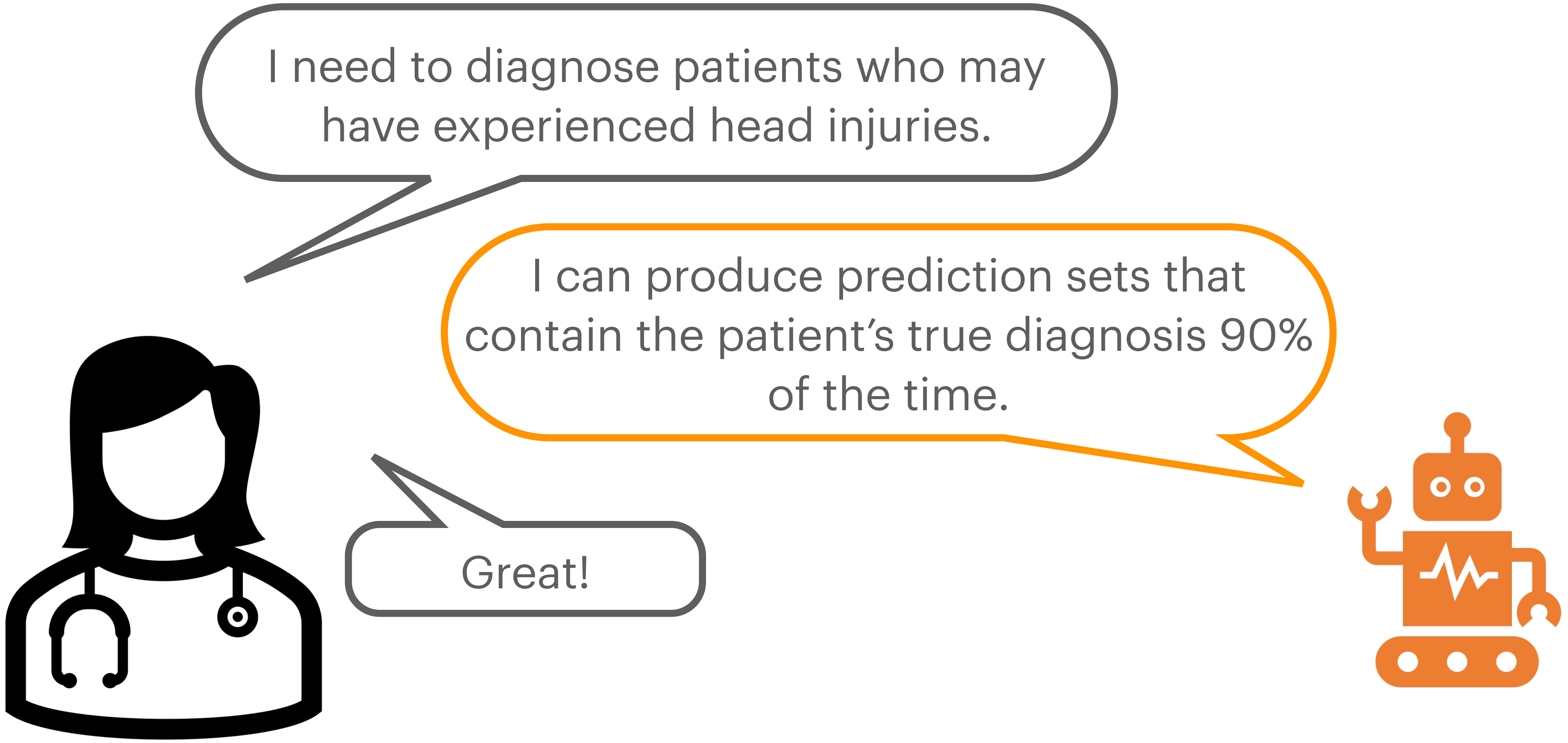
Imagine you're a doctor...

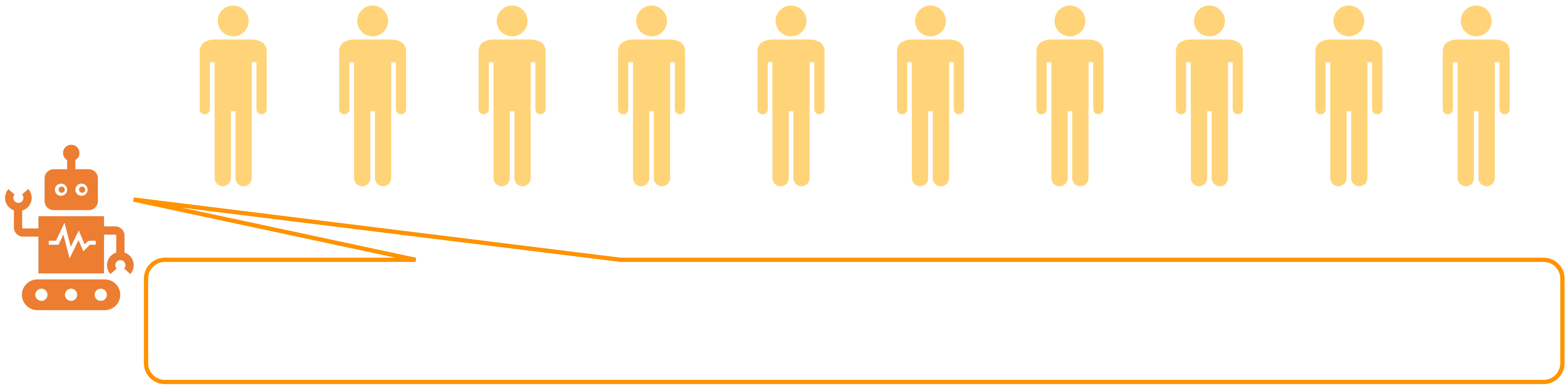


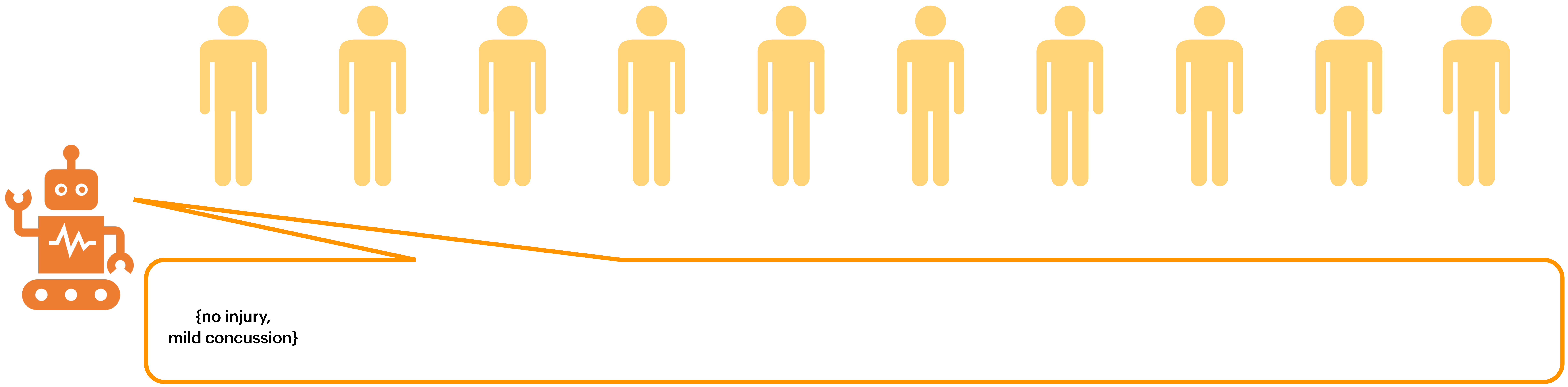
Imagine you're a doctor…

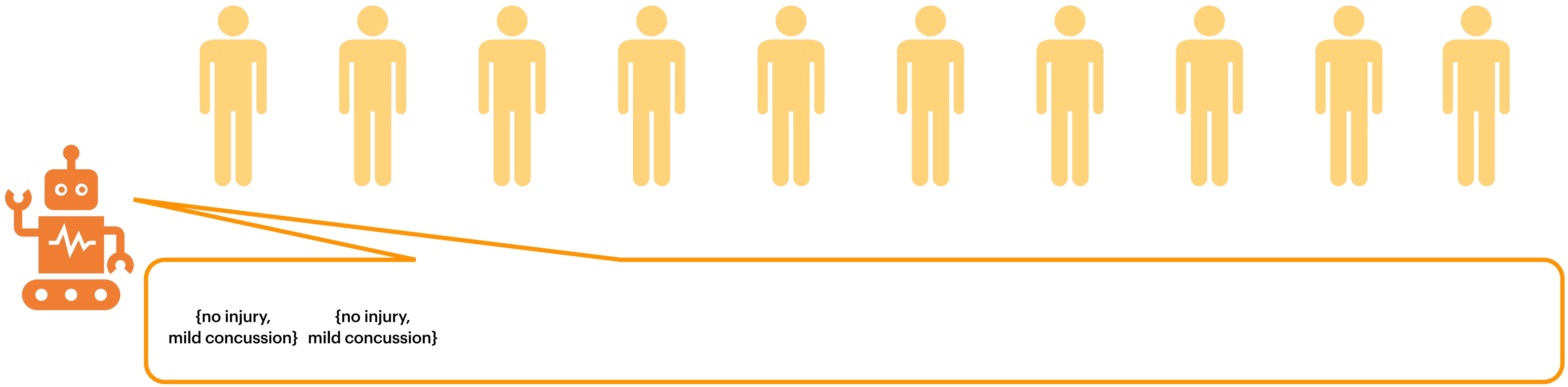


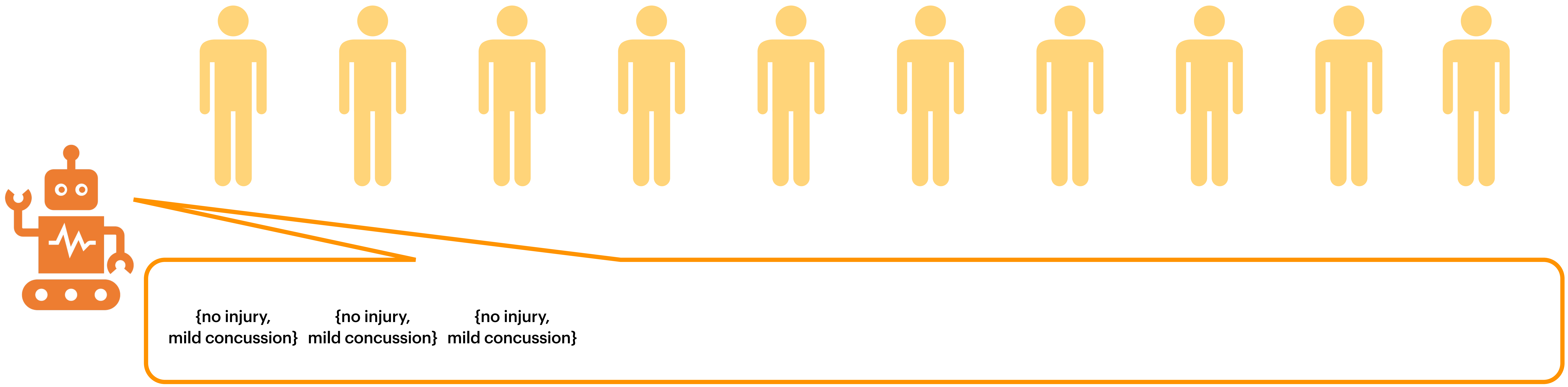
Imagine you're a doctor...

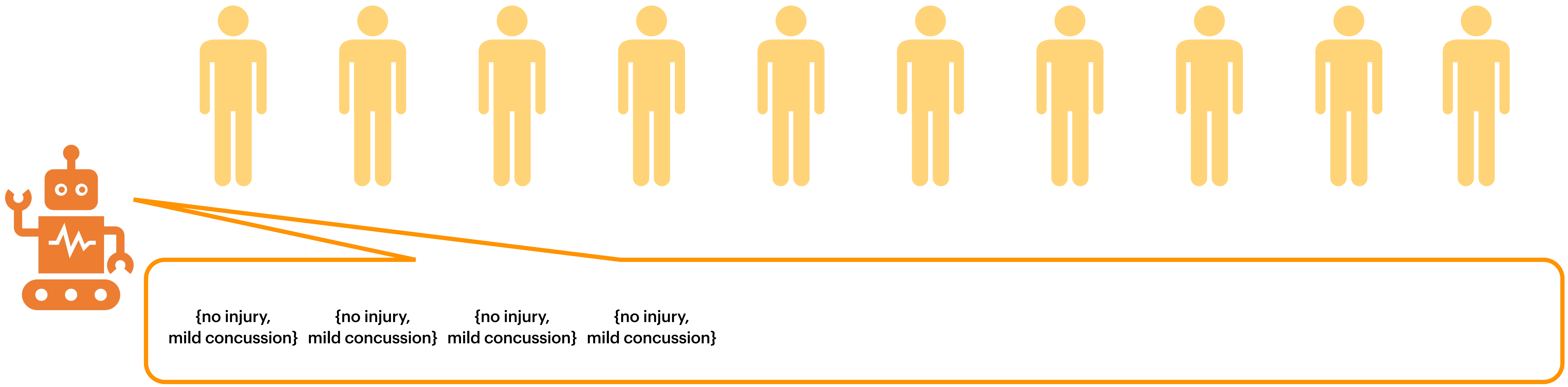


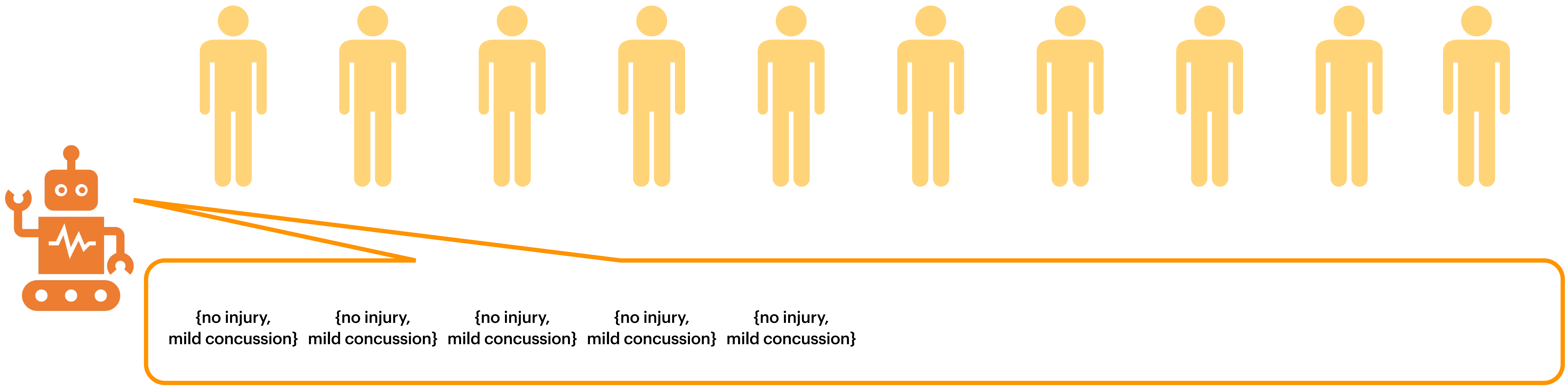


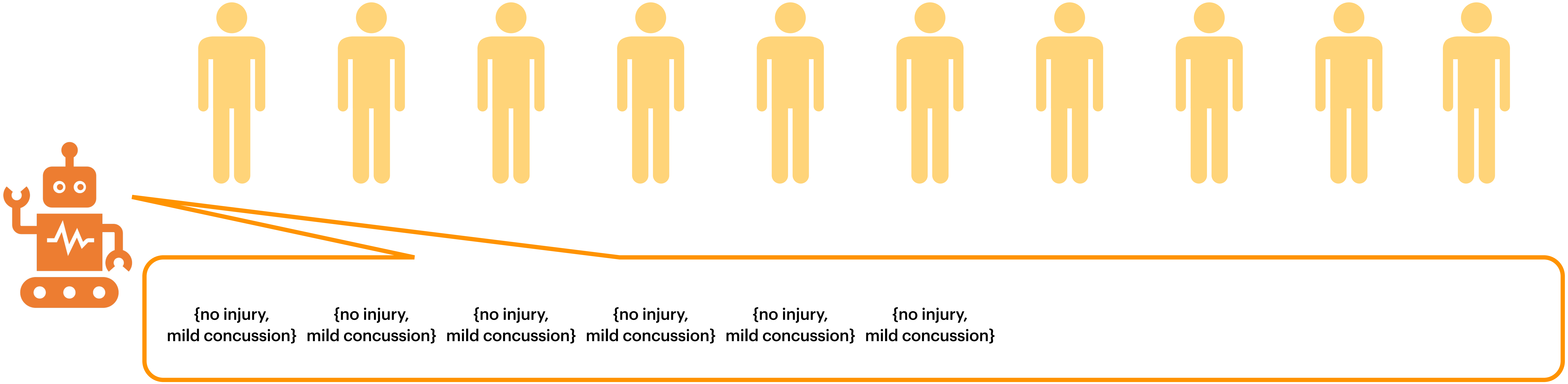

{no injury,
mild concussion}

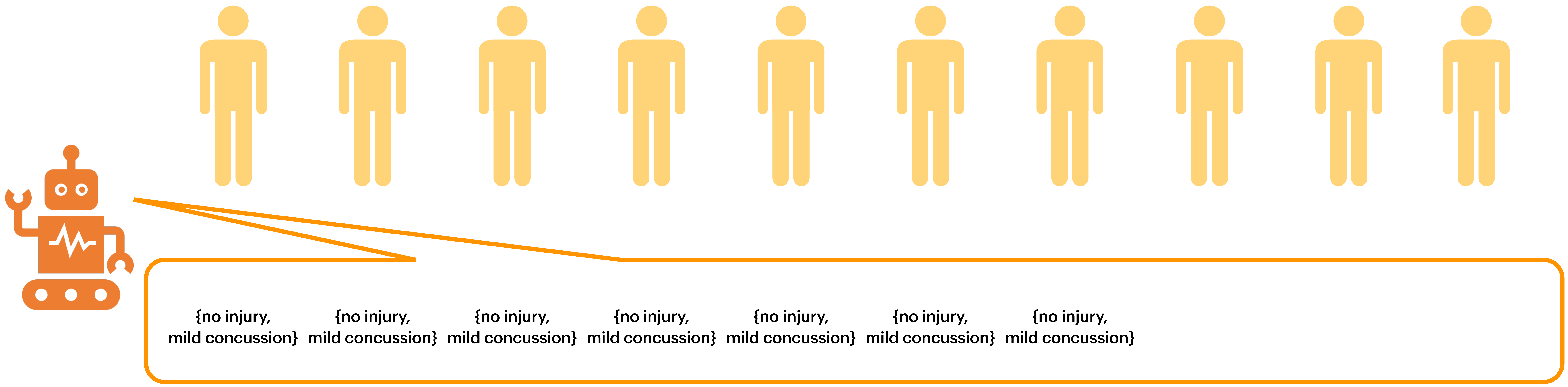

{no injury,
mild concussion}
{no injury,
mild concussion}

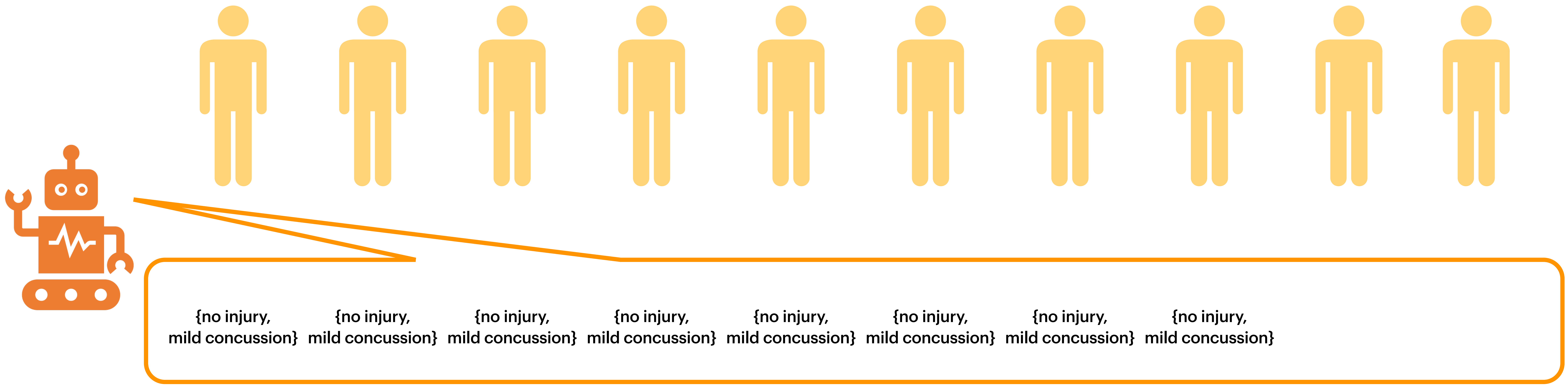

{no injury,
mild concussion}
{no injury,
mild concussion}
{no injury,
mild concussion}

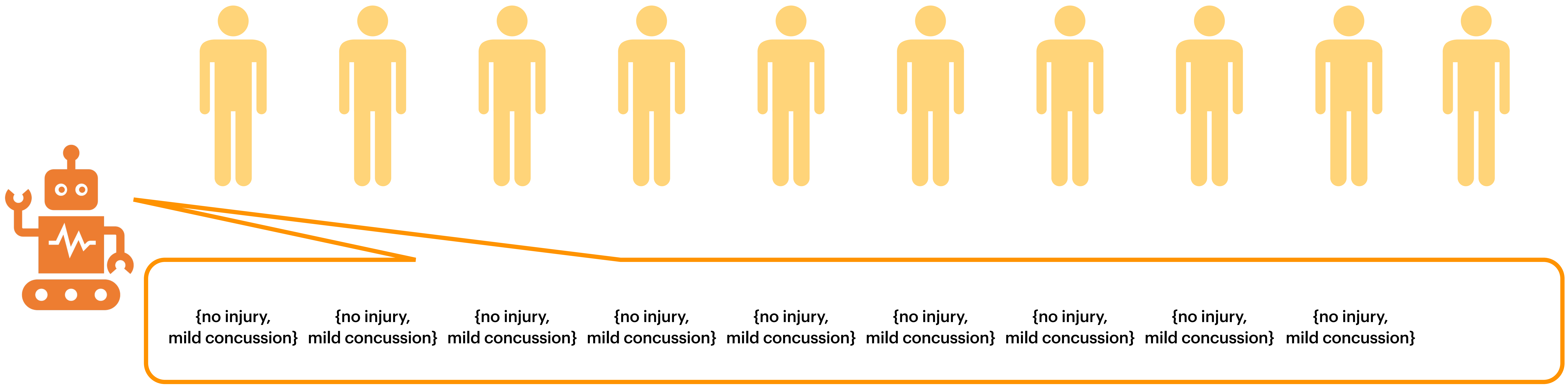

{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}

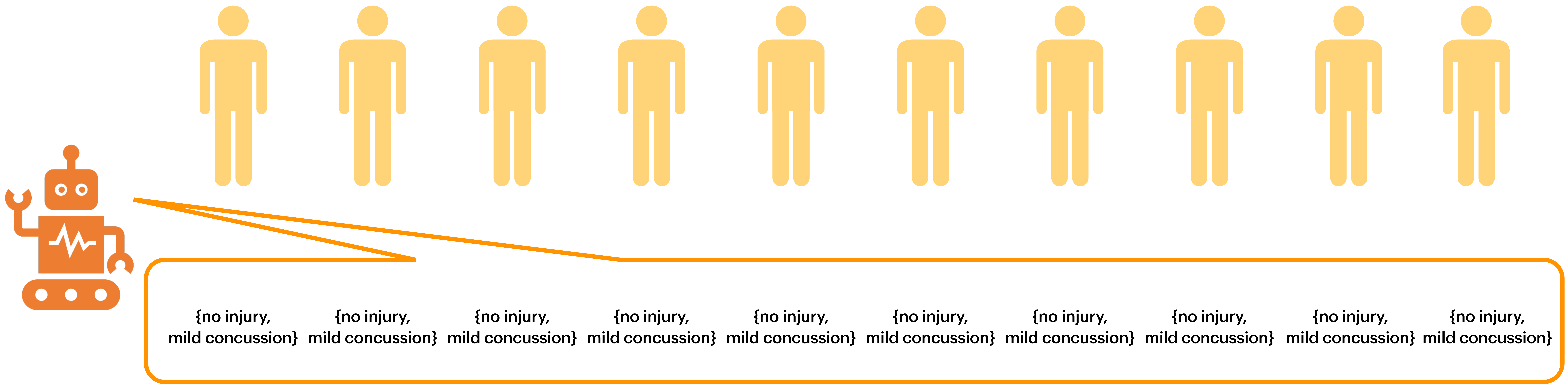

{no injury,
mild concussion}
{no injury,
mild concussion}
{no injury,
mild concussion}
{no injury,
mild concussion}
{no injury,
mild concussion}


{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}

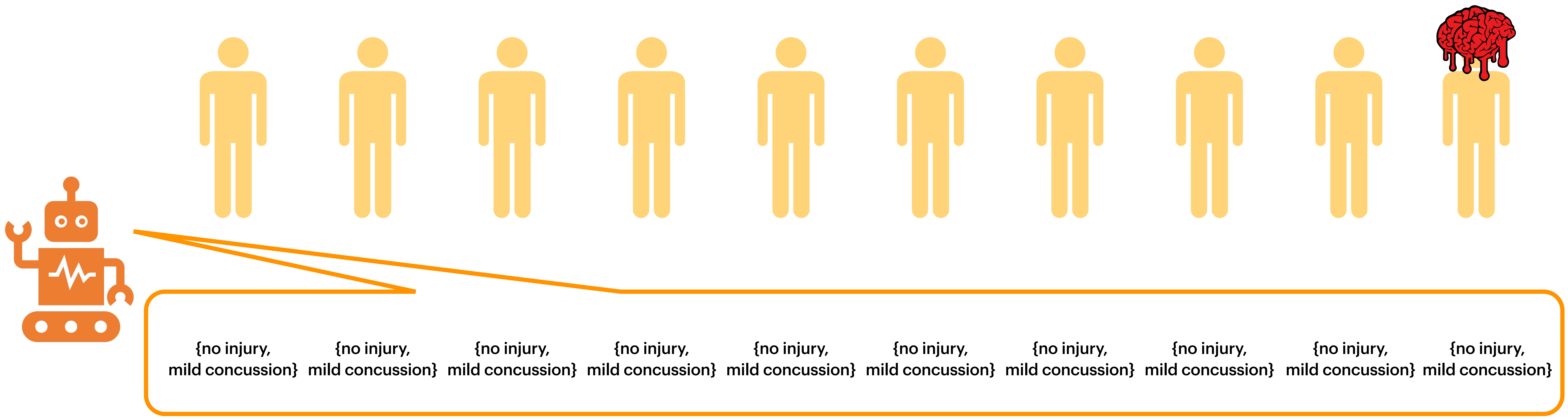

{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}


{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}

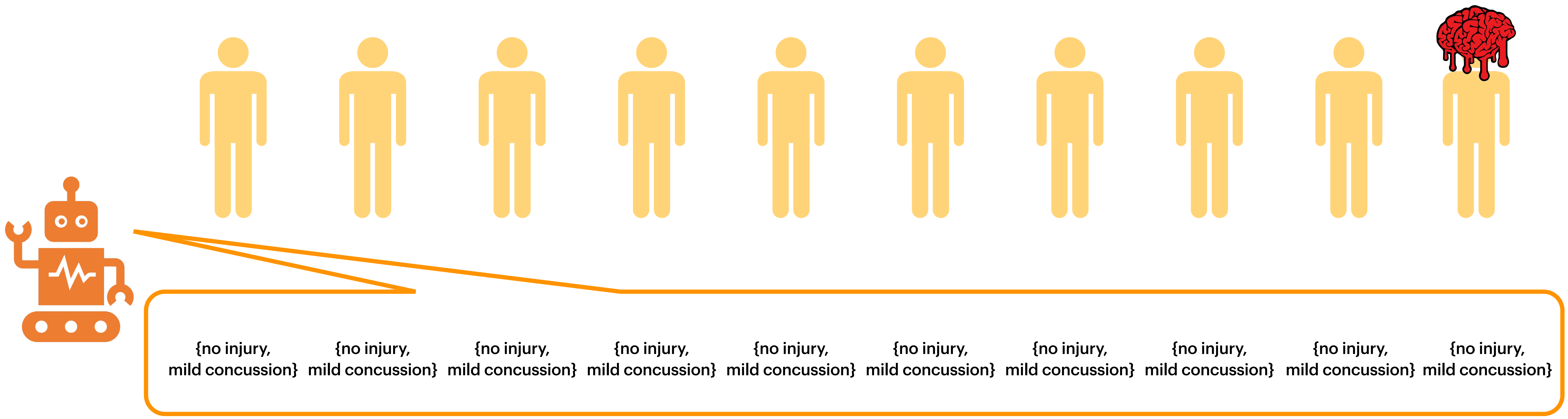

{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}


{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}
{no injury, mild concussion}

This patient actually has an **intracranial hemorrhage**

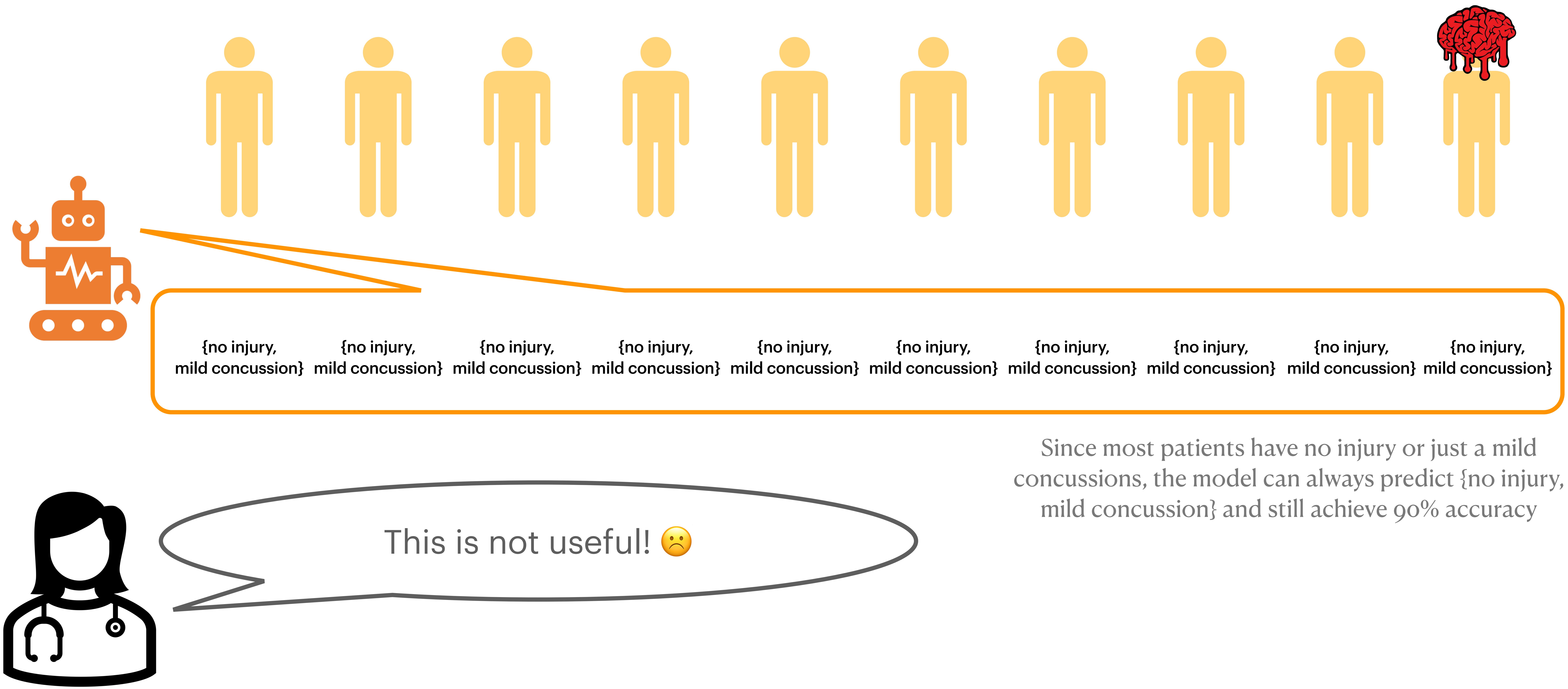


This patient actually has an **intracranial hemorrhage**



Since most patients have no injury or just a mild concussions, the model can always predict {no injury, mild concussion} and still achieve 90% accuracy

This patient actually has an **intracranial hemorrhage**



Since most patients have no injury or just a mild concussions, the model can always predict {no injury, mild concussion} and still achieve 90% accuracy

What do we actually want in this setting?

It's not enough for prediction sets to have coverage *on average across all patients.*

We want to have coverage *conditioned on the true label.*

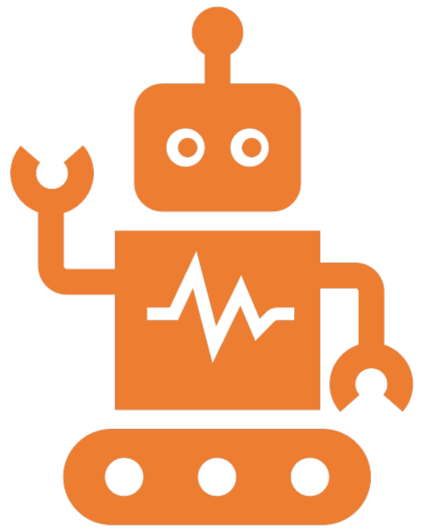
- e.g., with high probability,

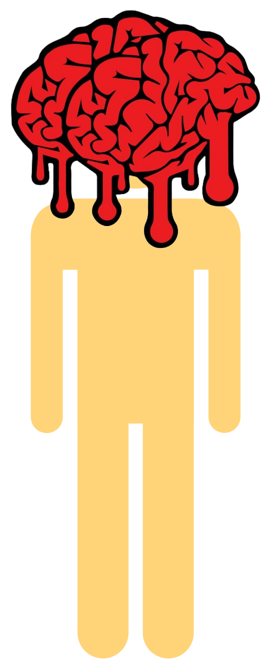
What do we actually want in this setting?

It's not enough for prediction sets to have coverage *on average across all patients.*

We want to have coverage *conditioned on the true label.*

- e.g., with high probability,



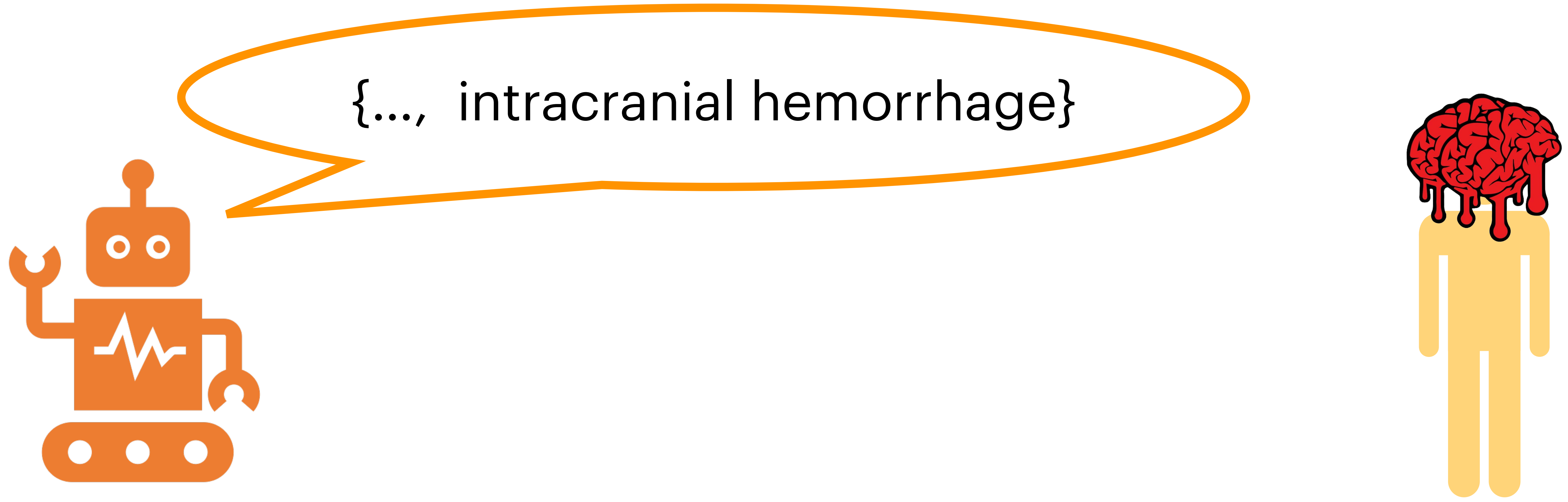


What do we actually want in this setting?

It's not enough for prediction sets to have coverage *on average across all patients.*

We want to have coverage *conditioned on the true label.*

- e.g., with high probability,



What do we actually want in this setting?

(in math)

What do we actually want in this setting?

(in math)

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

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

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

Q: Can we use *conformal prediction* to solve this problem?

Q: Can we use *conformal prediction* to solve this problem?

A: Yes, but naive methods struggle when there are many classes and/or limited labeled data.

In these situations, we must be a bit cleverer.

Review of conformal prediction (CP)

Standard CP

Review of conformal prediction (CP)

Standard CP

Black-box model

f

0) Use model f to define a *conformal score function* $s(x, y)$

e.g., if f outputs a vector of softmax scores, can use $s(x, y) = 1 - f_y(x)$

Review of conformal prediction (CP)

Standard CP

Black-box model

f

(X_1, Y_1)

(X_2, Y_2)

$\vdots$

(X_n, Y_n)

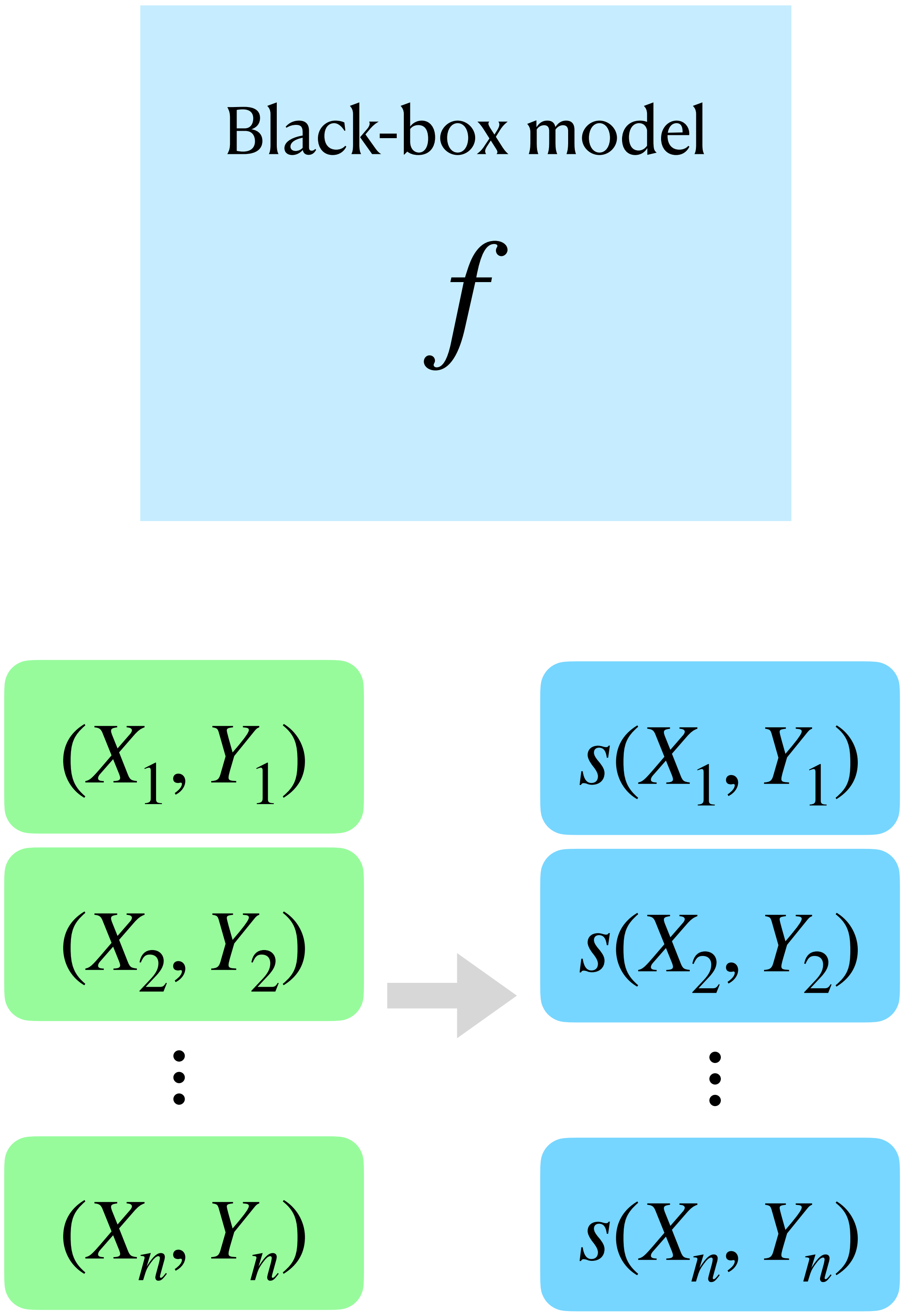
0) Use model f to define a *conformal score function* $s(x, y)$

e.g., if f outputs a vector of softmax scores, can use $s(x, y) = 1 - f_y(x)$

1) Apply $s(x, y)$ to n labeled calibration data points to get conformal scores

Review of conformal prediction (CP)

Standard CP



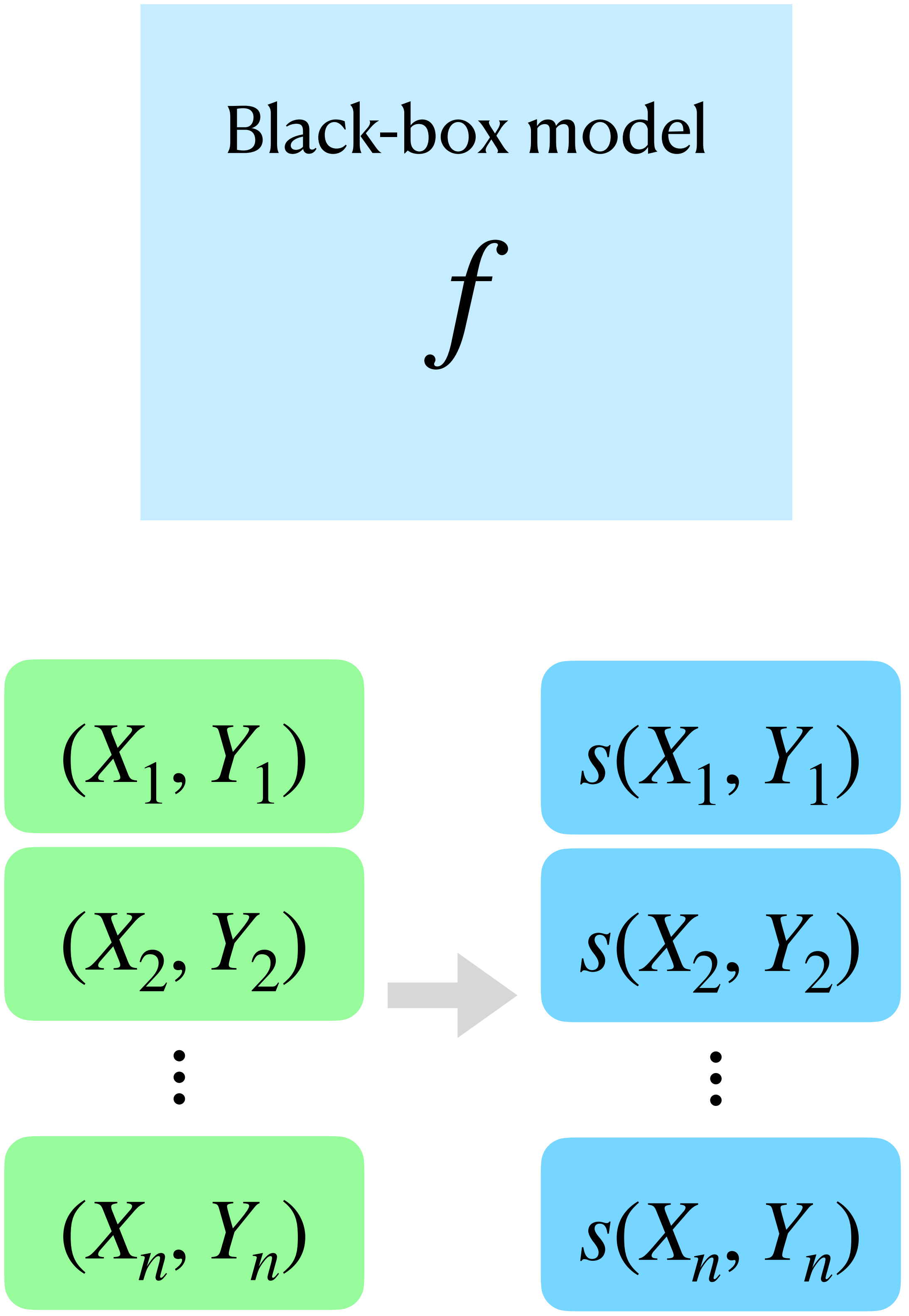
0) Use model f to define a *conformal score function* $s(x, y)$

e.g., if f outputs a vector of softmax scores, can use $s(x, y) = 1 - f_y(x)$

1) Apply $s(x, y)$ to n labeled calibration data points to get conformal scores

Review of conformal prediction (CP)

Standard CP



0) Use model f to define a *conformal score function* $s(x, y)$

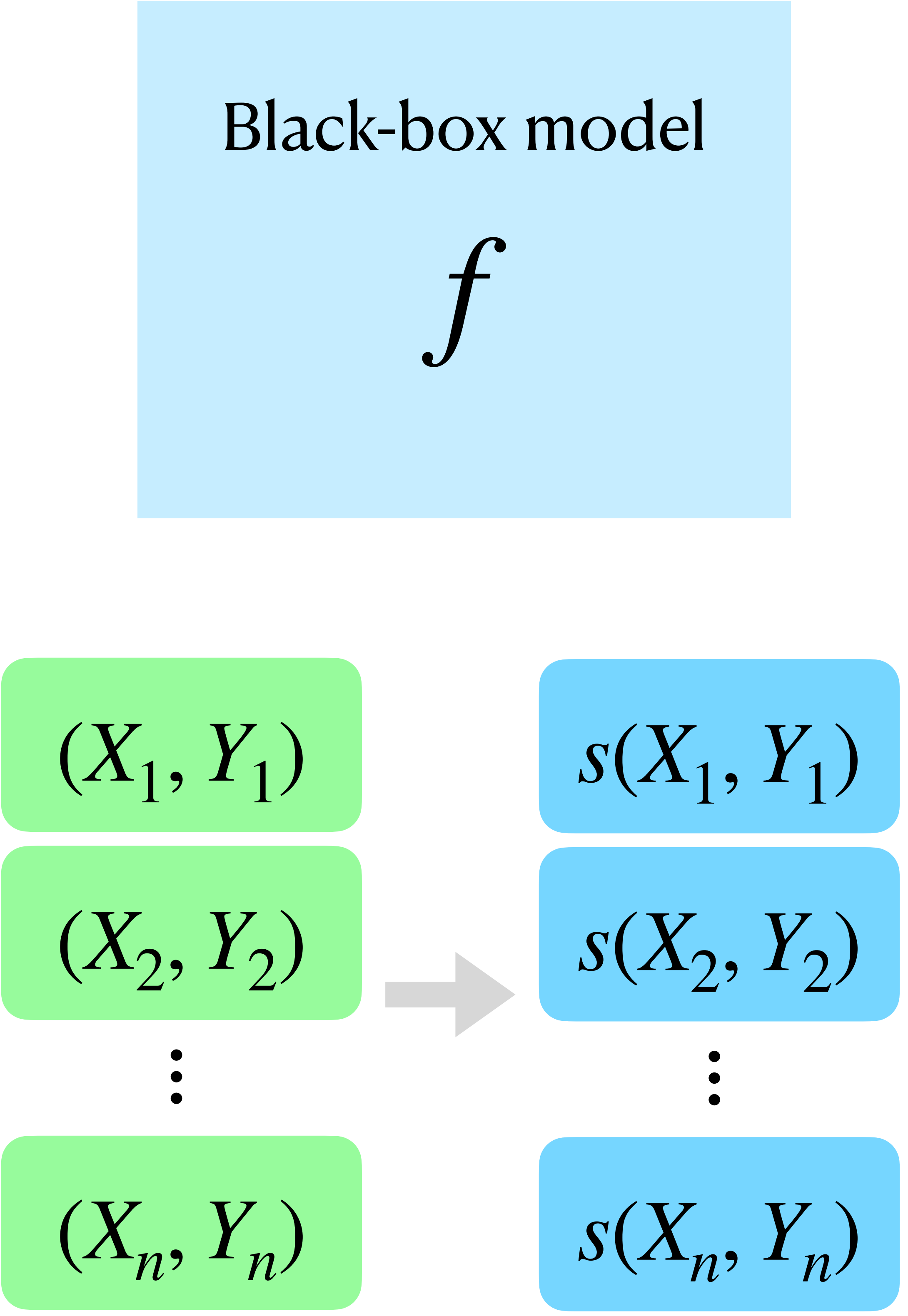
e.g., if f outputs a vector of softmax scores, can use $s(x, y) = 1 - f_y(x)$

1) Apply $s(x, y)$ to n labeled calibration data points to get conformal scores

2) Let $\hat{q} = \lceil (1 - \alpha)(n + 1) \rceil$ largest score

Review of conformal prediction (CP)

Standard CP



0) Use model f to define a *conformal score function* $s(x, y)$

e.g., if f outputs a vector of softmax scores, can use $s(x, y) = 1 - f_y(x)$

1) Apply $s(x, y)$ to n labeled calibration data points to get conformal scores

2) Let $\hat{q} = \lceil (1 - \alpha)(n + 1) \rceil$ largest score

At test time, construct prediction set as

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

Fact: As long as the calibration points and the test point are exchangeable, standard CP achieves *marginal coverage*:

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

Proof:

Fact: As long as the calibration points and the test point are exchangeable, standard CP achieves *marginal coverage*:

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

Proof:

$(X_1, Y_1), \ldots, (X_n, Y_n)$ and $(X_{\text{test}}, Y_{\text{test}})$ are exchangeable

Fact: As long as the calibration points and the test point are exchangeable, standard CP achieves *marginal coverage*:

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

Proof:

$(X_1, Y_1), \ldots, (X_n, Y_n)$ and $(X_{\text{test}}, Y_{\text{test}})$ are exchangeable

$$\implies s(X_1, Y_1), \ldots, s(X_n, Y_n) \text{ and } s(X_{\text{test}}, Y_{\text{test}}) \text{ are exchangeable}$$

Fact: As long as the calibration points and the test point are exchangeable, standard CP achieves *marginal coverage*:

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

Proof:

$(X_1, Y_1), \ldots, (X_n, Y_n)$ and $(X_{\text{test}}, Y_{\text{test}})$ are exchangeable

$\implies$ $s(X_1, Y_1), \ldots, s(X_n, Y_n)$ and $s(X_{\text{test}}, Y_{\text{test}})$ are exchangeable

$\implies$ any ordering of $s(X_1, Y_1), \ldots, s(X_n, Y_n)$ and $s(X_{\text{test}}, Y_{\text{test}})$ is equally likely

Fact: As long as the calibration points and the test point are exchangeable, standard CP achieves *marginal coverage*:

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

Proof:

$(X_1, Y_1), \ldots, (X_n, Y_n)$ and $(X_{\text{test}}, Y_{\text{test}})$ are exchangeable

$\implies$ $s(X_1, Y_1), \ldots, s(X_n, Y_n)$ and $s(X_{\text{test}}, Y_{\text{test}})$ are exchangeable

$\implies$ any ordering of $s(X_1, Y_1), \ldots, s(X_n, Y_n)$ and $s(X_{\text{test}}, Y_{\text{test}})$ is equally likely

$\implies$ $\mathbb{P}(s(X_{\text{test}}, Y_{\text{test}})$ is one of the $\lceil (1-\alpha)(n+1) \rceil$ smallest scores $) = \dfrac{\lceil (1-\alpha)(n+1) \rceil}{n+1} \geq 1 - \alpha$

Fact: As long as the calibration points and the test point are exchangeable, standard CP achieves *marginal coverage*:

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

Proof:

$(X_1, Y_1), \ldots, (X_n, Y_n)$ and $(X_{\text{test}}, Y_{\text{test}})$ are exchangeable

$\implies$ $s(X_1, Y_1), \ldots, s(X_n, Y_n)$ and $s(X_{\text{test}}, Y_{\text{test}})$ are exchangeable

$\implies$ any ordering of $s(X_1, Y_1), \ldots, s(X_n, Y_n)$ and $s(X_{\text{test}}, Y_{\text{test}})$ is equally likely

$\implies$ $\underbrace{\mathbb{P}(s(X_{\text{test}}, Y_{\text{test}}) \text{ is one of the } \lceil (1-\alpha)(n+1) \rceil \text{ smallest scores })}$ $= \dfrac{\lceil (1-\alpha)(n+1) \rceil}{n+1} \geq 1 - \alpha$

$= \mathbb{P}(s(X_{\text{test}}, Y_{\text{test}}) \leq \hat{q}) = \mathbb{P}(Y_{\text{test}} \in C(X_{\text{test}}))$

Fact: As long as the calibration points and the test point are exchangeable, standard CP achieves *marginal coverage*:

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

Proof:

$(X_1, Y_1), \ldots, (X_n, Y_n)$ and $(X_{\text{test}}, Y_{\text{test}})$ are exchangeable

$\implies$ $s(X_1, Y_1), \ldots, s(X_n, Y_n)$ and $s(X_{\text{test}}, Y_{\text{test}})$ are exchangeable

Note: all we need for valid coverage is **exchangeable conformal scores**

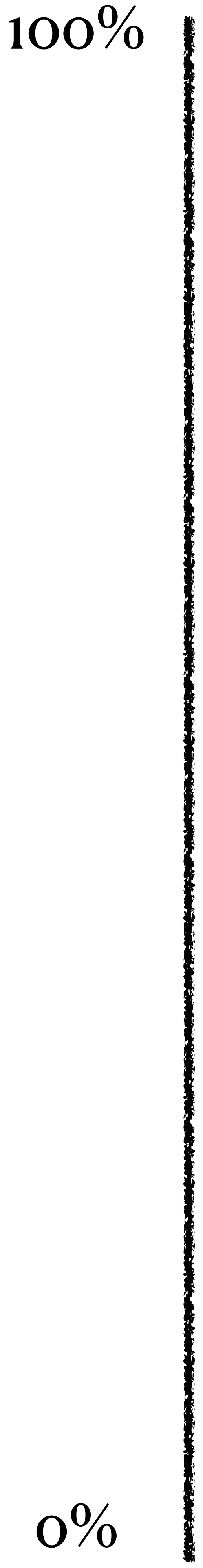
$\implies$ any ordering of $s(X_1, Y_1), \ldots, s(X_n, Y_n)$ and $s(X_{\text{test}}, Y_{\text{test}})$ is equally likely

$\implies$ $\mathbb{P}(s(X_{\text{test}}, Y_{\text{test}})$ is one of the $\lceil (1-\alpha)(n+1) \rceil$ smallest scores $) = \dfrac{\lceil (1-\alpha)(n+1) \rceil}{n+1} \geq 1 - \alpha$

$= \mathbb{P}(s(X_{\text{test}}, Y_{\text{test}}) \leq \hat{q}) = \mathbb{P}(Y_{\text{test}} \in C(X_{\text{test}}))$

Marginal coverage ↛ class-conditional coverage

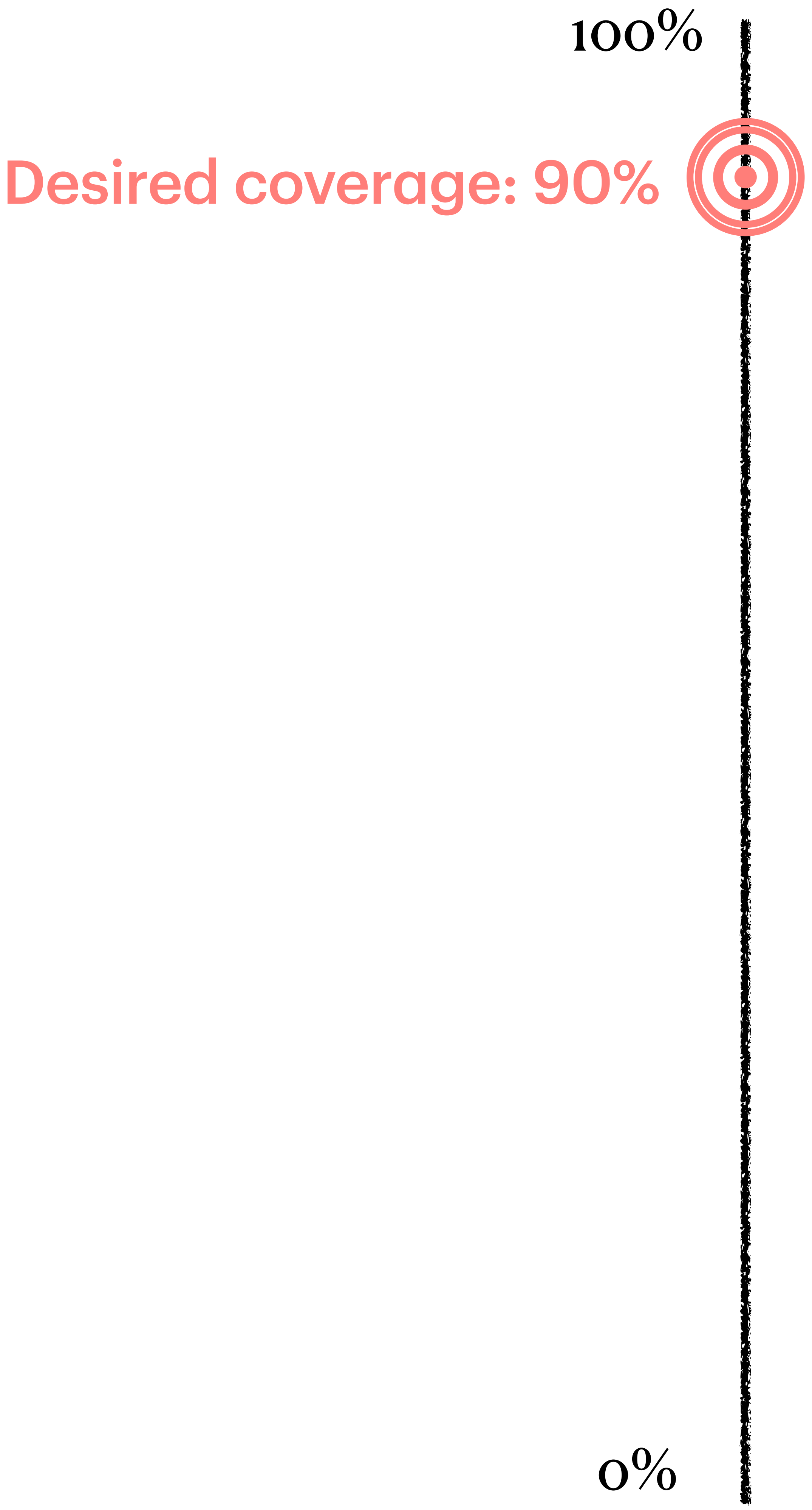
An ImageNet case study



Standard CP on ImageNet using 50,000 calibration images

Marginal coverage ↛ class-conditional coverage

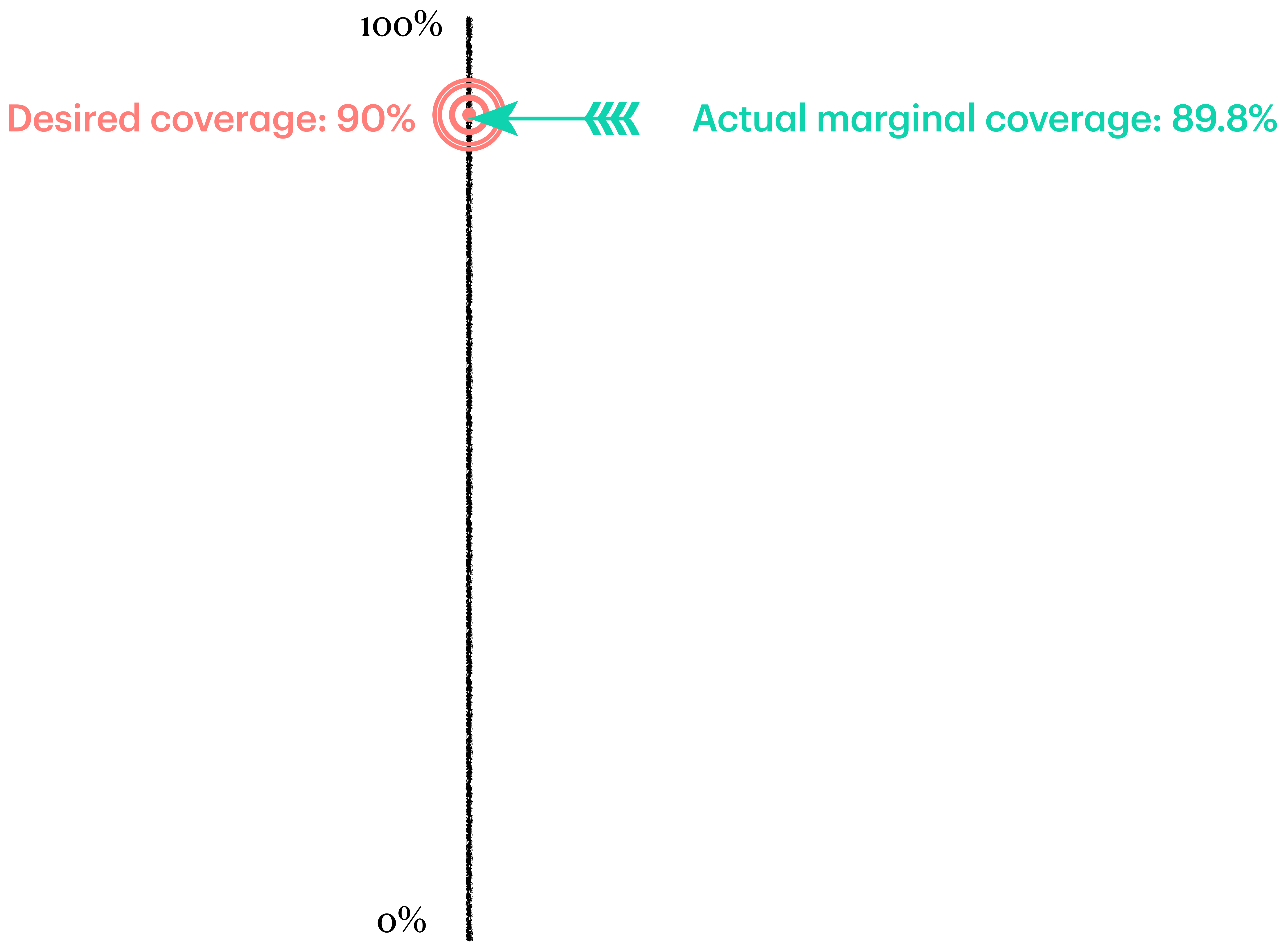
An ImageNet case study



Standard CP on ImageNet using 50,000 calibration images

Marginal coverage $\nrightarrow$ class-conditional coverage

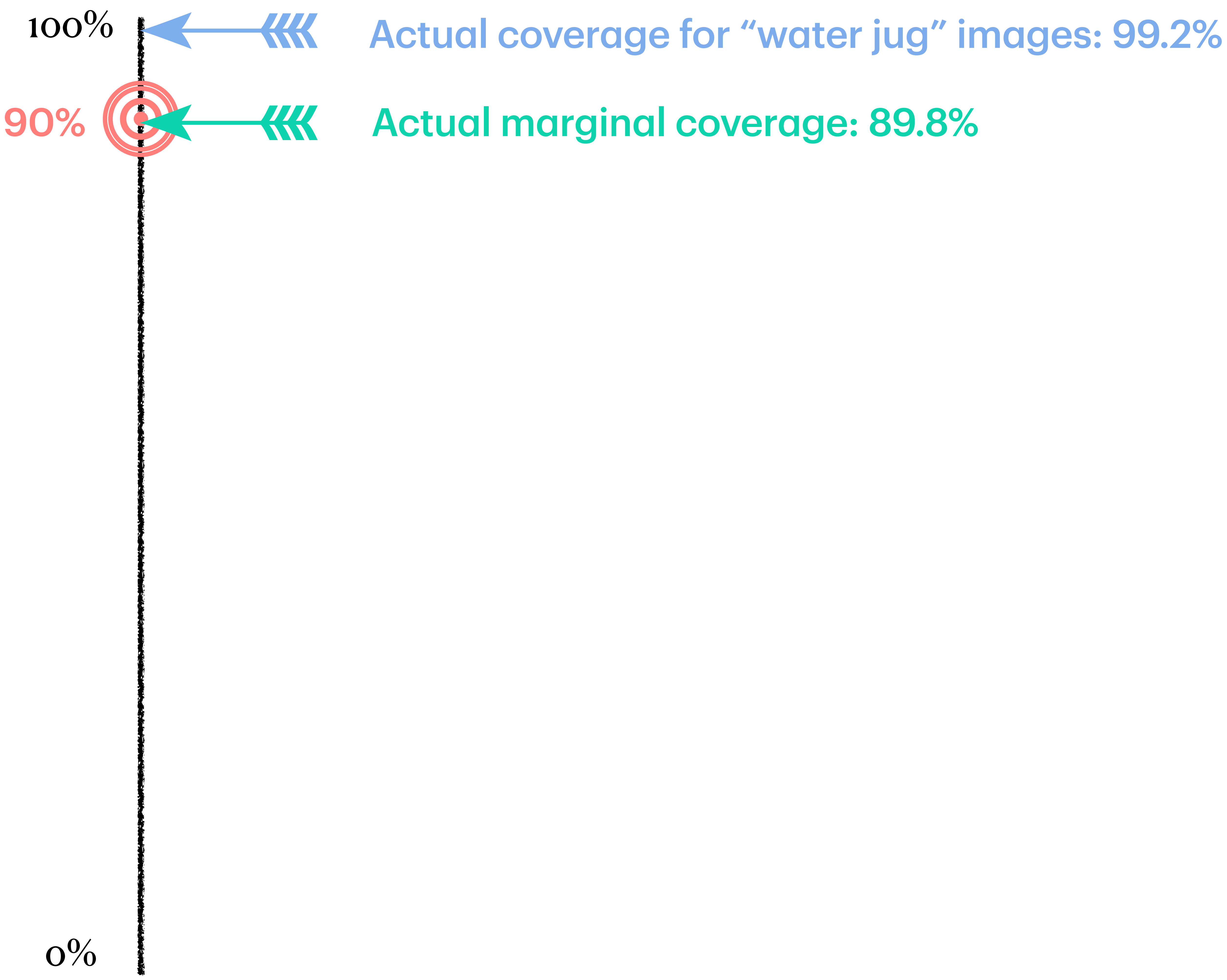
An ImageNet case study



Standard CP on ImageNet using 50,000 calibration images

Marginal coverage ↛ class-conditional coverage

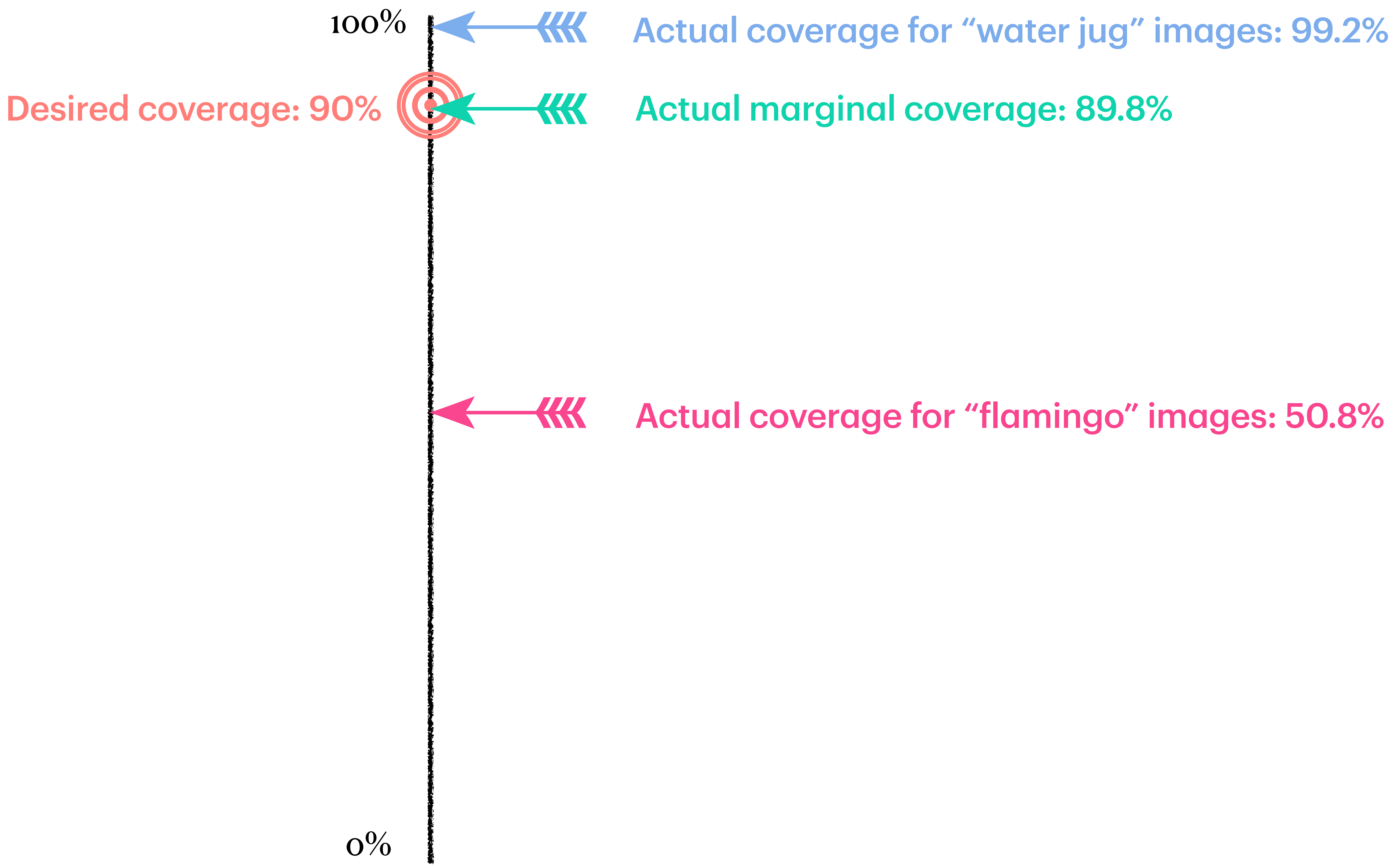
An ImageNet case study



Standard CP on ImageNet using 50,000 calibration images

Marginal coverage $\nrightarrow$ class-conditional coverage

An ImageNet case study



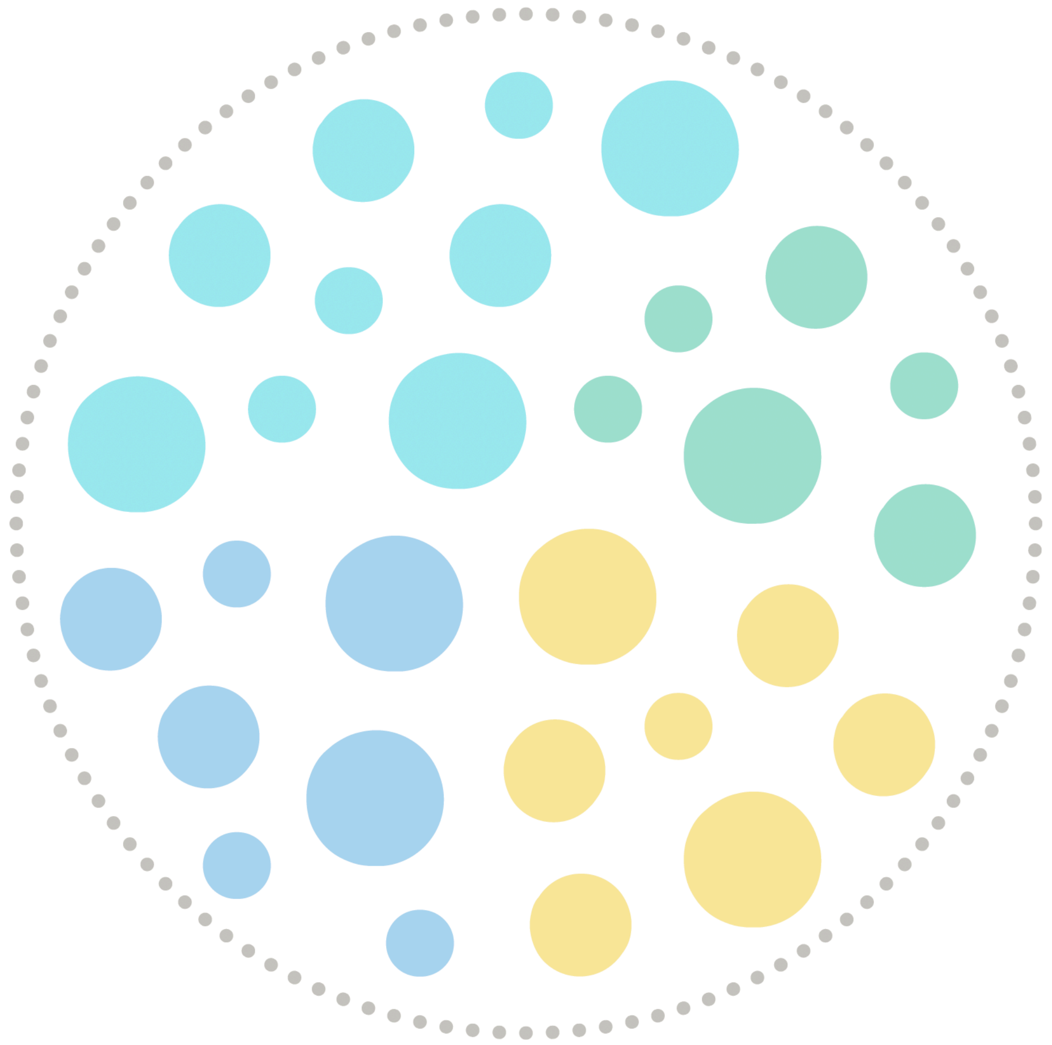
Standard CP on ImageNet using 50,000 calibration images

Classwise CP

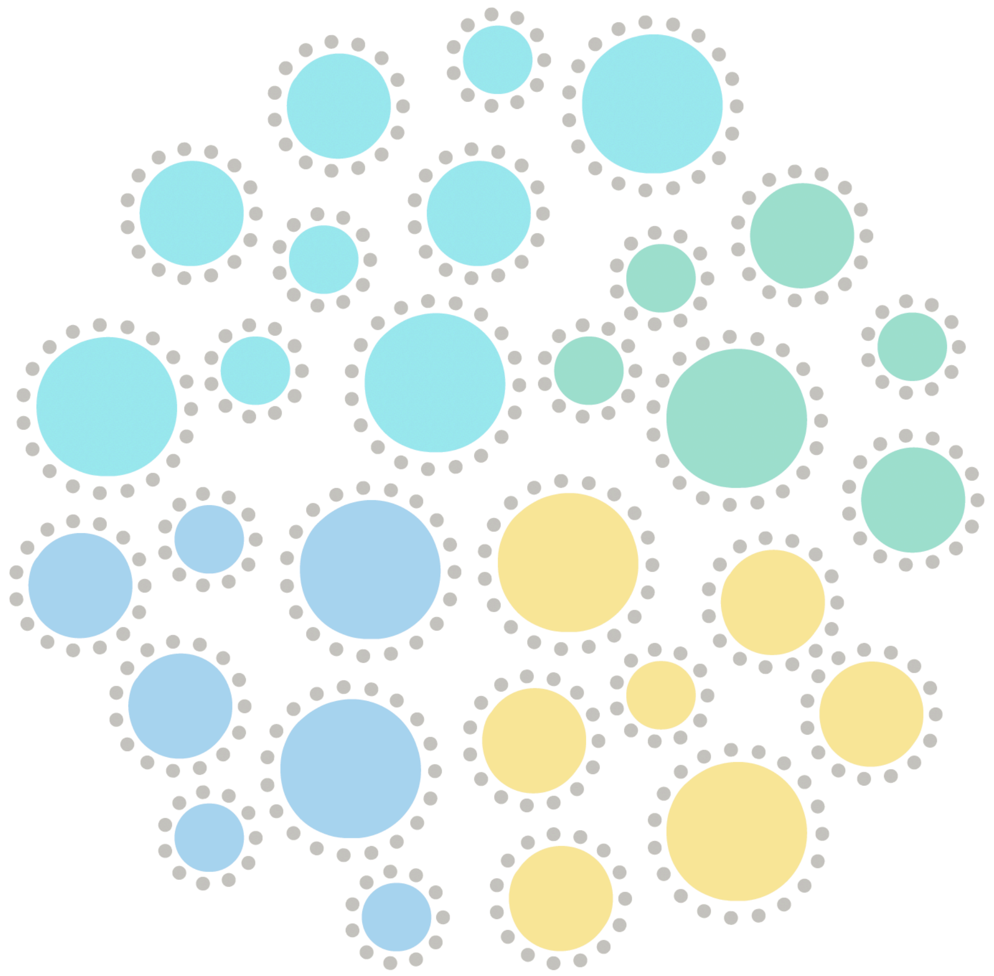
A naive adaptation of CP that achieves class-conditional coverage

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

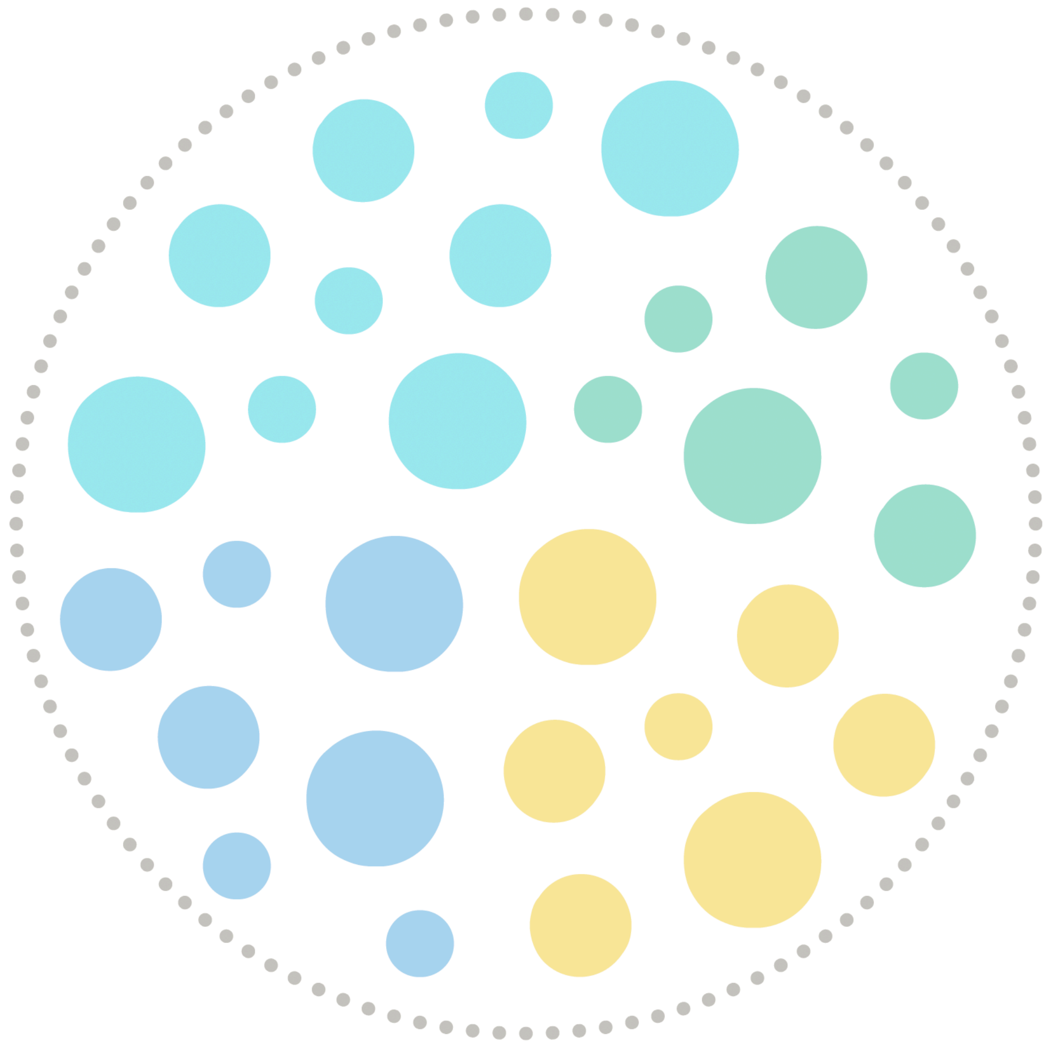
$C_{\text{CLASSWISE}}(X_{\text{test}})$ will have class-conditional coverage, but requires a lot of data per class.



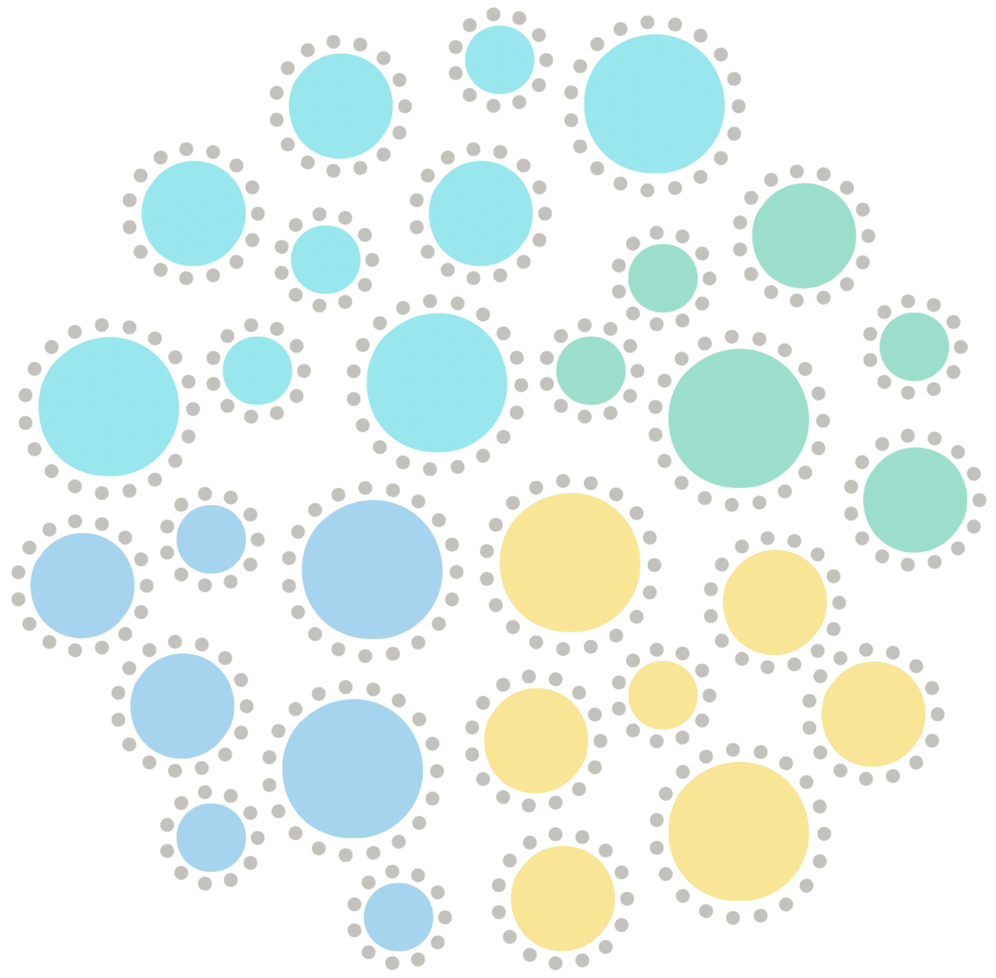
Standard CP



Classwise CP



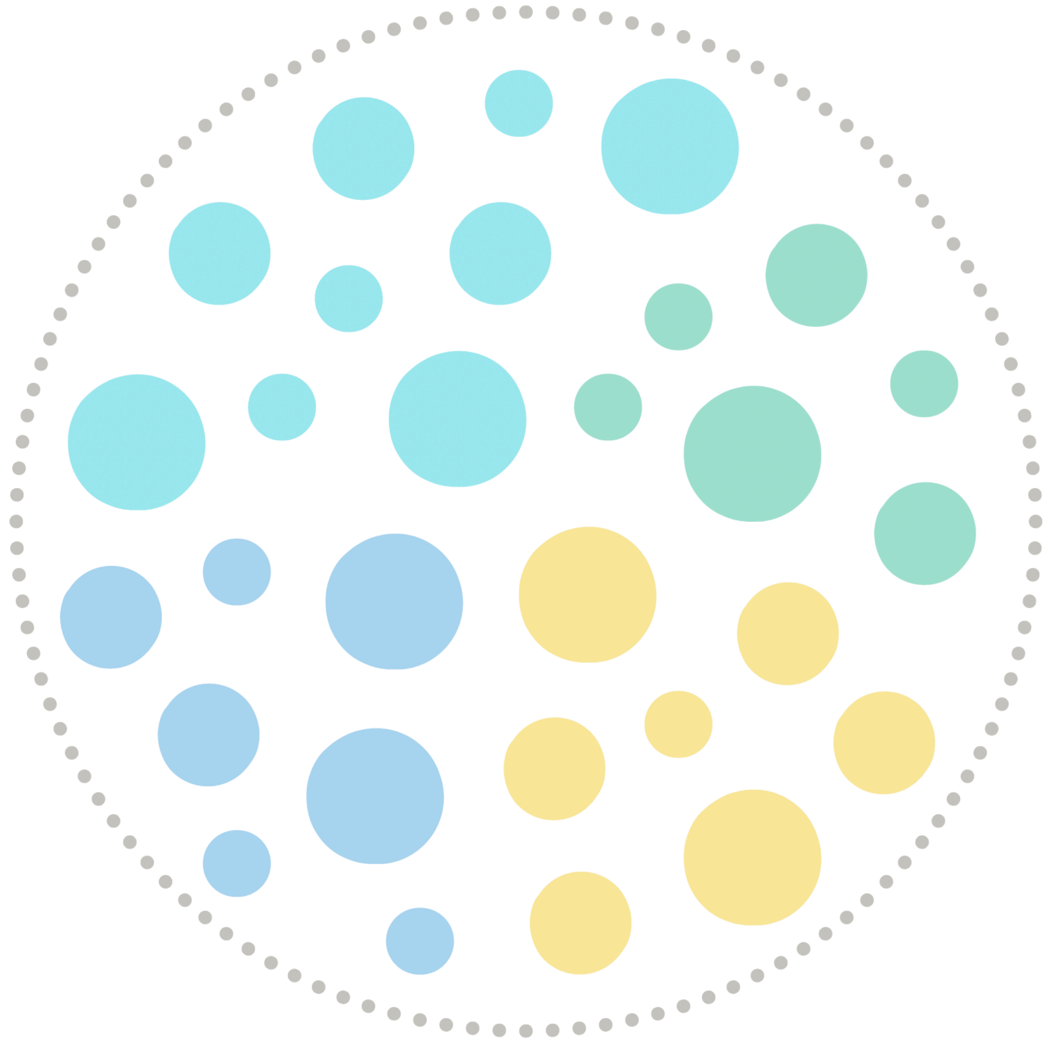
Standard CP



Classwise CP

✅ Low variance

☹️ No class-conditional coverage guarantee



Standard CP

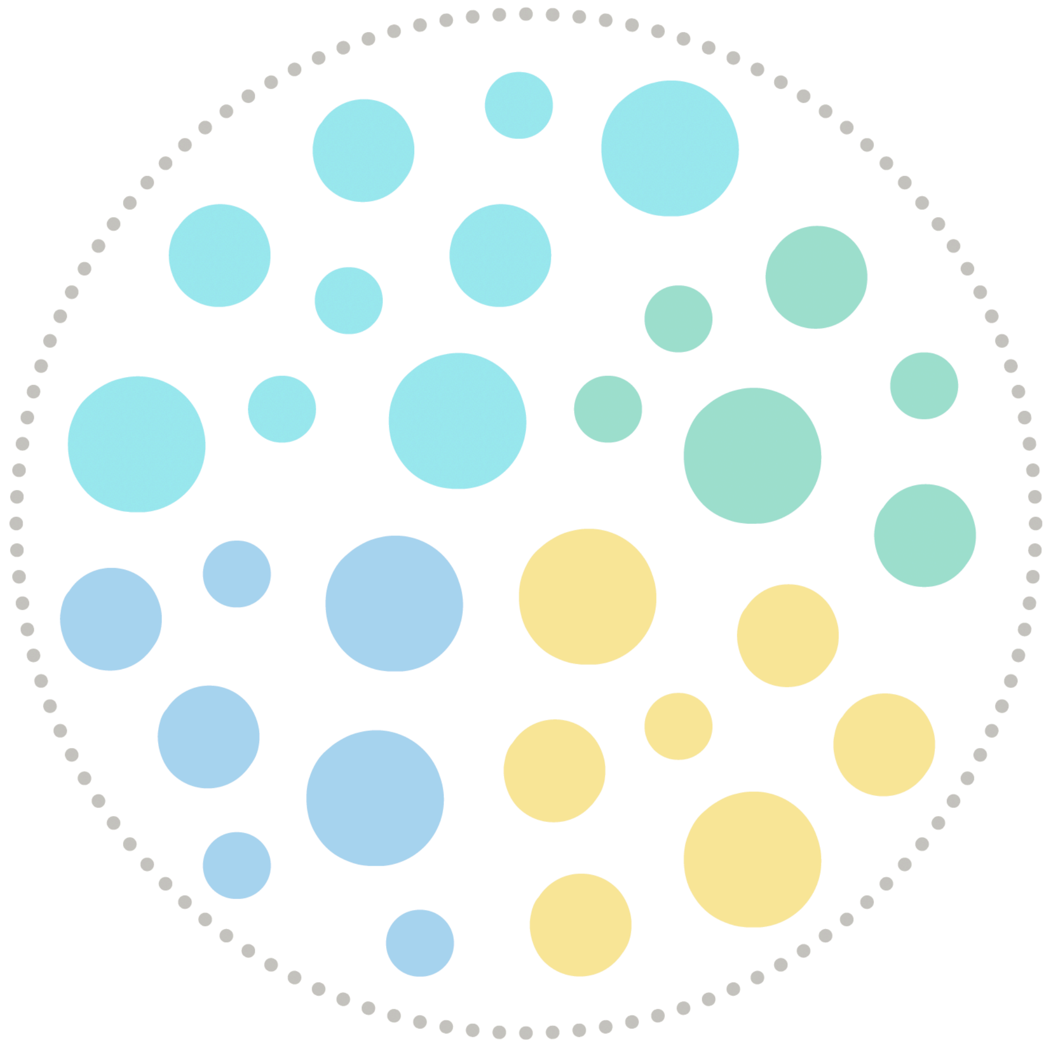
✅ Low variance

☹️ No class-conditional coverage guarantee

Classwise CP

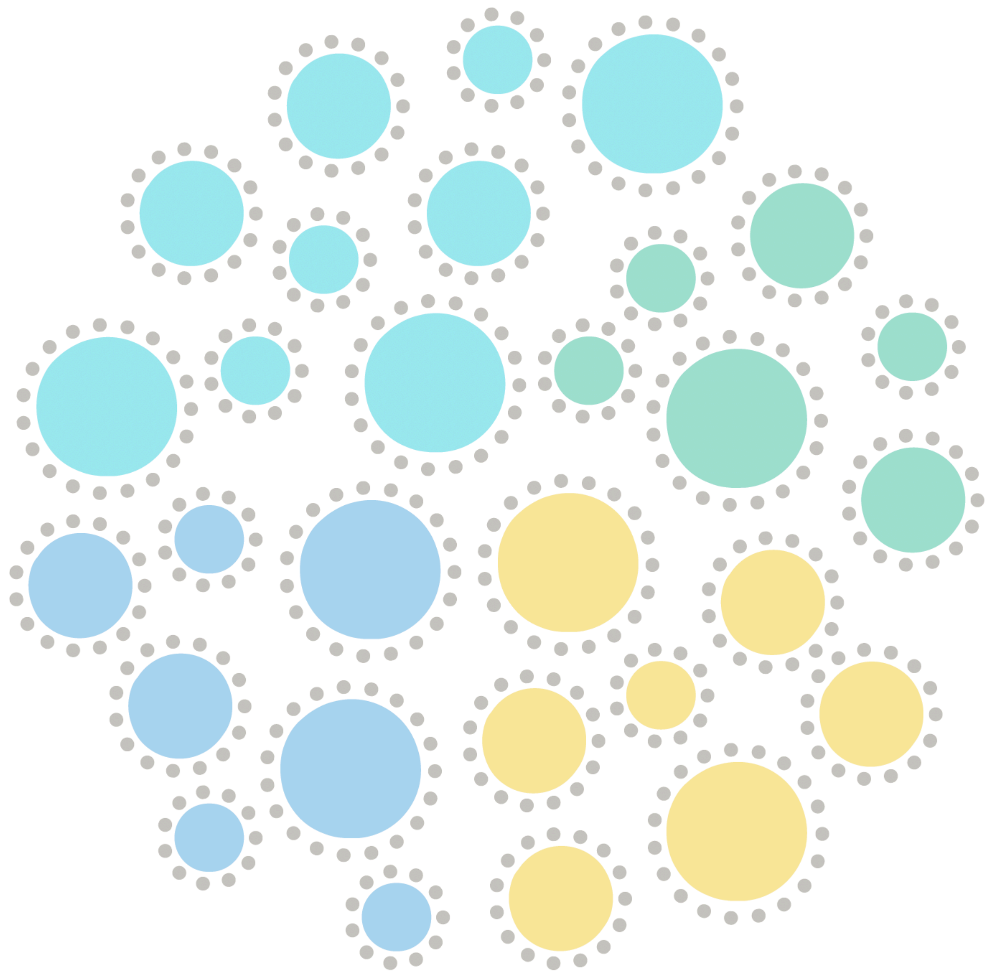
☹️ High variance

✅ Class-conditional coverage guarantee



Standard CP

Can we get the best
of both worlds?



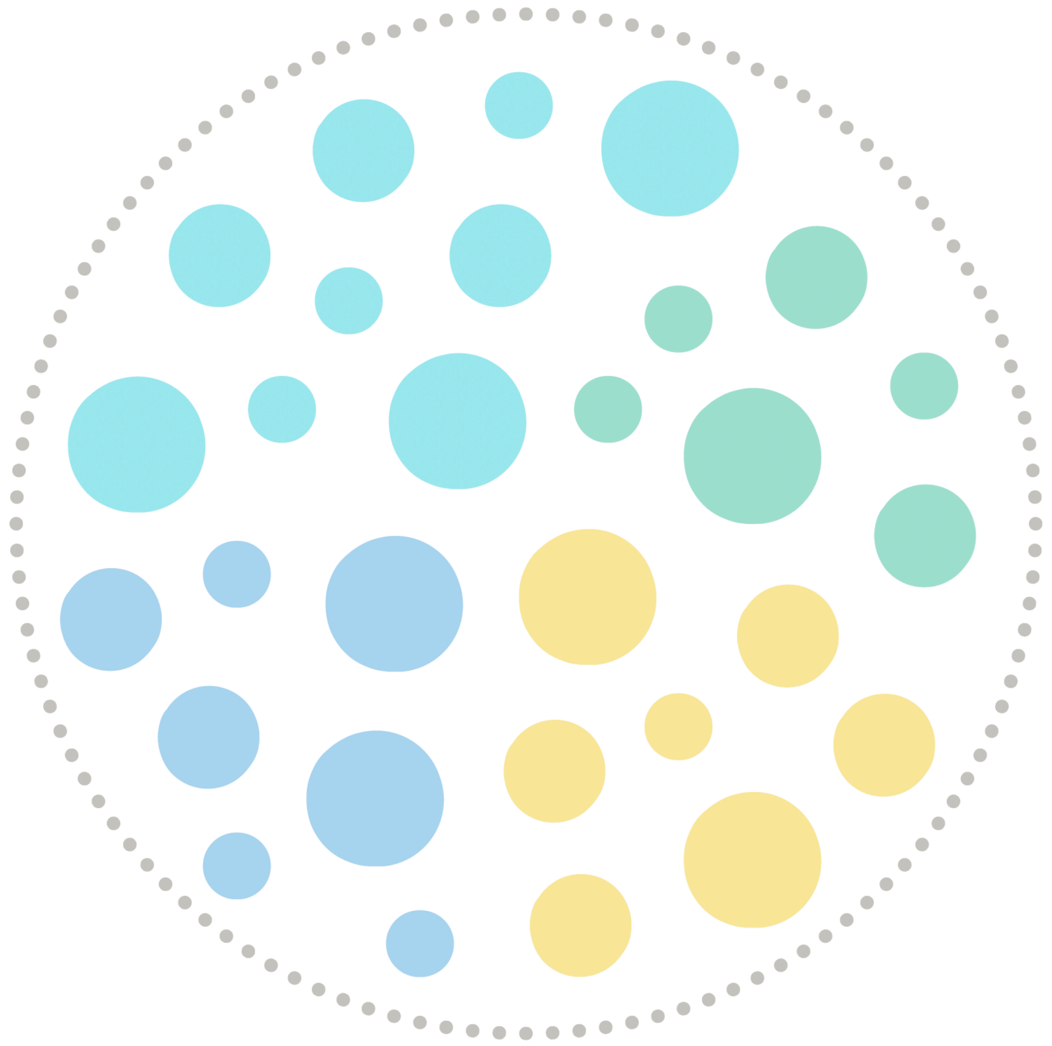
Classwise CP

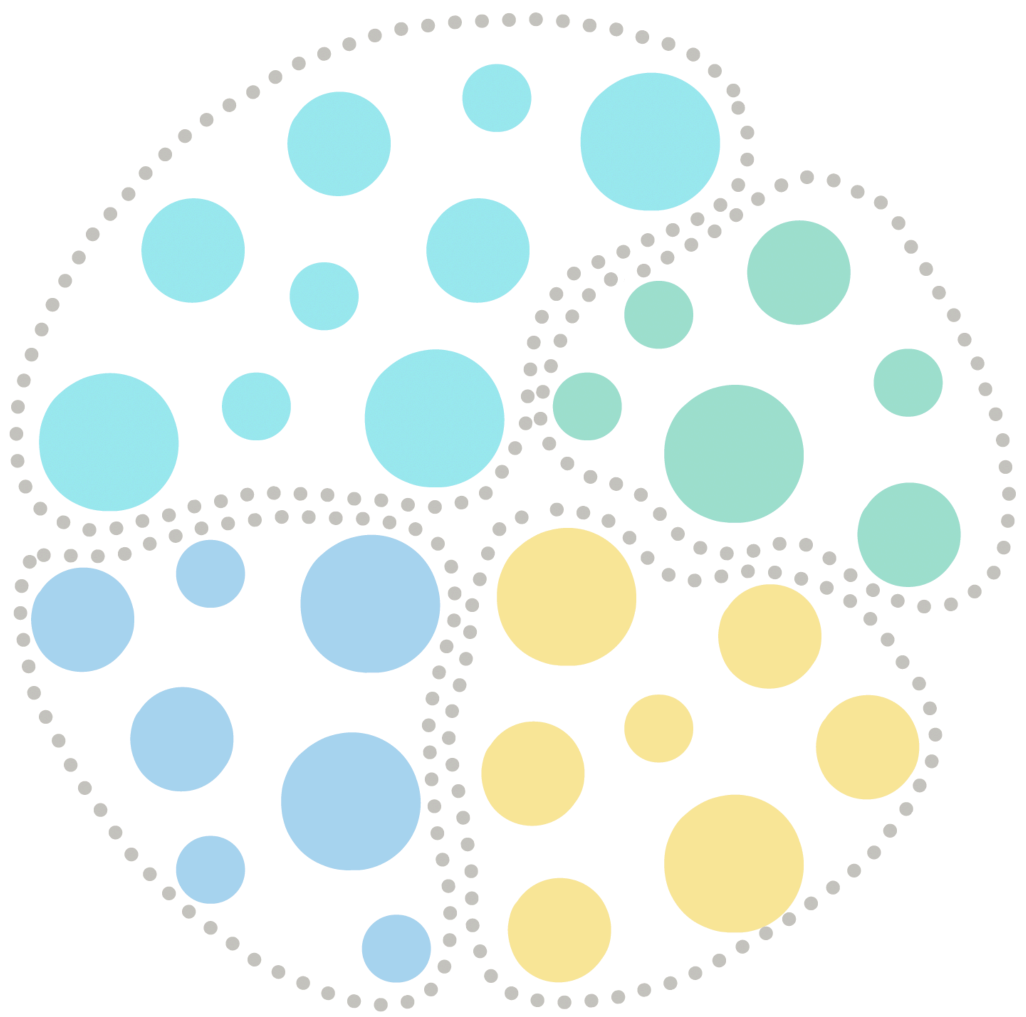
✅ Low variance

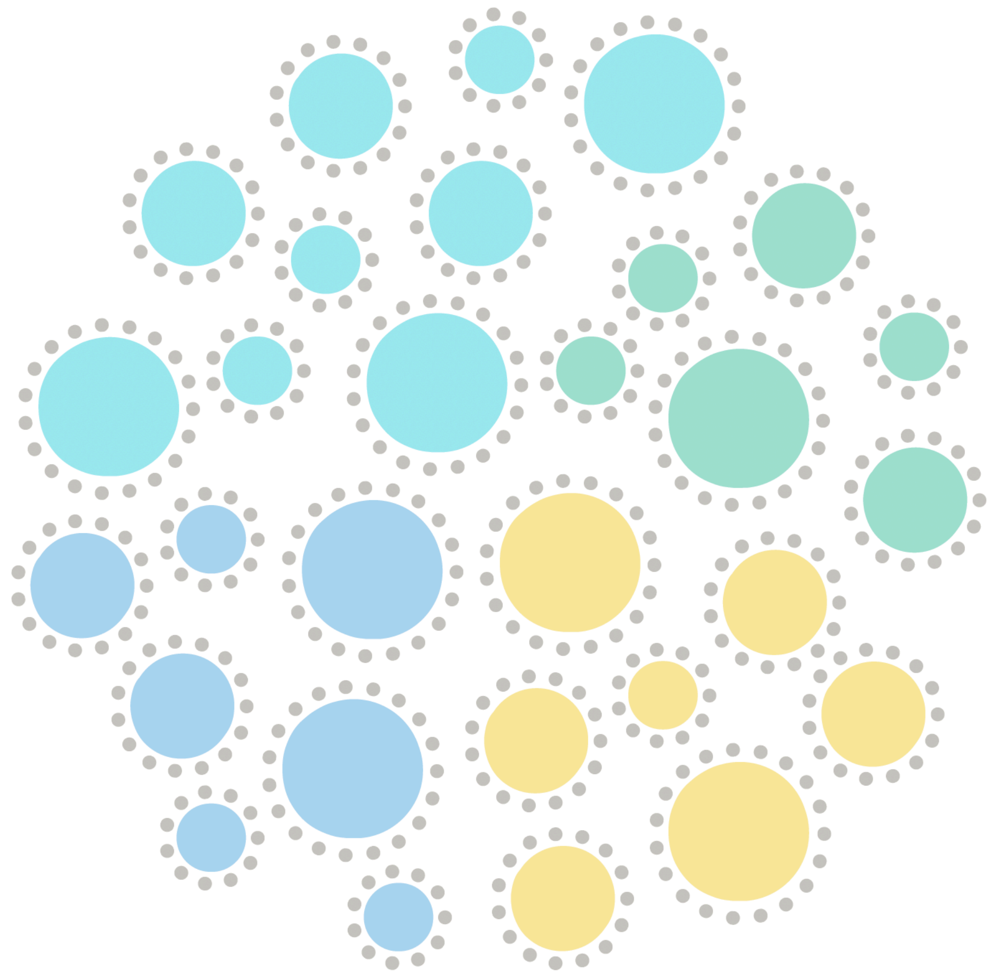
☹️ No class-conditional coverage guarantee

☹️ High variance

✅ Class-conditional coverage guarantee







Standard CP | **Our method: Clustered CP** | **Classwise CP**

✅ Low variance

☹️ No class-conditional coverage guarantee

🔑 **Key idea:**

Combine data from classes that are "similar"

☹️ 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 $\hat{h}$?

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

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

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

for all clusters $m = 1, \ldots, M$

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

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

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

for all clusters $m = 1, \ldots, M$

But our goal is to get class-conditional coverage:

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

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

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

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

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

for all clusters $m = 1, \ldots, M$

?

But our goal is to get class-conditional coverage:

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

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.

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**.

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

Clustered CP

(as a diagram)

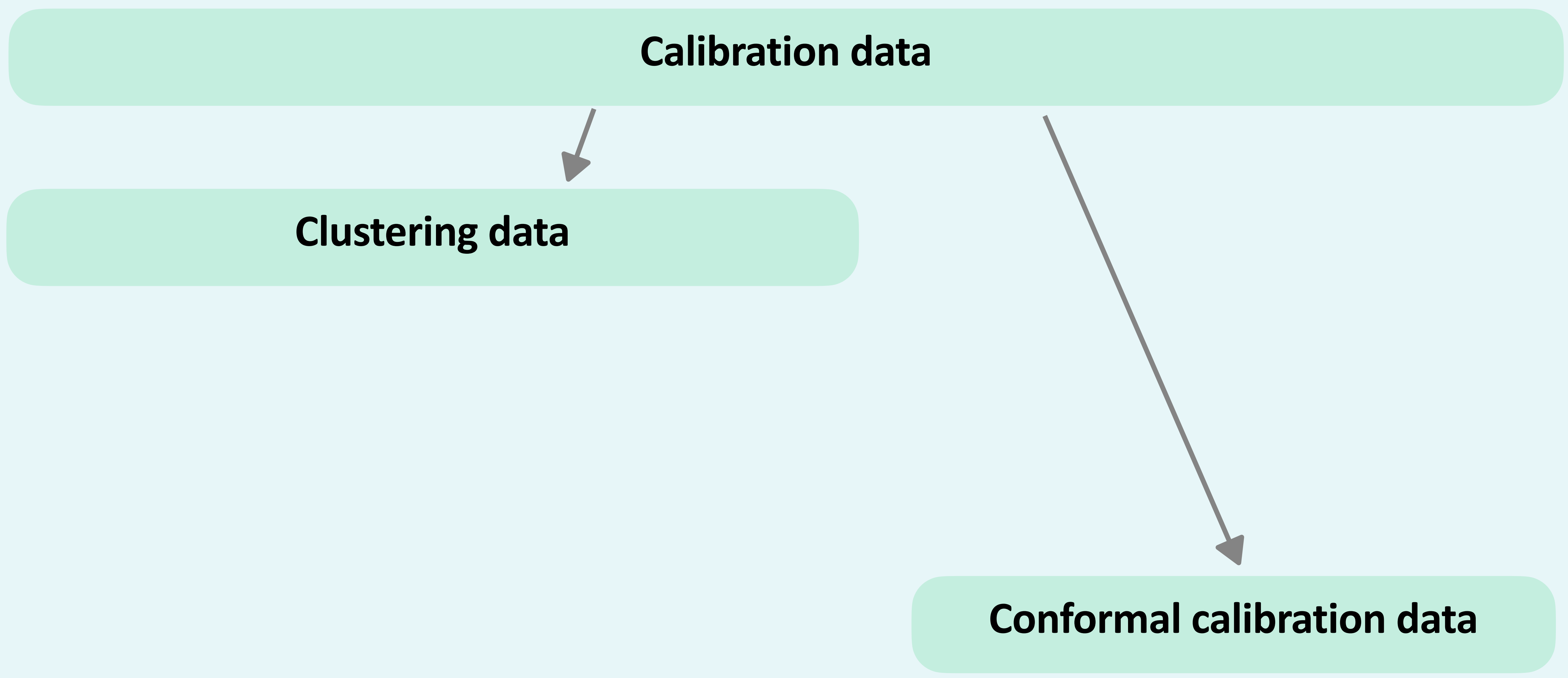
Clustered CP

(as a diagram)

Calibration data

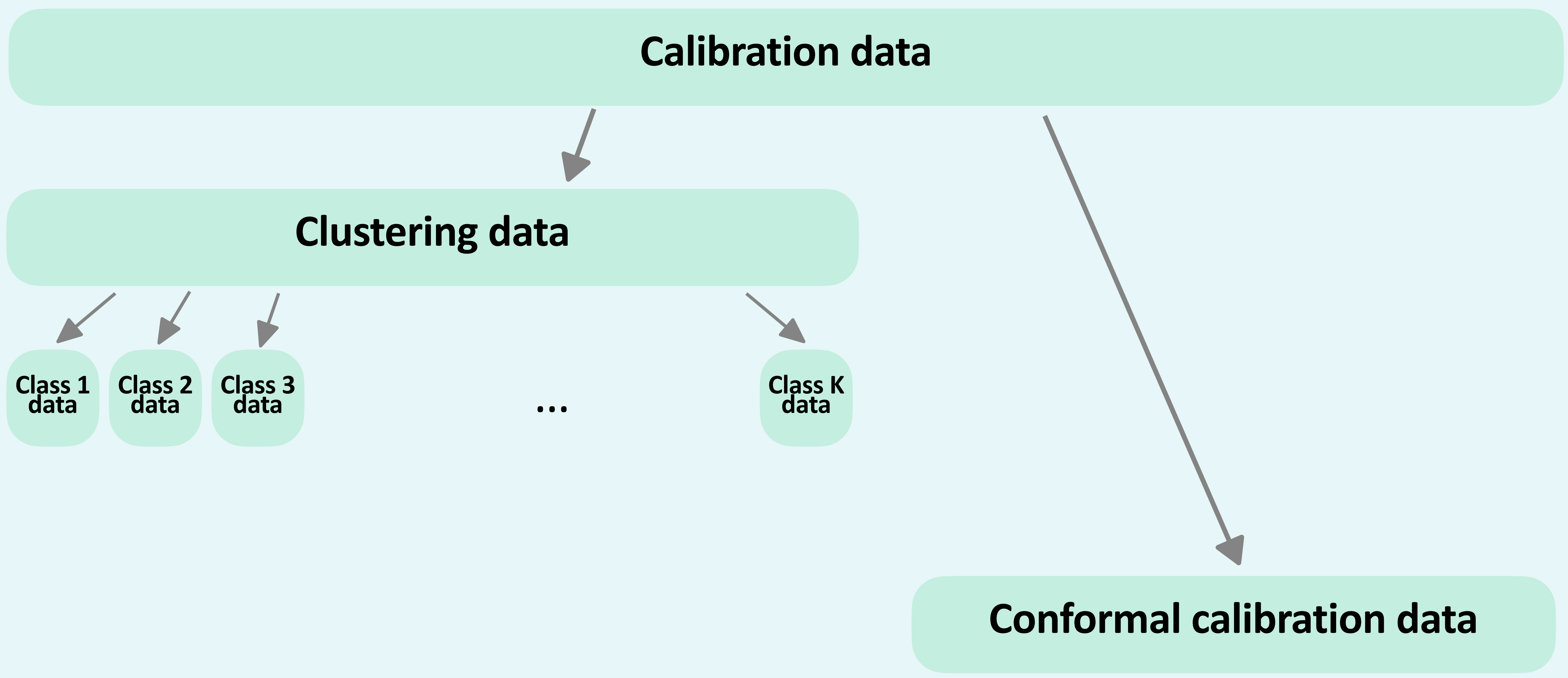
Clustered CP

(as a diagram)



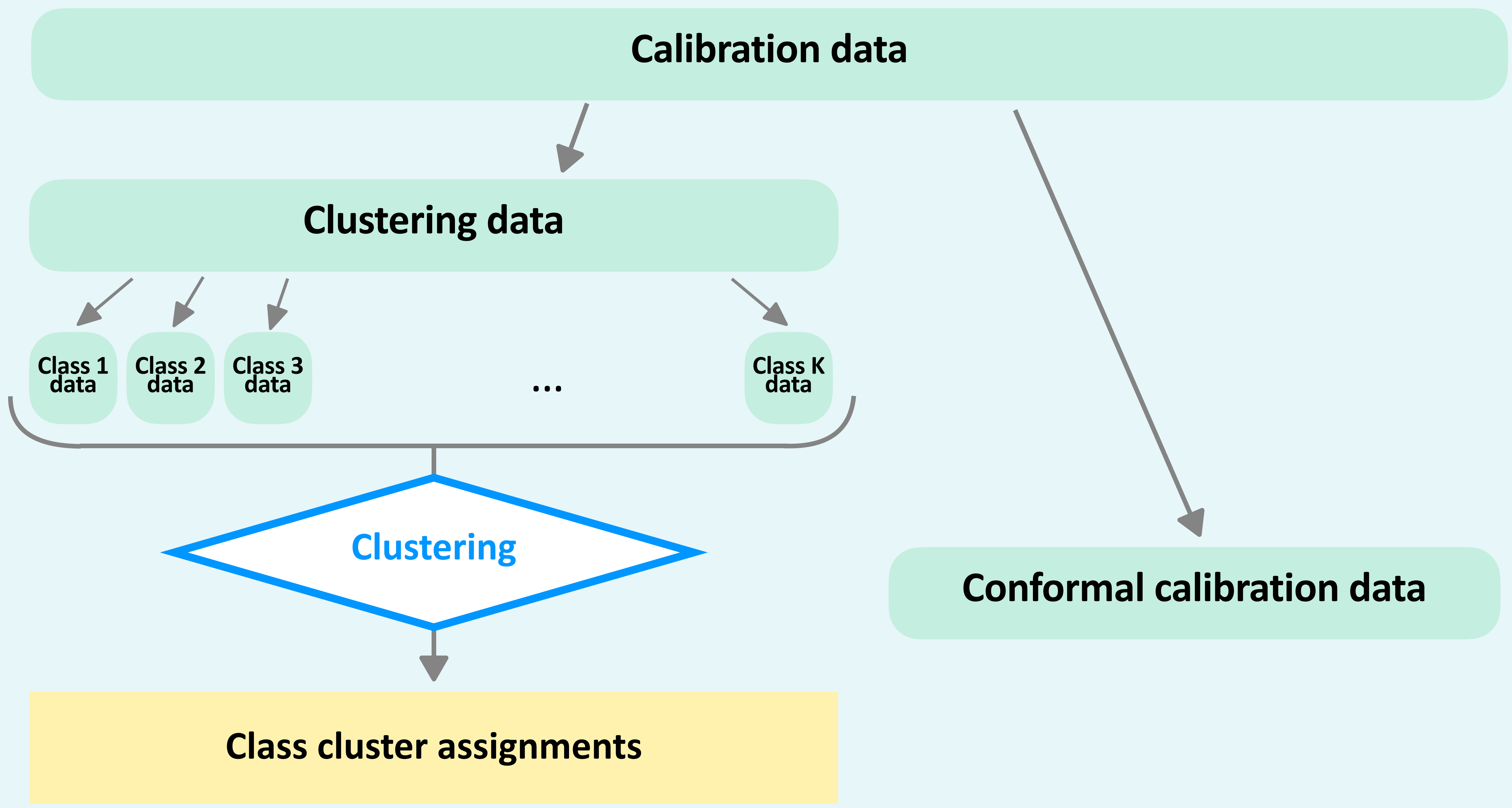
Clustered CP

(as a diagram)



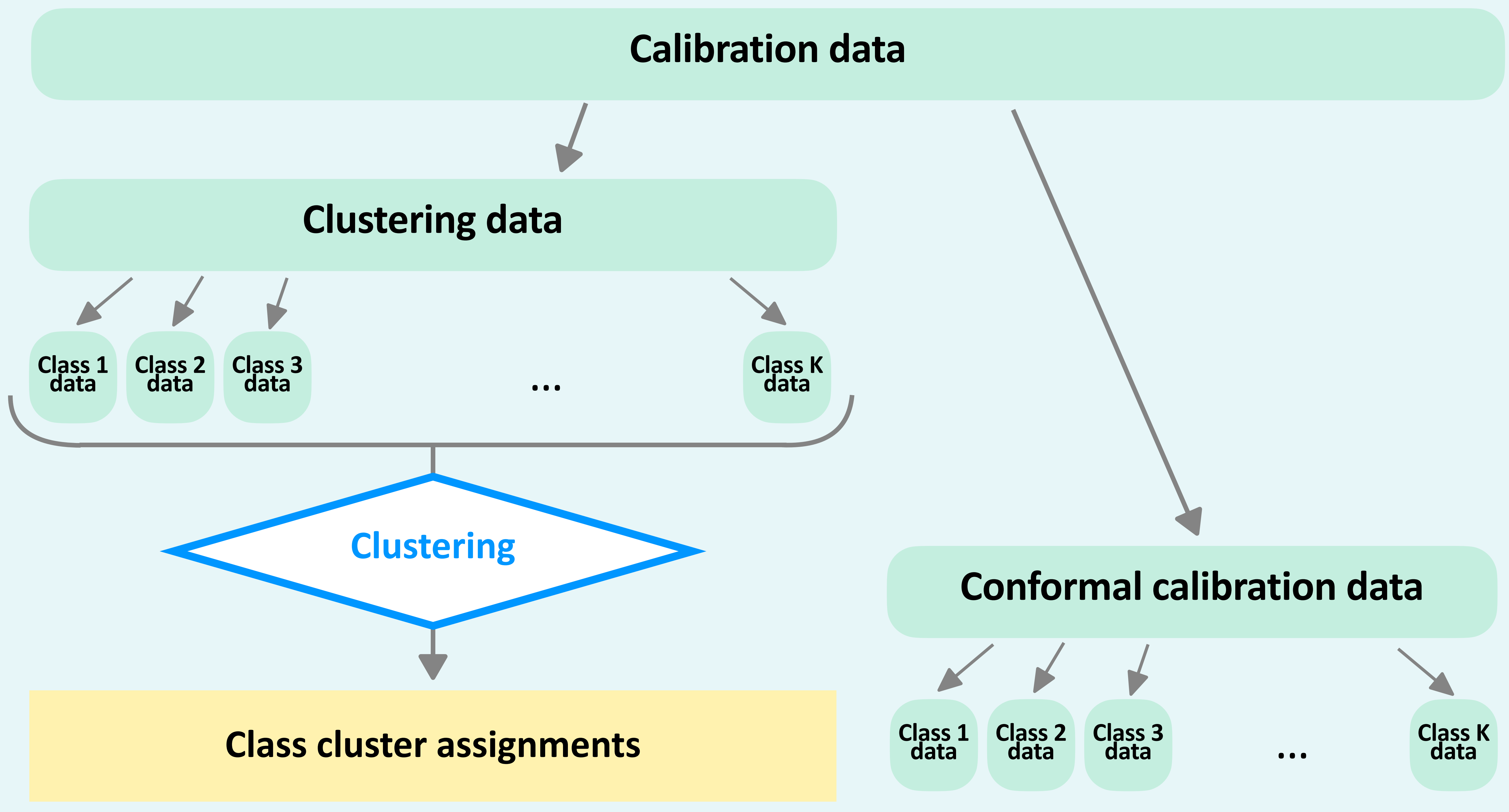
Clustered CP

(as a diagram)



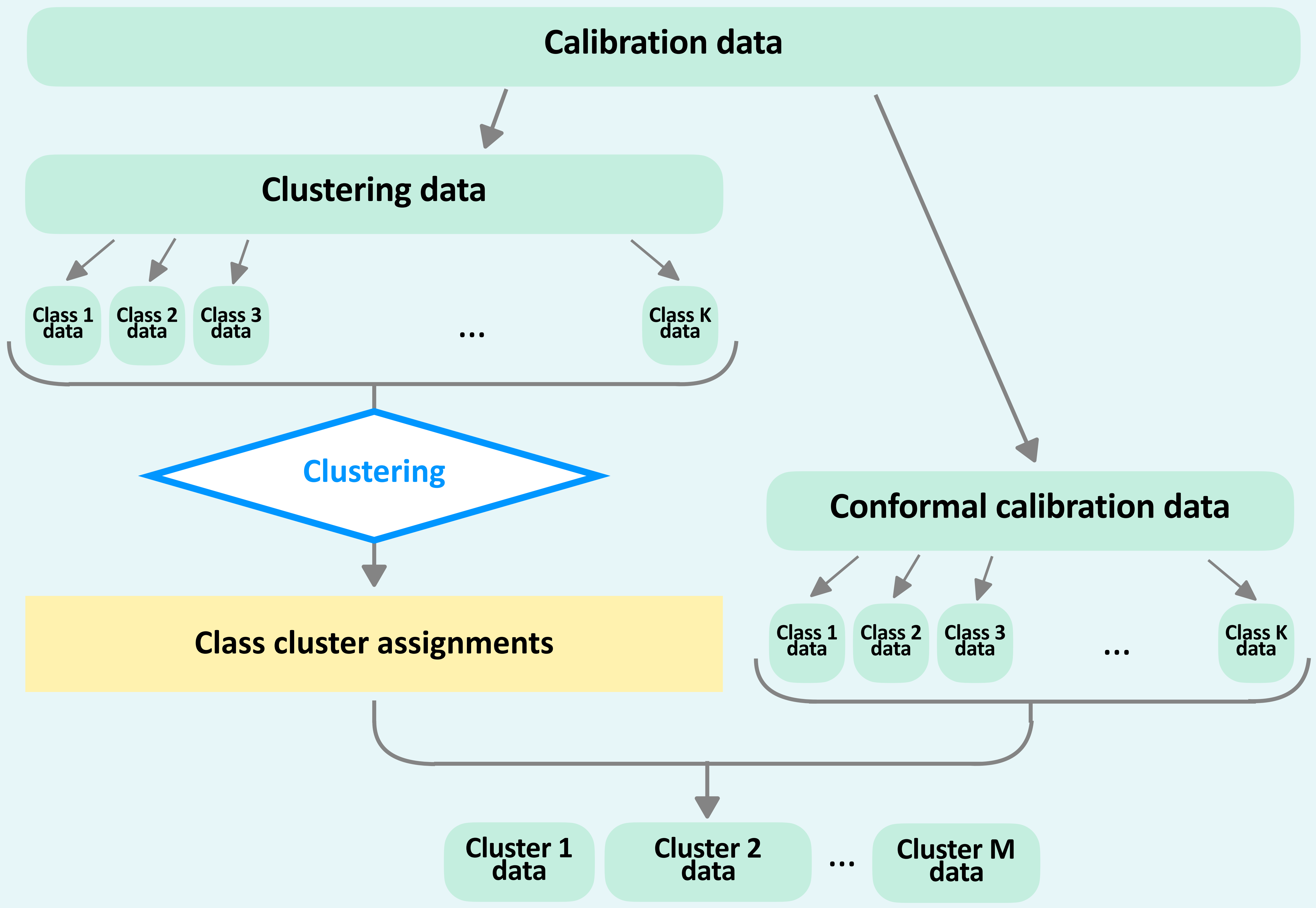
Clustered CP

(as a diagram)



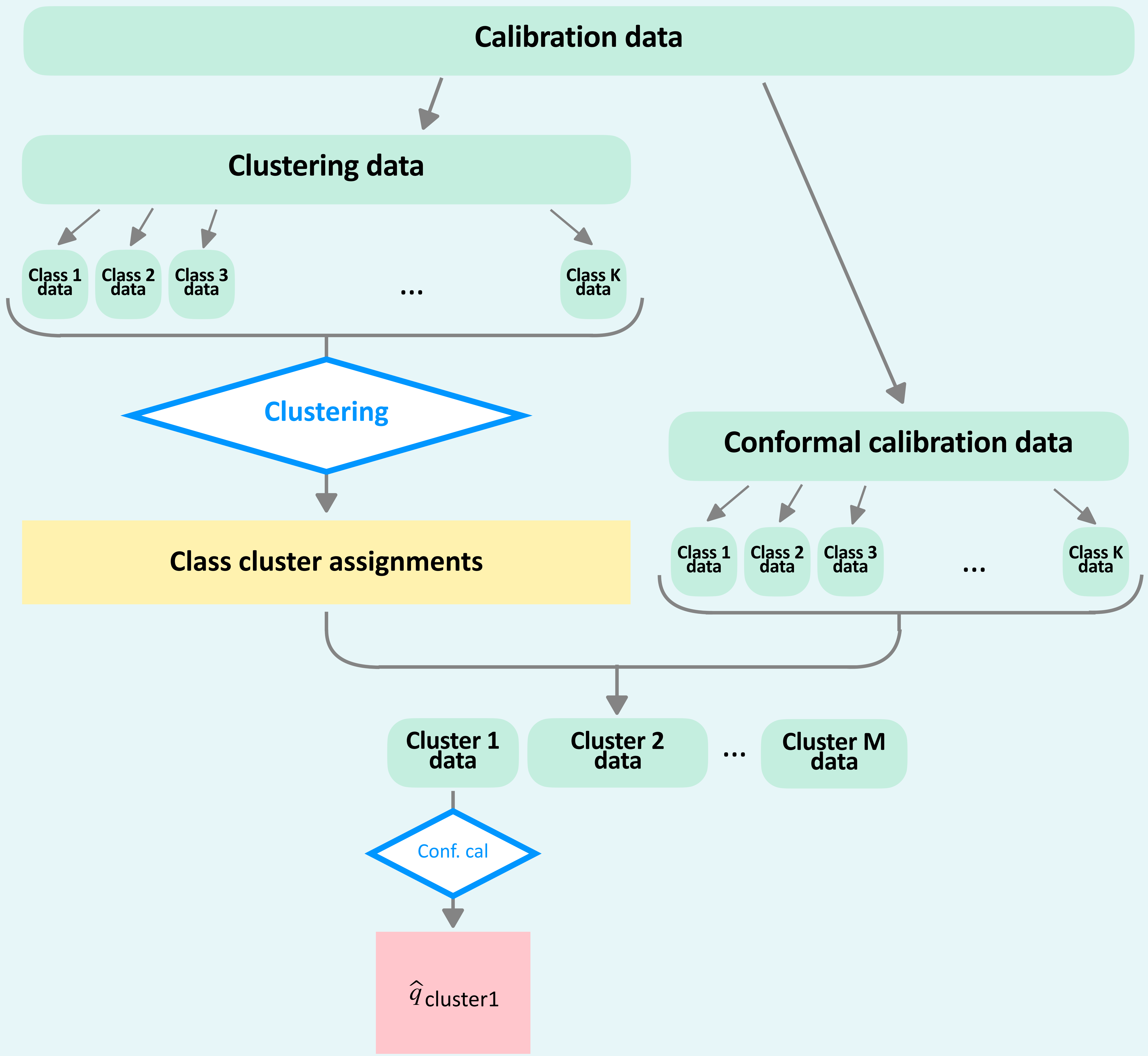
Clustered CP

(as a diagram)



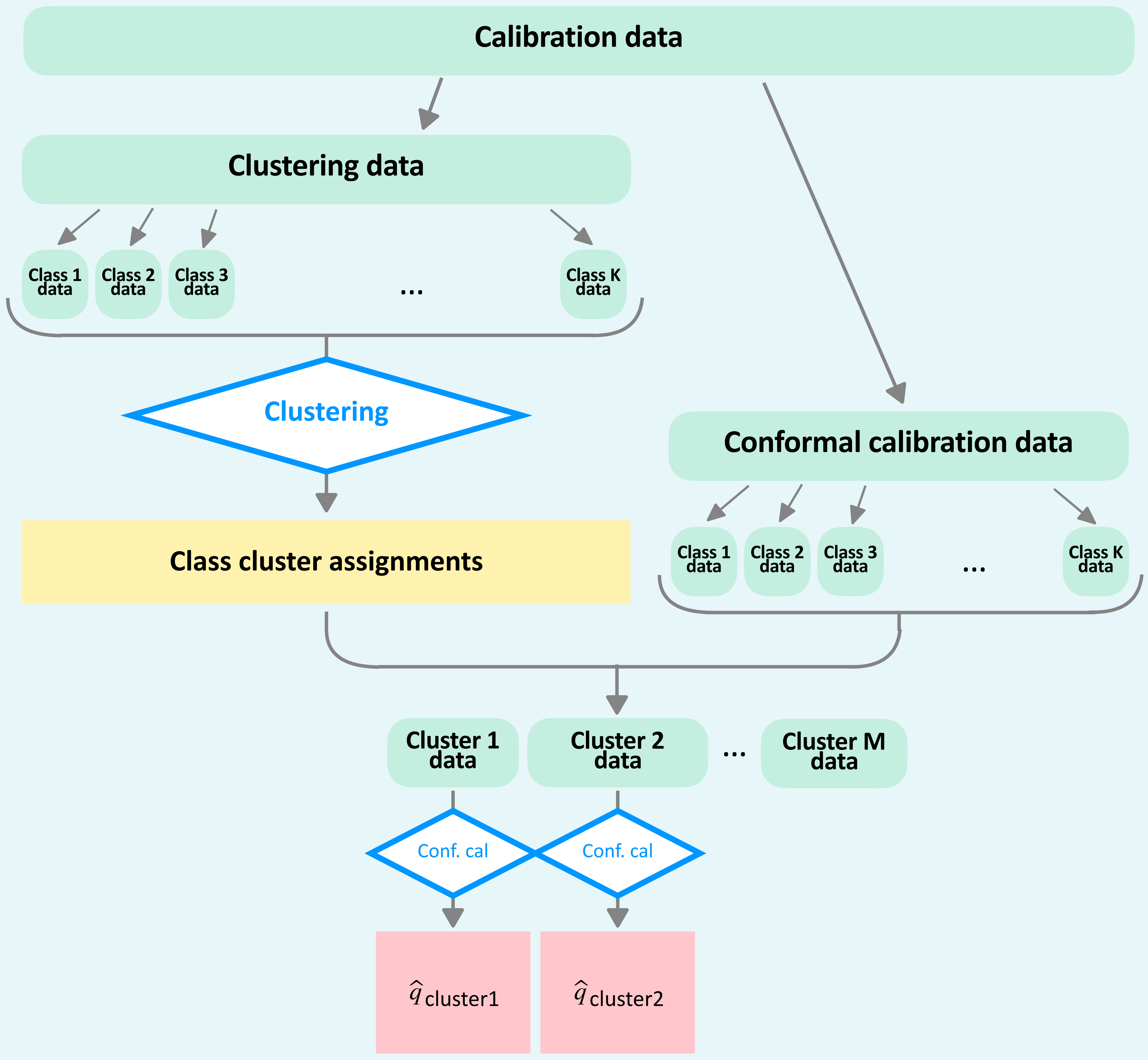
Clustered CP

(as a diagram)



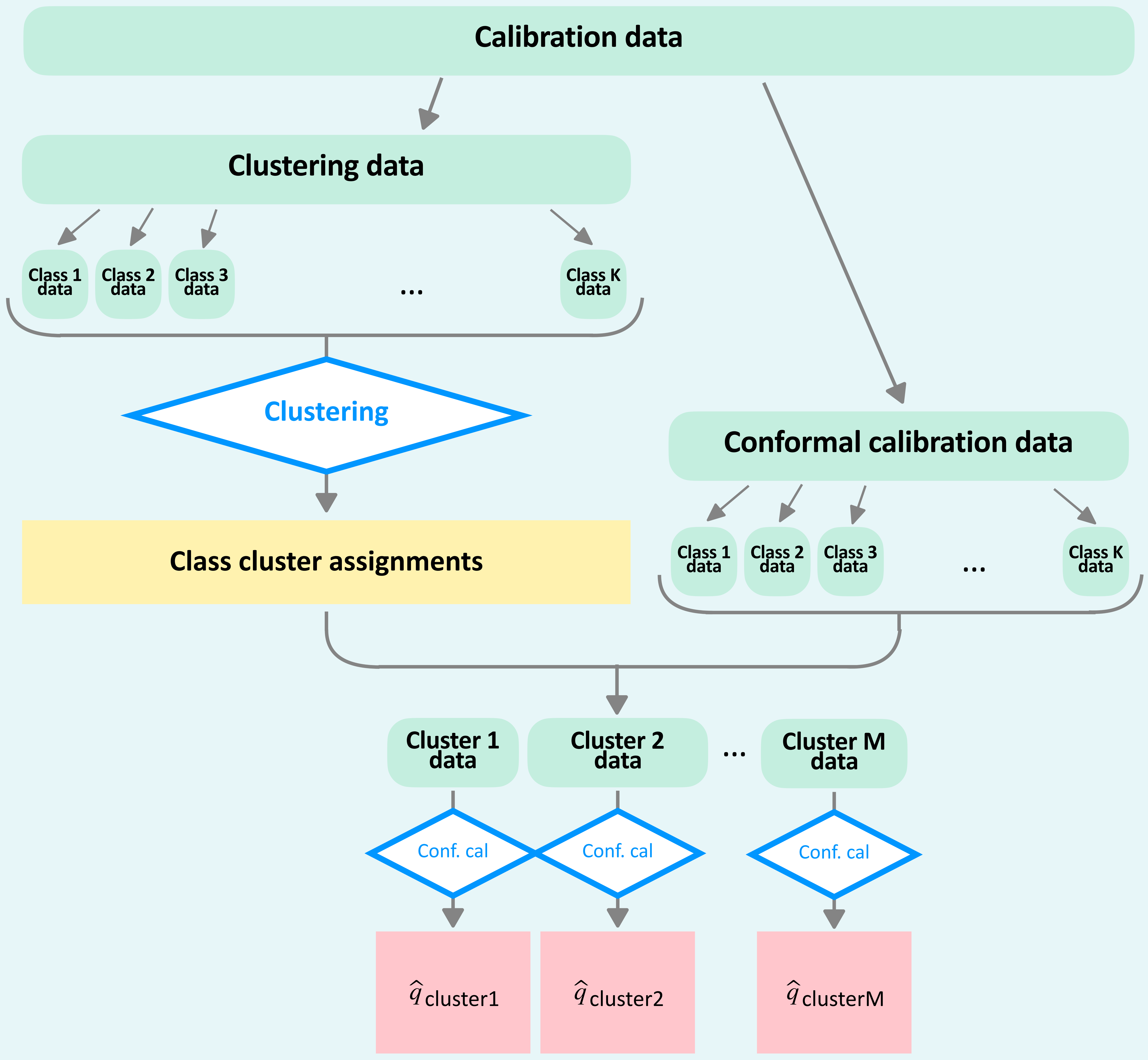
Clustered CP

(as a diagram)



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 $\hat{h}$ satisfy

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

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.

Conformal score functions

softmax: 1 - (softmax score of base classifier)

APS: designed to achieve better X-conditional coverage

RAPS: regularized version of APS that often produces smaller sets

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* $(1-\alpha)$*?*

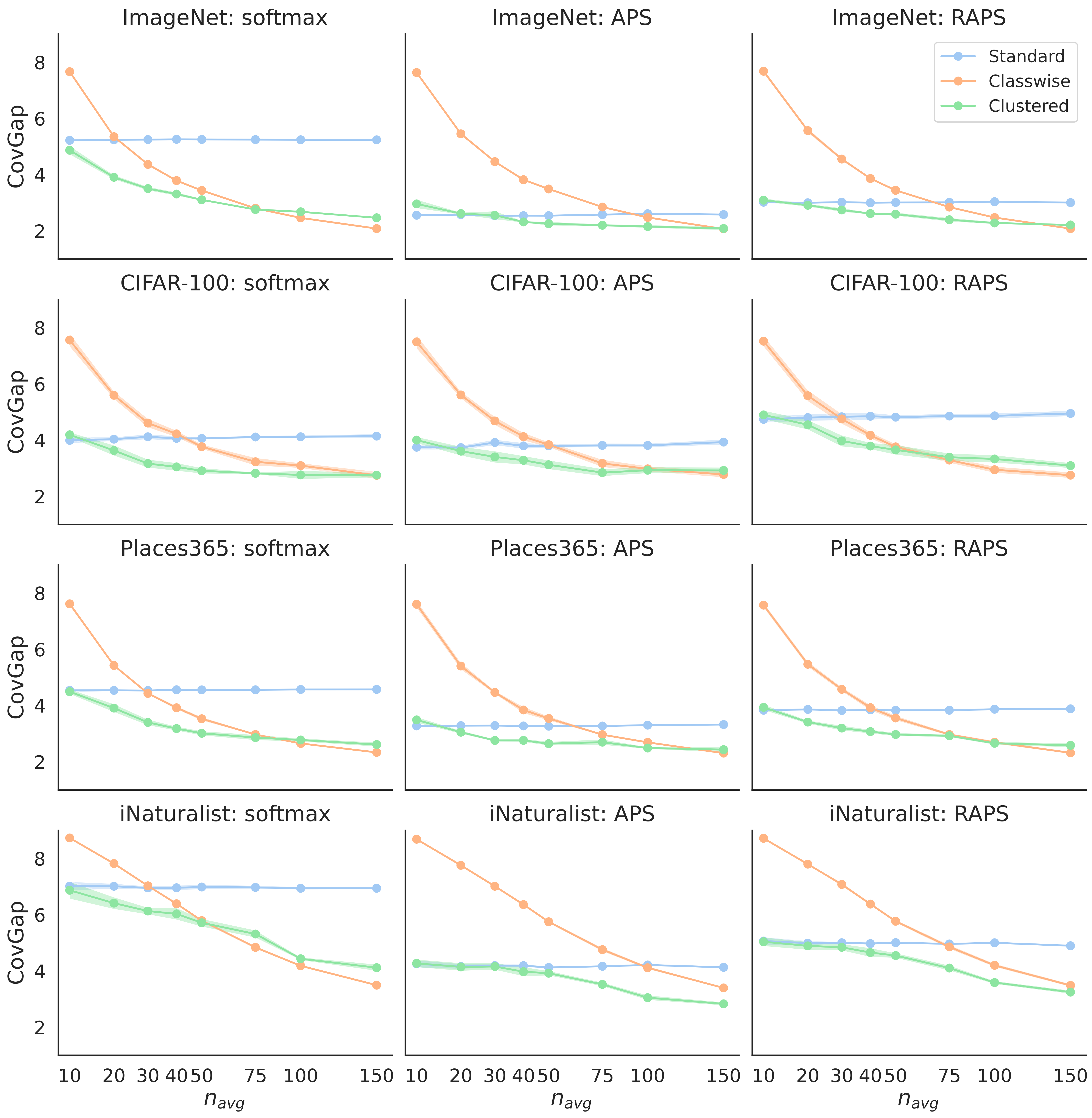
$$\text{CovGap} = 100 \times \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} |\hat{c}_y - (1-\alpha)|$$

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

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

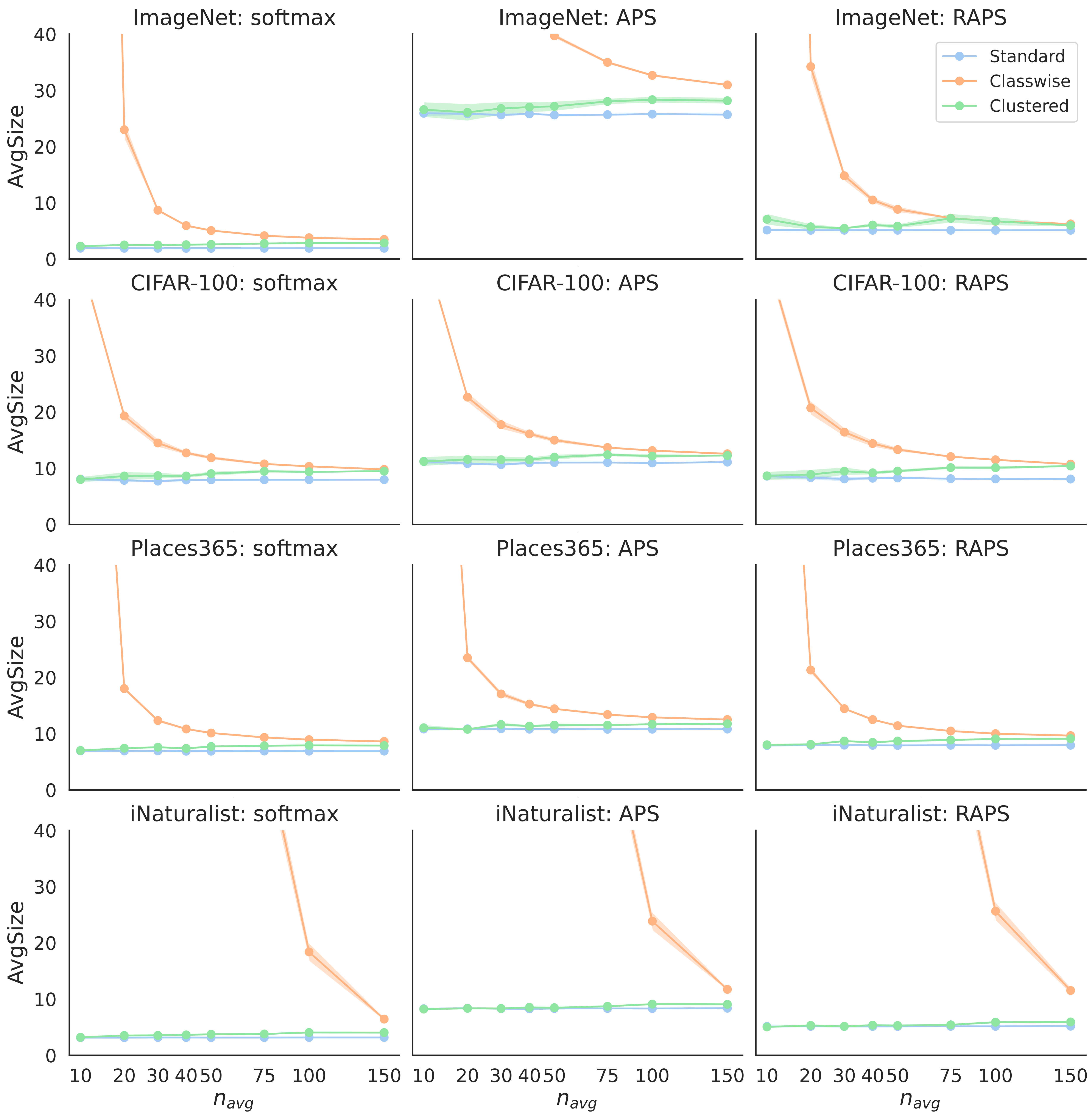
$$\text{CovGap} = 100 \times \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} |\hat{c}_y - (1-\alpha)|$$

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?*

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



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

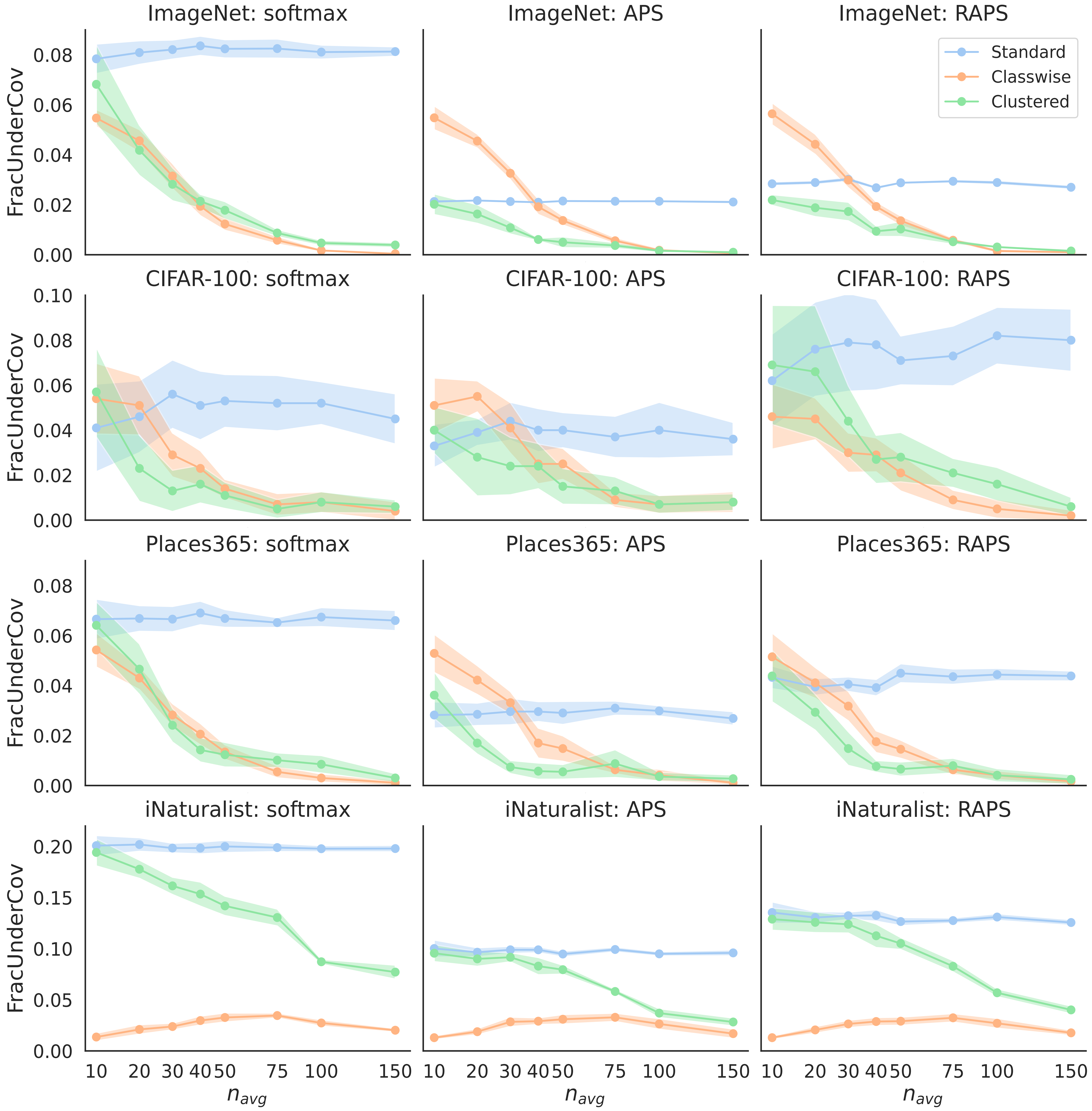
$$\text{FracUnderCov} = \frac{1}{|\mathcal{Y}|}\sum_{y=1}^{|\mathcal{Y}|} \mathbf{1}\{\hat{c}_y \leq 1-\alpha-0.1\}$$

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

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

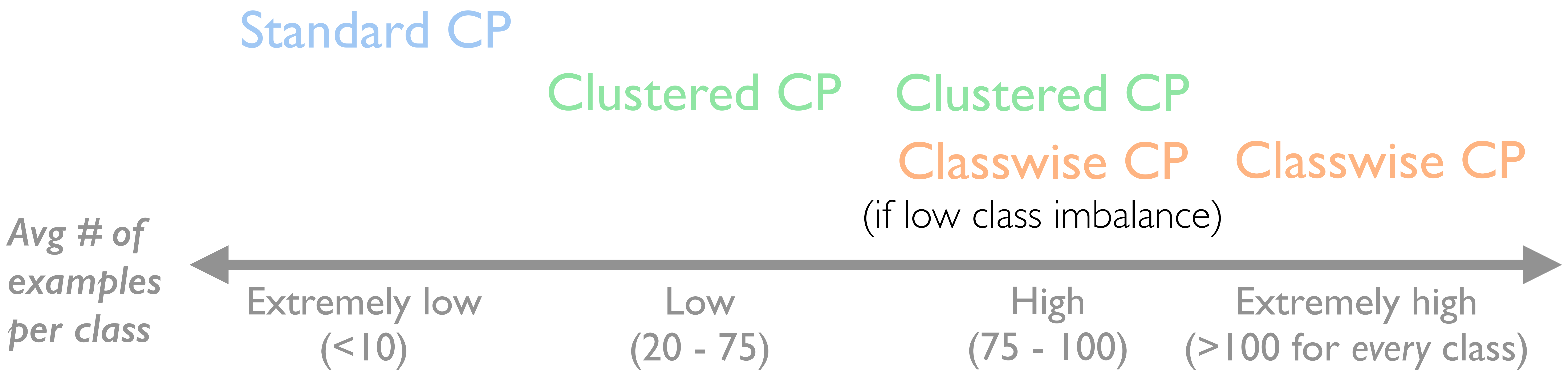
$$\text{FracUnderCov} = \frac{1}{|\mathcal{Y}|}\sum_{y=1}^{|\mathcal{Y}|} \mathbf{1}\{\hat{c}_y \leq 1 - \alpha - 0.1\}$$

* 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?*



Conclusion

Summary

1. Marginal coverage is not enough. In many settings, we want to have class-conditional coverage.
2. Class-conditional coverage is hard to achieve when there are many classes and limited data per class.
3. Clustering classes with similar score distributions allows us to share data between classes in a way that will achieve good class-conditional coverage

Future directions?

Generalizing our clustering approach to achieve group-conditional coverage for any grouping.

Thanks!

For more details:

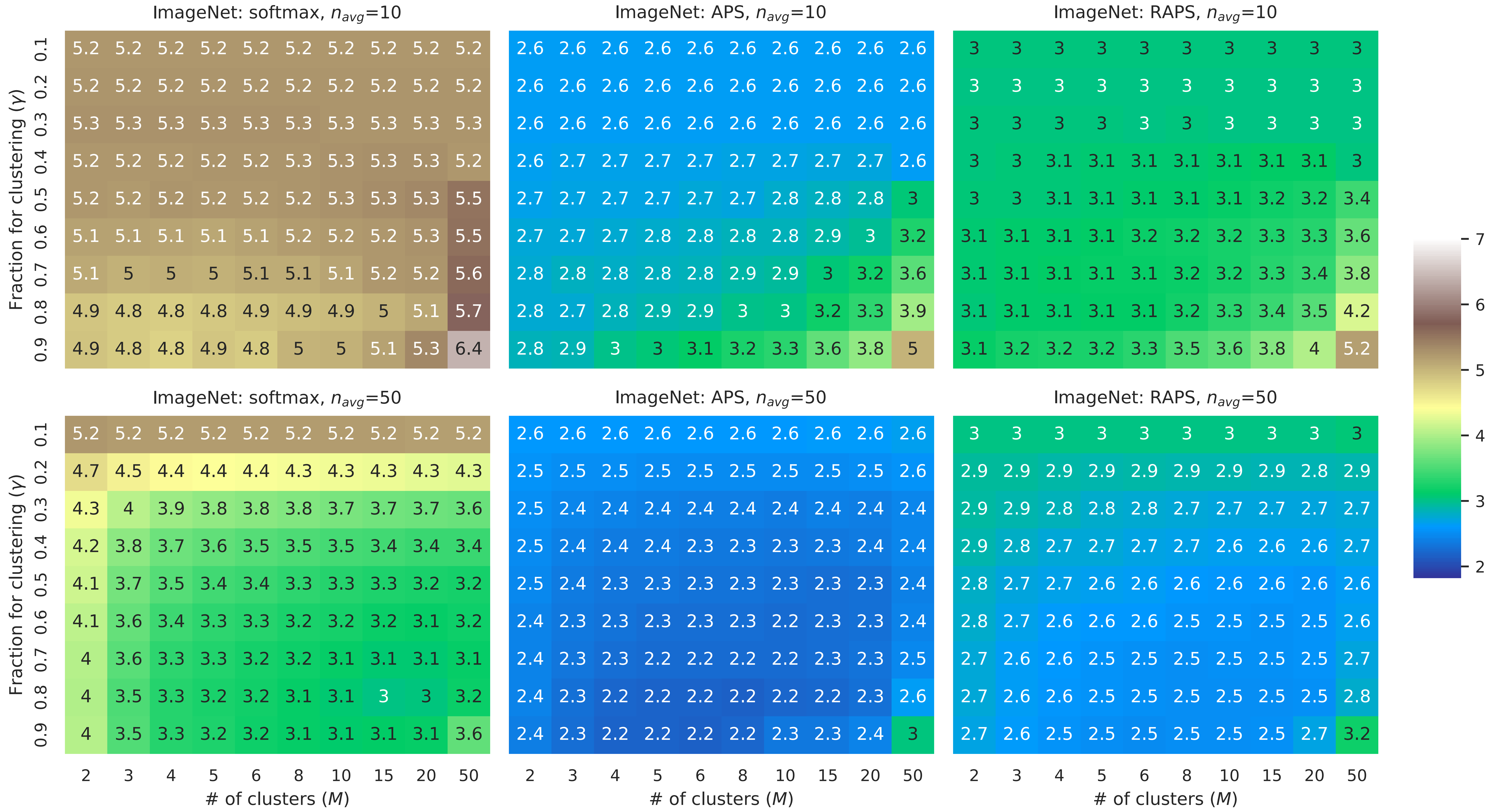
arxiv.org/abs/2306.09335

To try it yourself:

Paper code: github.com/tiffanyding/class-conditional-conformal

PyTorch implementation by SUSTech: github.com/ml-stat-Sustech/TorchCP

Sensitivity analysis for Clustered CP parameters



Randomized versions to achieve exact $1-\alpha$ coverage