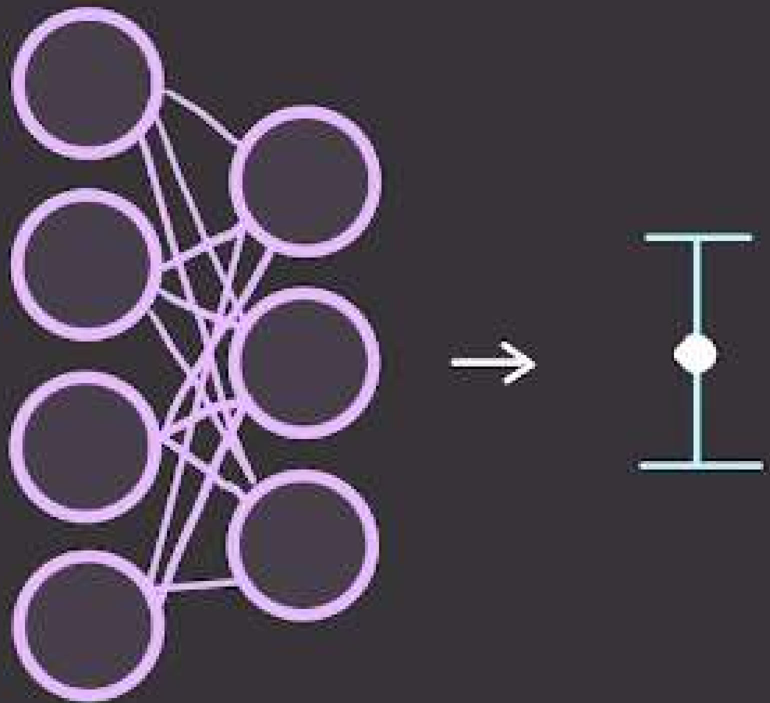
A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification



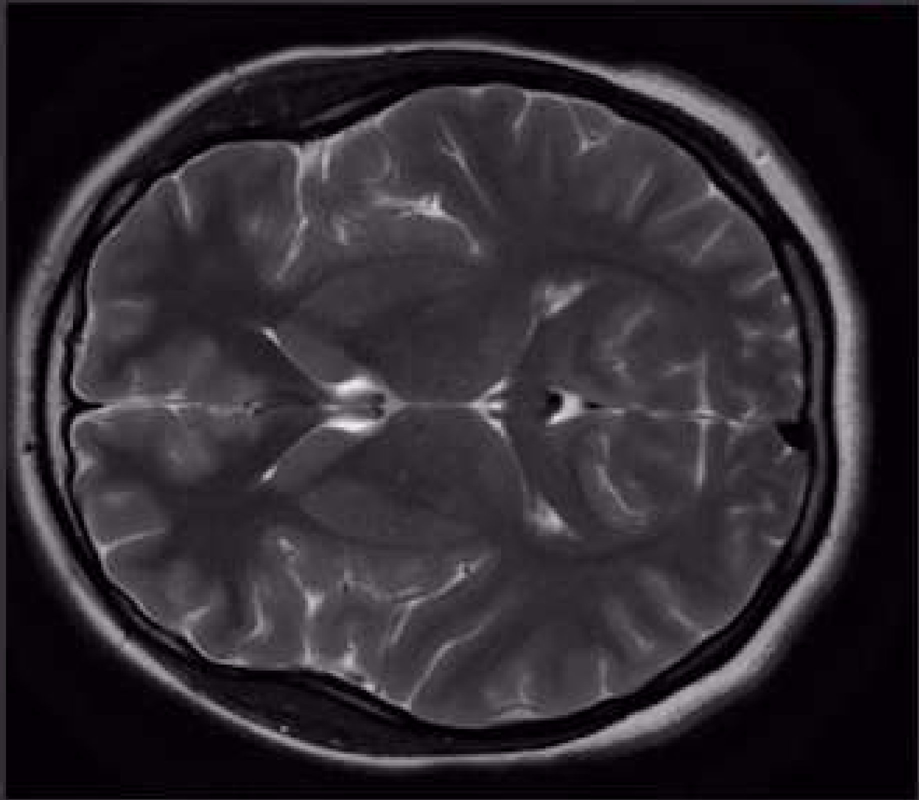
Rigorous, finite-sample CIs for any model and any dataset for free.

+ Black box. (No analysis of model, no training modification.)

Tutorial style, starting w/conformal prediction.

Motivation:

Downstream decision-making tasks.



Accuracy is not enough!!

Need instance-wise UQ.

rule in or rule out harmful diagnoses

❌ ignore NNs

⚠️ Analyze them

➕ treat them as black box

$$\mathbb{P}\left[\text{true diagnosis} \in \{\text{Normal}, \text{concussion}, \text{cancer}\}\right] \geq 90\%$$

Let's start with classification.

Goal:

"image" $\in \mathbb{R}^d$ "class" $\in \mathcal{Y} = \{1, \dots, K\}$

Given a "calib dataset" $\{(X_i, Y_i)\}_{i=1}^{n} \sim \mathbb{P}$ i.i.d.,

a "model" $\hat{\pi}_y(x)$, estimating $\pi_y(x) = \mathbb{P}[Y=y \mid X=x]$.

Notation: the true conditional distribution.

a "new image" X_{n+1}

$$\boxed{\text{predict a SET, } \mathcal{T}(X_{n+1}) \subseteq \mathcal{Y},}$$

which contains the true class Y_{n+1} with high $\mathbb{P}$.

Definition "COVERAGE"

$$\mathbb{P}\left[Y_{n+1} \in \mathcal{T}(X_{n+1})\right] \geq 1 - \alpha.$$

"Conservative coverage."

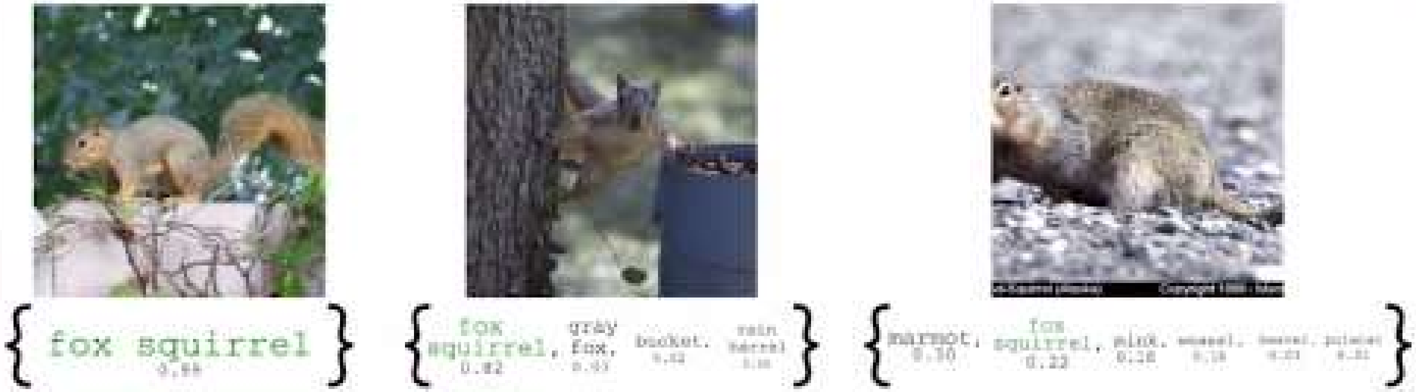


Figure 1: **Prediction set examples on Imagenet.** We show three examples of the class `fox squirrel` and the 95% prediction sets generated by `RAPS` to illustrate how the size of the set changes as a function of the difficulty of a test-time image.

(Angelopoulos, Bates, Malik, Jordan)

Objectives:

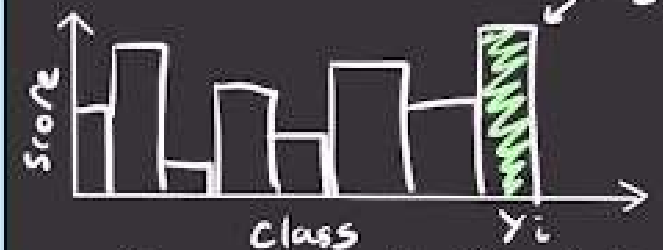
- Exact coverage $\circ \circ \bigcirc \left\{ \begin{array}{ll} y & \text{w/p } 1-\alpha \\ \emptyset & \text{w/p } \alpha \end{array} \right.$?
- Small size $|T(X)|$
- "Adaptive" i.e. smaller for easy
  bigger for hard

Conformal Prediction (Vovk et al.)

What if we **LEARN** a rule to predict sets?

Using the calibration dataset $\{(X_i, Y_i)\}_{i=1}^{n}$

1. Get score of correct class



for each of (X_i, Y_i).

2. Take the 10% quantile.

$\hat{q}$ = np.quantile(
    $[E_1, \dots, E_n]$, 0.1, 'lower')

— small correction

"At least 90% of examples have the class score above $\hat{q}$."

3. Form prediction sets

{All classes whose score exceeds $\hat{q}$ when X_{n+1} input}.

= Valid Prediction Set!

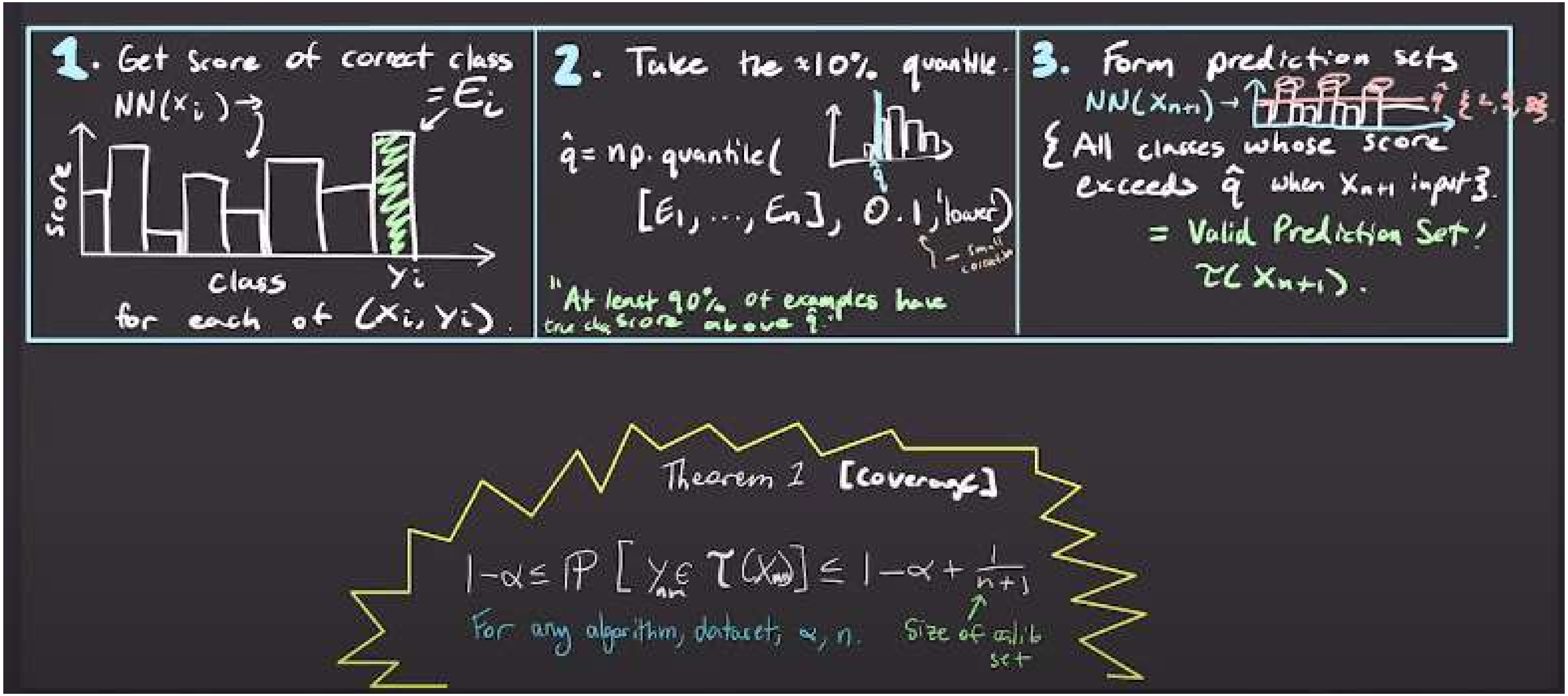
$\mathcal{T}(X_{n+1})$.

Theorem 1 **[Coverage]**

$$1-\alpha \le \mathbb{P}\left[Y_{n+1} \in \mathcal{T}(X_{n+1})\right] \le 1-\alpha + \frac{1}{n+1}$$

For any algorithm, dataset, α, n.

Size of calib set


1. Get score of correct class
NN(xi) →
= Ei
score
class
yi
for each of (Xi, Yi).
2. Take the ≈10% quantile.
q̂ = np.quantile(
[E1, ..., En], 0.1, 'lower')
← small correction
"At least 90% of examples have true class score above q̂."
3. Form prediction sets
NN(Xn+1) →
q̂
{All classes whose score exceeds q̂ when Xn+1 input}.
= Valid Prediction Set!
τ(Xn+1).
Theorem 1 [Coverage]
1−α ≤ P [Yn+1 ∈ τ(Xn+1)] ≤ 1−α + 1/(n+1)
For any algorithm, dataset, α, n.
Size of calib set

Conformal Prediction (General case]

0) Identify a heuristic notion of uncertainty

1) Define a score function $s(x,Y)\in\mathbb{R}$

2) Compute $\hat{q}$: $\frac{\lceil (n+1)(1-\alpha)\rceil}{n}$ quantile of $\underbrace{s(X_1,Y_1),\dots s(X_n,Y_n)}_{\text{calibration points}}$

3) Deploy

$$T(x)=\{y: s(x,y)\le \hat{q}\}$$

recall

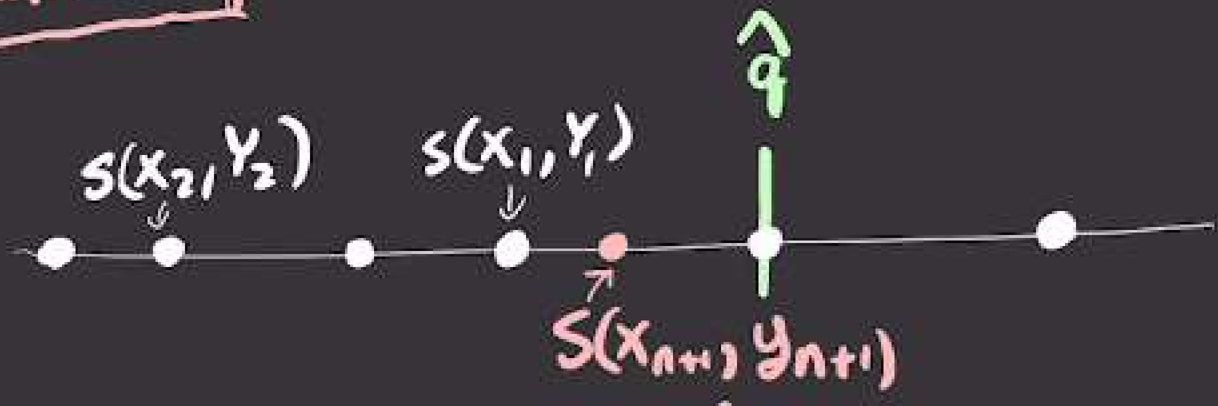
Theorem 1 [coverage]

$$1-\alpha \le \mathbb{P}\left[Y_{n+1} \in \mathcal{T}(X_{n+1})\right] \le 1-\alpha+\frac{1}{n+1}$$

For any algorithm, dataset, α, n.

size of calib set

why? symmetry



Only make mistake if ↑ is too big

Recap

heuristic uncertainty → Conformal → rigorous uncertainty

- ✓ correct coverage
- ✓ any model
- ✓ any data set

Score function matters!
(Important design choice)

Example: Adaptive Prediction Sets

(Romano et al.)

	Previous method	This method	
✓	smallest average size	usually larger size	✗
✗	not very adaptive	designed to be adaptive	✓
✗	only uses output of true class	uses output of all classes	✓

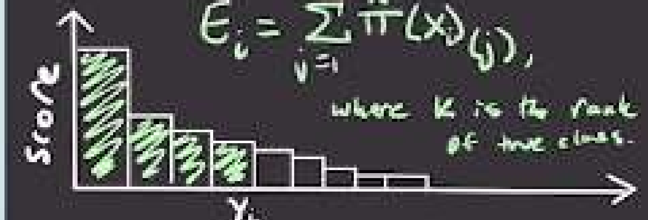
$NN(X_{n+1}) \rightarrow$ 

1. Get score of correct class

$NN(X_i) \rightarrow$ sort $\rightarrow$

$$E_i = \sum_{j=1}^{k} \hat{\pi}(X_i)_{(j)},$$

where k is the rank of true class.



for each of (X_i, Y_i)

2. Take the ≈90% quantile.

$$\hat{q} = \text{np.quantile}\Big([E_1, \ldots, E_n],\ \underbrace{0.9}_{\frac{\lceil (n+1)(1-\alpha)\rceil}{n}},\ \text{'upper'}\Big)$$

3. Form prediction sets

{the k most likely classes, where $\sum_{j=1}^{k} \hat{\pi}(X_{n+1})_{Y_{n+1}} \geq \hat{q}$}.

= Valid Prediction Set! $\tau(X_{n+1})$.

Example: Conformalized Quantile Regression

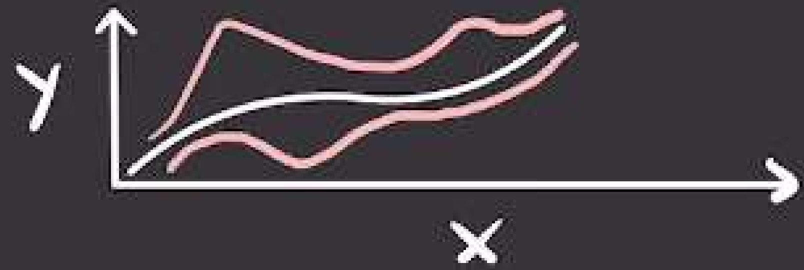
(Romano et al.)

Continuous Y.

Two models: $\hat{t}_{\alpha/2}(X)$ and $\hat{t}_{1-\alpha/2}(X)$,

5% quantile

95% quantile



e.g. NN with pinball loss

The quantile model may be wrong.

IDEA: Inflate by constant to fix their coverage.

1. Get score of correct class

E_i = Proj of Y_i onto $[\hat{t}_{\alpha/2}(X_i), \hat{t}_{1-\alpha/2}(X_i)]$,

for each of (X_i, Y_i).

2. Take the ≈90% quantile.

```
q̂ = np.quantile(
      [E1, ..., En], 0.9,'upr')
```

3. Form prediction sets

$$\mathcal{T}(X_{n+1}) = [\hat{t}_{\alpha/2}(X_{n+1}) - \hat{q}, \hat{t}_{1-\alpha/2}(X_{n+1}) + \hat{q}]$$

= Valid Prediction Set! $\mathcal{T}(X_{n+1})$.

MORE EXAMPLES IN OUR PAPER!

Summary: Conformal prediction is a flexible and powerful tool for UQ.

See RCPS for complicated, high-dimensional Y.