Quantifying the Uncertainty in Model Predictions

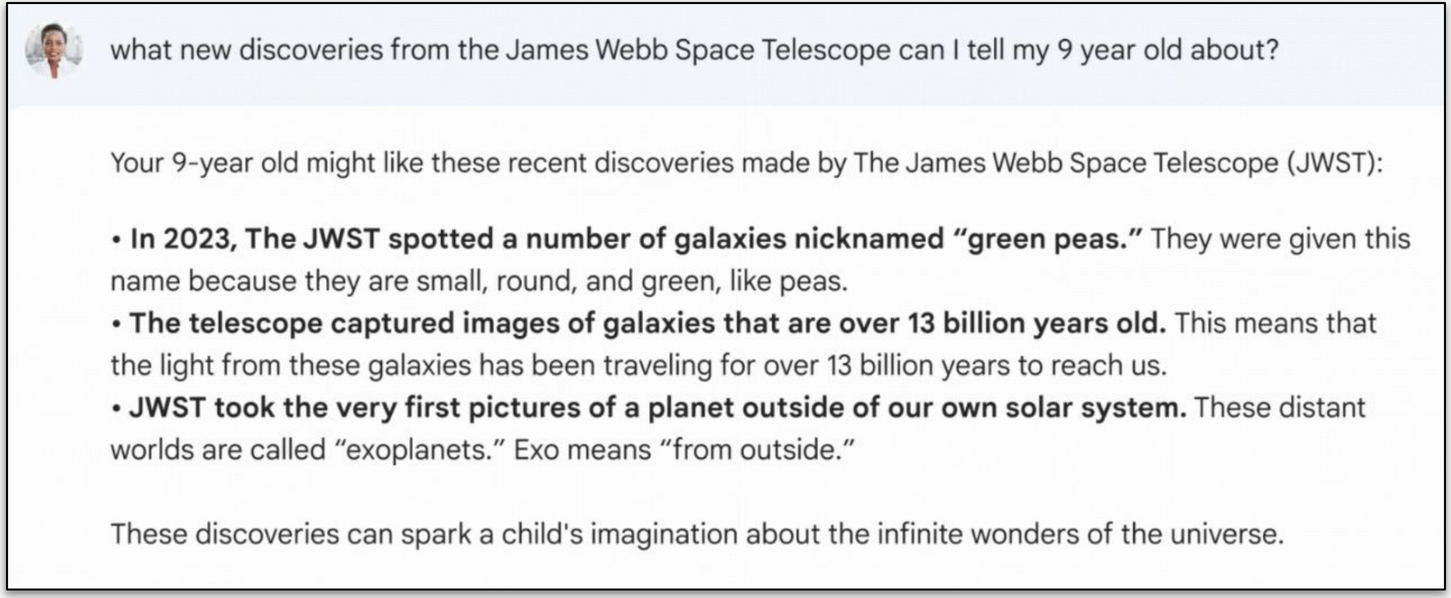
Jesse Cresswell, PhD
Senior Machine Learning Scientist
Layer 6 AI at TD

June 13 2023 - TMLS

Agenda

- Uncertainty Quantification
- Conformal Prediction
- Applications
- Conclusion

Why should we quantify uncertainty?



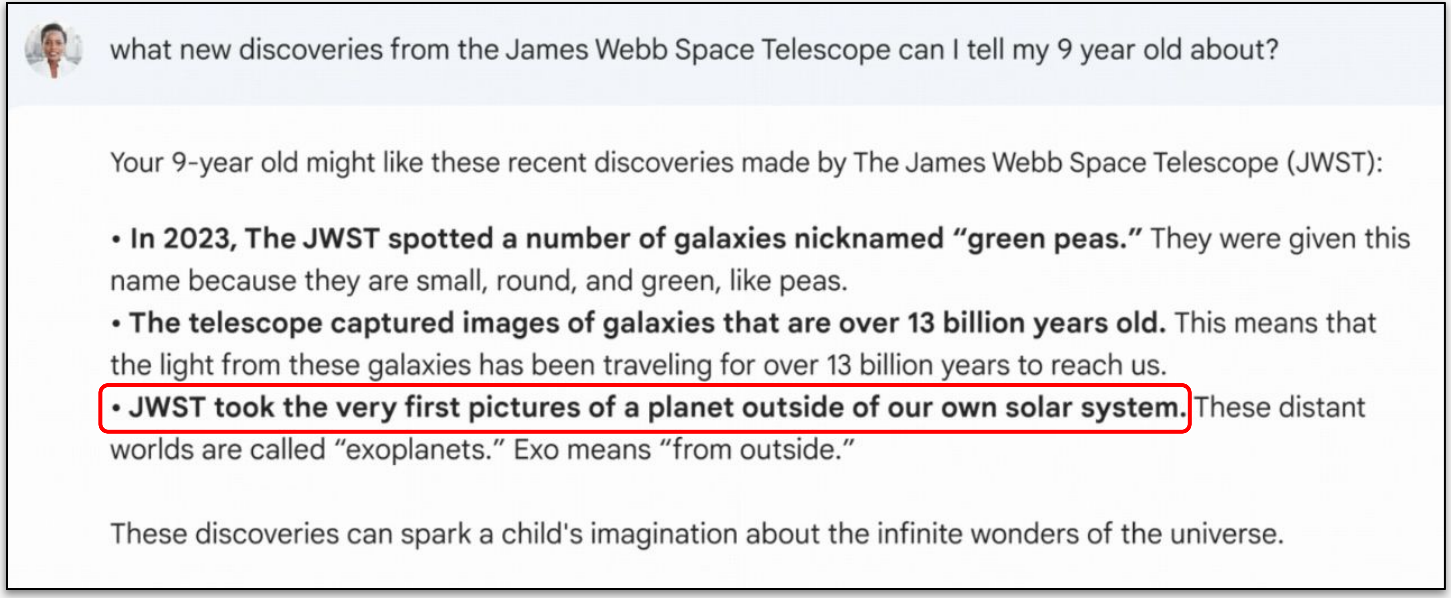
what new discoveries from the James Webb Space Telescope can I tell my 9 year old about?

Your 9-year old might like these recent discoveries made by The James Webb Space Telescope (JWST):

- **In 2023, The JWST spotted a number of galaxies nicknamed "green peas."** They were given this name because they are small, round, and green, like peas.
- **The telescope captured images of galaxies that are over 13 billion years old.** This means that the light from these galaxies has been traveling for over 13 billion years to reach us.
- **JWST took the very first pictures of a planet outside of our own solar system.** These distant worlds are called "exoplanets." Exo means "from outside."

These discoveries can spark a child's imagination about the infinite wonders of the universe.

Why should we quantify uncertainty?



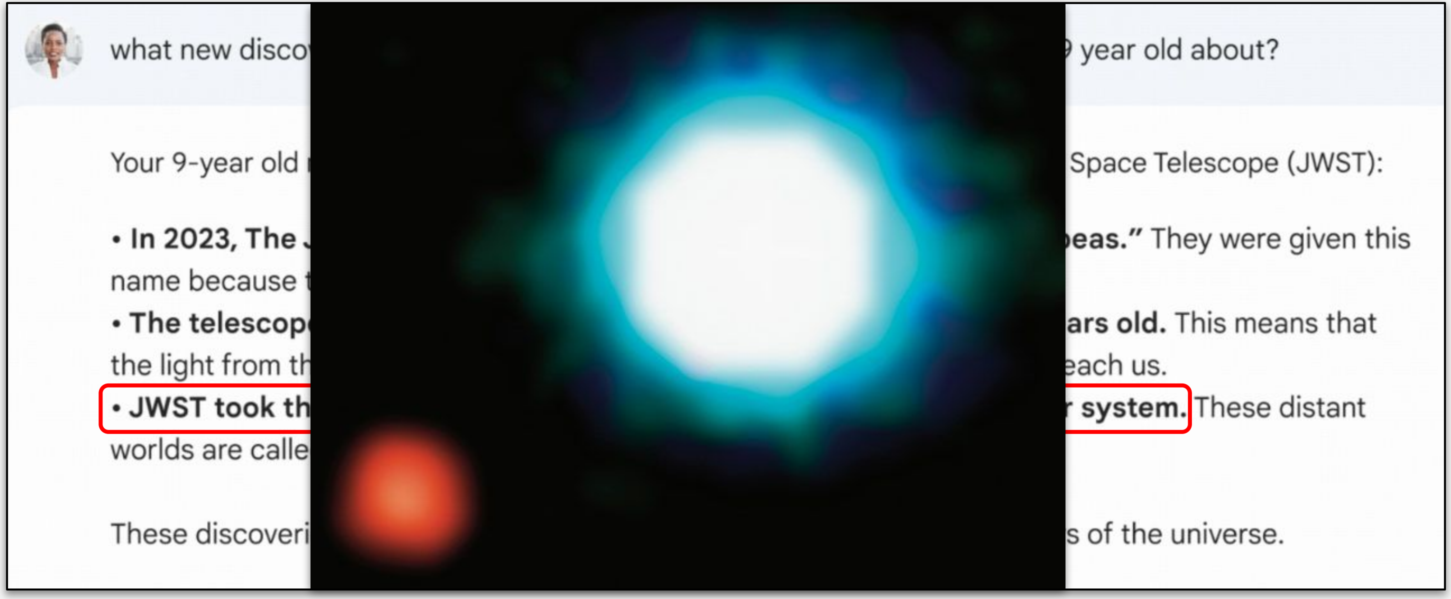
what new discoveries from the James Webb Space Telescope can I tell my 9 year old about?

Your 9-year old might like these recent discoveries made by The James Webb Space Telescope (JWST):

- **In 2023, The JWST spotted a number of galaxies nicknamed "green peas."** They were given this name because they are small, round, and green, like peas.
- **The telescope captured images of galaxies that are over 13 billion years old.** This means that the light from these galaxies has been traveling for over 13 billion years to reach us.
- **JWST took the very first pictures of a planet outside of our own solar system.** These distant worlds are called "exoplanets." Exo means "from outside."

These discoveries can spark a child's imagination about the infinite wonders of the universe.

Why should we quantify uncertainty?



what new disco 9 year old about?

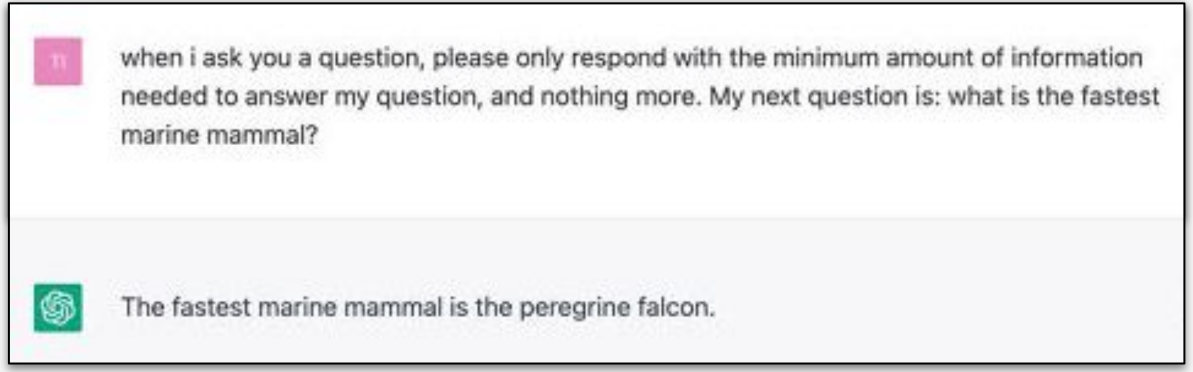
Your 9-year old Space Telescope (JWST):

- **In 2023, The J** **eas."** They were given this name because t
- **The telescop** **ars old.** This means that the light from th each us.
- **JWST took th** **r system.** These distant worlds are calle

These discoveri s of the universe.

Actual first image of an exoplanet, 2M1207-b, captured by the Very Large Telescope (European Southern Observatory)

Why should we quantify uncertainty?



when i ask you a question, please only respond with the minimum amount of information needed to answer my question, and nothing more. My next question is: what is the fastest marine mammal?

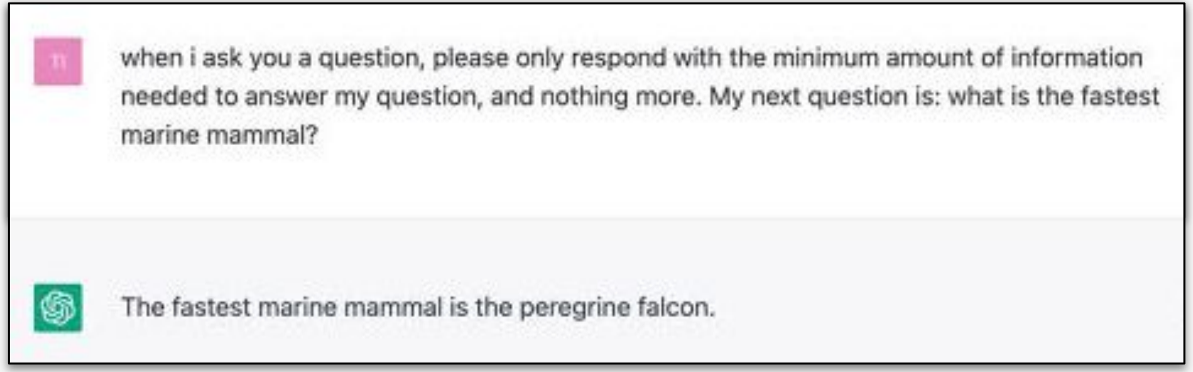
The fastest marine mammal is the peregrine falcon.

Luckily, we have a snappy name for wrong answers now: **Hallucinations**.

Usually, they're pretty good though!

The issue is, these models don't indicate how confident they are.

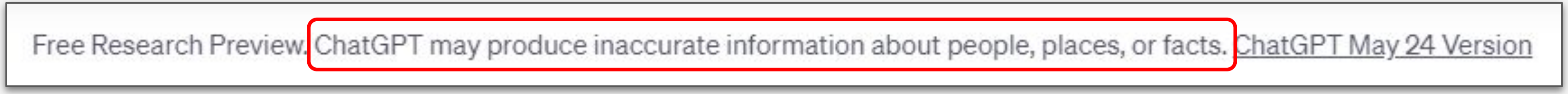
Why should we quantify uncertainty?



when i ask you a question, please only respond with the minimum amount of information needed to answer my question, and nothing more. My next question is: what is the fastest marine mammal?

The fastest marine mammal is the peregrine falcon.

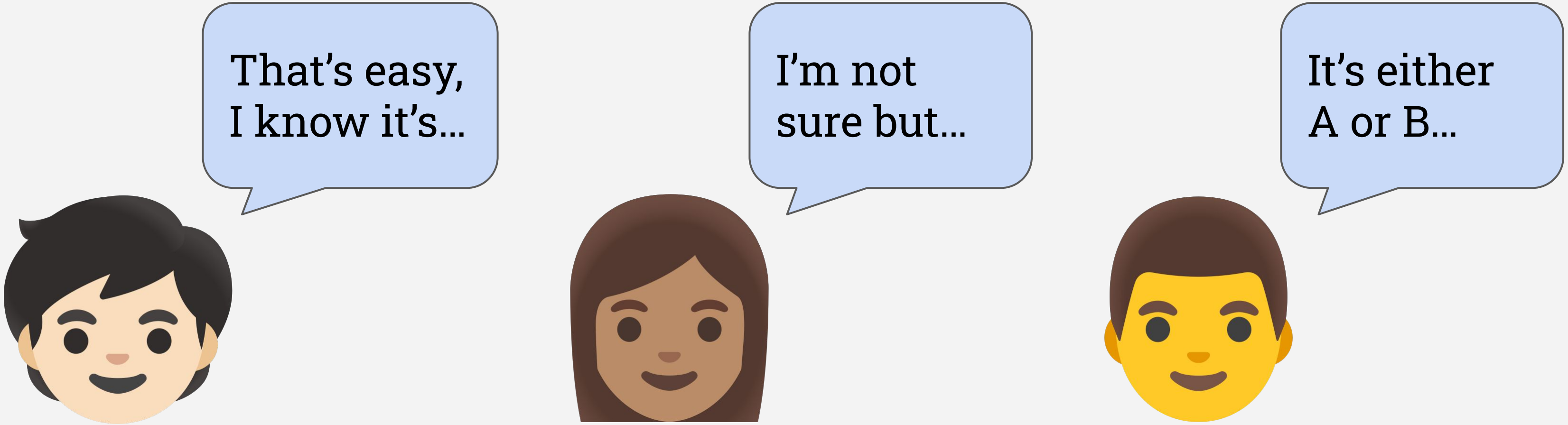
Luckily, we have a snappy name for wrong answers now: **Hallucinations**.
Usually, they're pretty good though!
The issue is, these models don't indicate how confident they are.



Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT May 24 Version

Why should we quantify uncertainty?

When humans answer questions, we naturally state how confident we are.
It’s a crucial aspect of decision making.



We signal unconfidence, and offer alternatives.

Why should we quantify uncertainty?

Real-world decisions involve uncertainty

Regression example: House price prediction

Listed for: $1,189,000

We use **AI** to estimate home value

Estimated value: **$1,376,595**

Quantifying uncertainty allows us to make confident decisions.

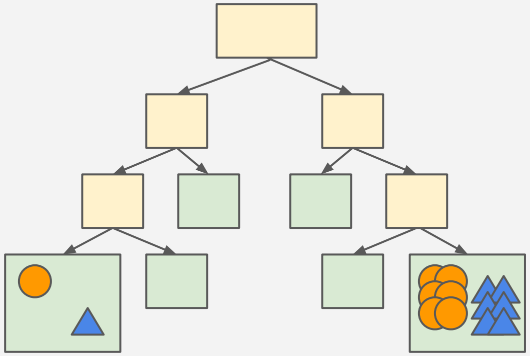
There is a 95% chance the sale price will be **$1,260,000 - $1,480,000**.

There is a 95% chance the sale price will be **$1,340,000 - $1,390,000**.

Why should we quantify uncertainty?

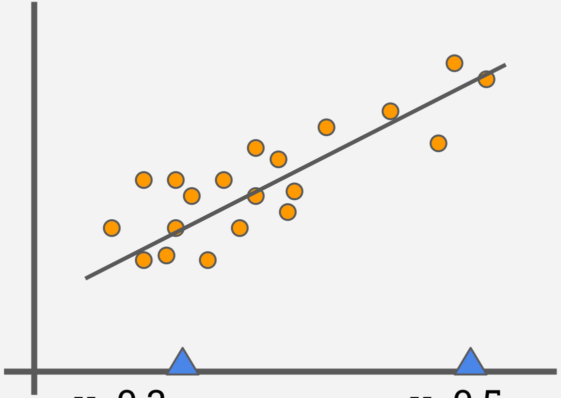
Models are trained on limited data.

By quantifying uncertainty across regimes, we can evaluate when the predictions are robust and stable, or identify where more data is needed.



y=0.5
Low Confidence

y=0.5
High Confidence



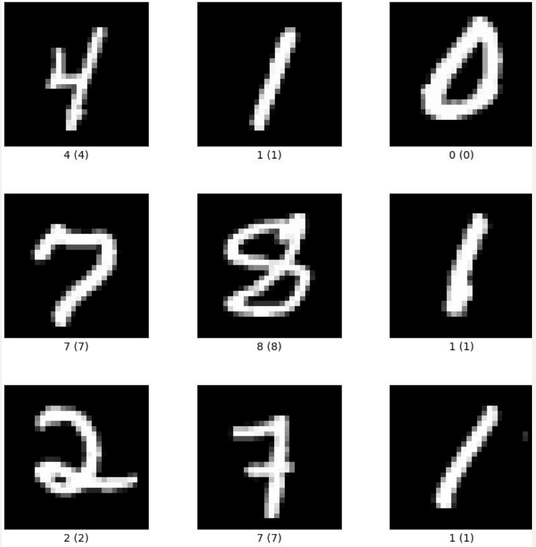
y=0.3
High Confidence

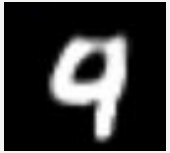
y=0.5
Low Confidence

Why should we quantify uncertainty?

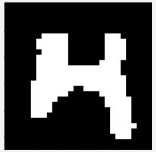
Most models interpolate, and may perform poorly on out-of-distribution data.
Models should indicate low confidence on data they have not seen before.

In distribution ⟷ Out of distribution

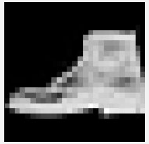




4 or 9?
Is a digit.



4 or H?
Might be
a digit.



Not a digit
at all.

Quantifying uncertainty

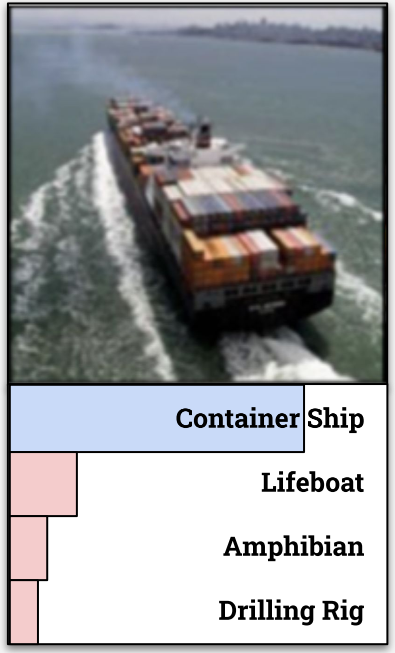
Classifiers have an inbuilt notion of uncertainty.

Normally, a classifier outputs a number between 0 and 1 for each of *k* classes, with all outputs summing to one.

$$y_i = f(x_i)_k \in [0, 1],\ \sum_k f(x)_k = 1$$

A **higher raw score** for class *k* indicates **more confidence** in that prediction.

Is a classifier's score a **good** notion of uncertainty?



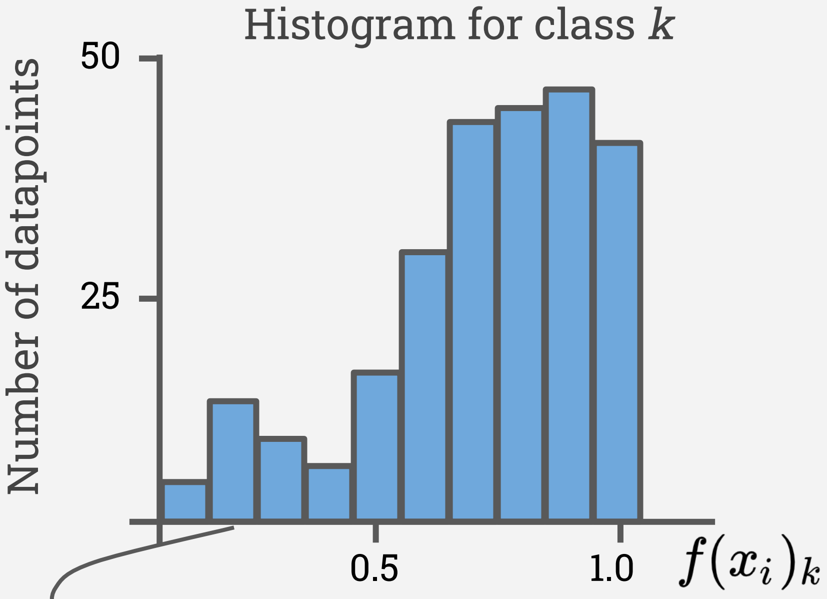
Calibration

We expect a raw score of 0.2 to mean “There is an 20% percent chance that the correct class is *k*.”

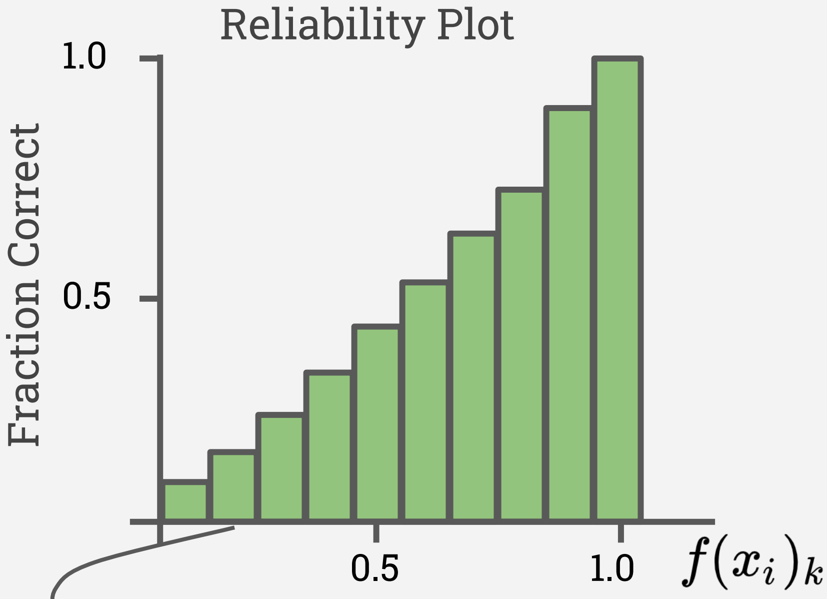
But for any given input, there should only be one correct answer. Hence, the 20% chance must be in reference to a group of inputs.

“If we consider all inputs with raw score of 0.2, then the predicted class should be correct 20% of the time.”

Calibration



Take all points with predictions close to 0.2 and **count**.



Take all points with predictions close to 0.2, and **compute fraction correct**.

layer 6
AI at TD

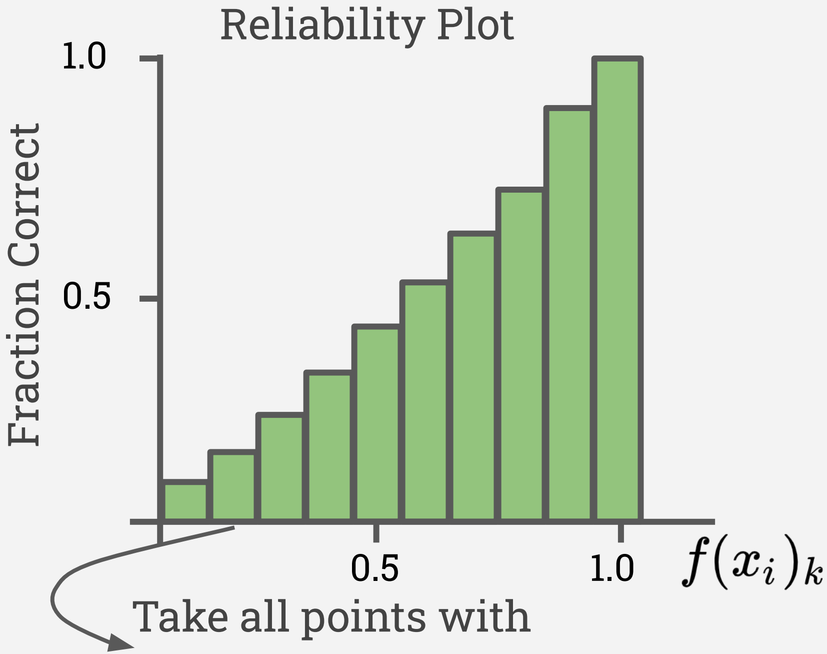
Calibration

“If we consider all inputs with raw score of 0.2, then the predicted class should be correct 20% of the time.”

This should be true for all ranges of raw scores.

$$P(y_i = k|x_i) = f(x_i)_k$$

When a model is well-calibrated, it’s predictions are more meaningful.



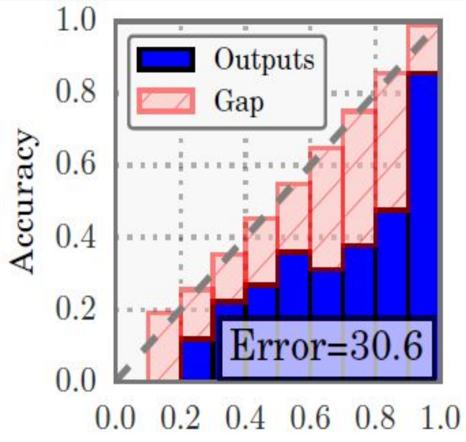
Take all points with predictions close to 0.2, and **compute fraction correct**.

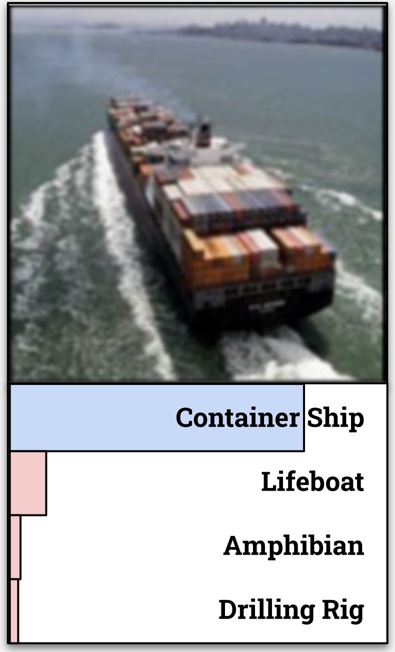
layer 6
AI at TD

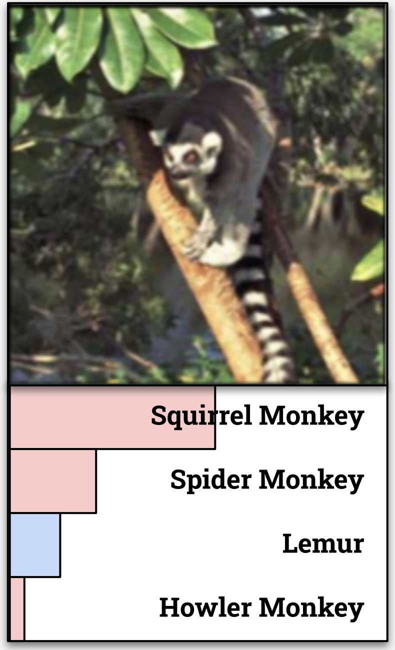
Confident, but Wrong

Some models typically have fair or good calibration (linear regression, decision trees).

However, deep neural networks have poor calibration, and tend to be very overconfident when they are wrong.







Conformal Prediction

Conformal Prediction

Conformal prediction is a general purpose method for transforming heuristic notions of uncertainty into rigorous ones.

Instead of outputting a single prediction, conformal prediction returns a **set**.

Conformal Prediction

Conformal prediction is a general purpose method for transforming heuristic notions of uncertainty into rigorous ones.

Instead of outputting a single prediction, conformal prediction returns a **set**.

{Container Ship}

Conformal Prediction

Conformal prediction is a general purpose method for transforming heuristic notions of uncertainty into rigorous ones.

Instead of outputting a single prediction, conformal prediction returns a **set**.

{Container Ship}

{Squirrel Monkey, Spider Monkey, Lemur}

Conformal Prediction

{Container Ship}

**{Squirrel Monkey,
Spider Monkey, Lemur}**

The **size** of a prediction set **quantifies how uncertain** the model is.

When the model is more uncertain, the prediction set is larger, and this **provides alternatives** to the point prediction.

Conformal Prediction

Conformal prediction provides a statistical guarantee:

“The correct answer is in the prediction set with probability at least 1-α.”

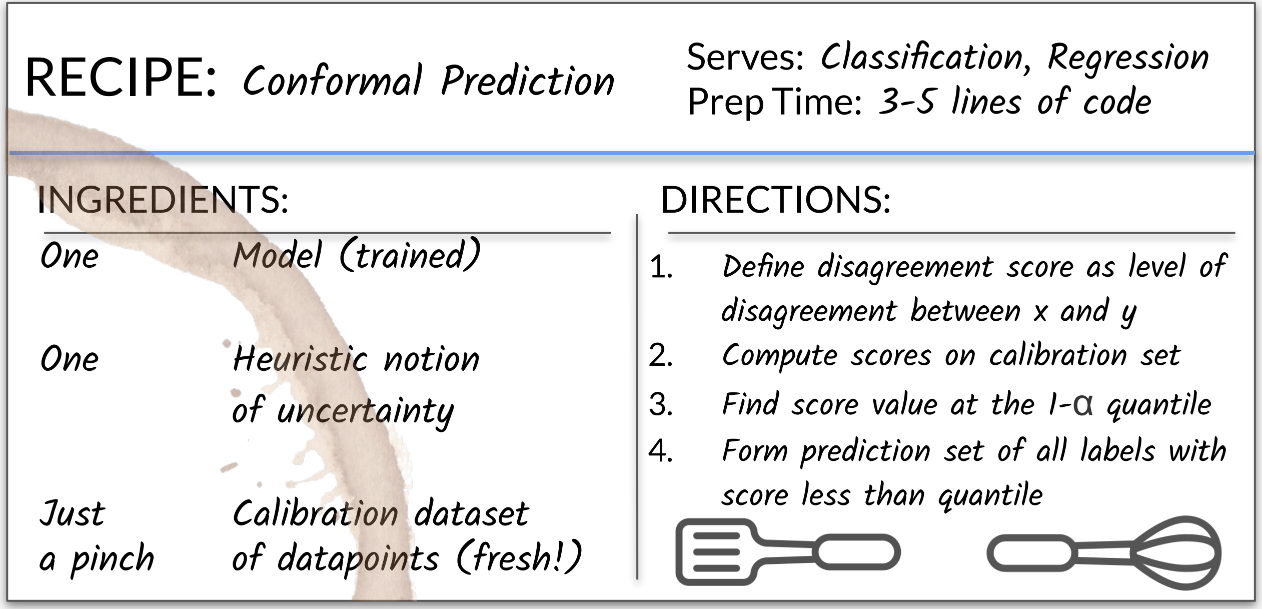
$$P(y_i \in S(x_i)) \geq 1 - \alpha$$

1-α can be thought of as the **success rate -** we choose it based on our error tolerance. Mistakes are reduced by making prediction sets larger (and therefore less useful).

Along with **statistical rigour**, conformal prediction is **versatile**, and **simple** to apply.

The Conformal Recipe

Before looking at what to do with a prediction set, let's discuss how to construct one.



RECIPE: *Conformal Prediction*

Serves: *Classification, Regression*
Prep Time: *3-5 lines of code*

INGREDIENTS:

One	*Model (trained)*
One	*Heuristic notion of uncertainty*
Just a pinch	*Calibration dataset of datapoints (fresh!)*

DIRECTIONS:

1. *Define disagreement score as level of disagreement between x and y*
2. *Compute scores on calibration set*
3. *Find score value at the 1-α quantile*
4. *Form prediction set of all labels with score less than quantile*

The Conformal Recipe

1. From the heuristic uncertainty measure, define the **disagreement score** $d(x, y)$. Larger scores mean worse agreement between *x* and *y*.

Classification:

The score could simply be the model's output for the correct class

$$d(x, y) = 1 - f(x)_y$$

Larger scores mean the model thought the correct class was unlikely.

Regression:

The Conformal Recipe

1. From the heuristic uncertainty measure, define the **disagreement score** $d(x, y)$. Larger scores mean worse agreement between *x* and *y*.

Classification:

The score could simply be the model's output for the correct class

$$d(x, y) = 1 - f(x)_y$$

Larger scores mean the model thought the correct class was unlikely.

Regression:

The score could be difference between the model prediction and true value

$$d(x, y) = |y - f(x)|$$

Larger scores mean the model prediction was far from the true value.

The Conformal Recipe

1. From the heuristic uncertainty measure, define the **disagreement score** $d(x, y)$. Larger scores mean worse agreement between *x* and *y*.

Classification:

The score could simply be the model's output for the correct class

$$d(x, y) = 1 - f(x)_y$$

Larger scores mean the model thought the correct class was unlikely.

Regression:

The score could be difference between the model prediction and true value

$$d(x, y) = |y - f(x)|$$

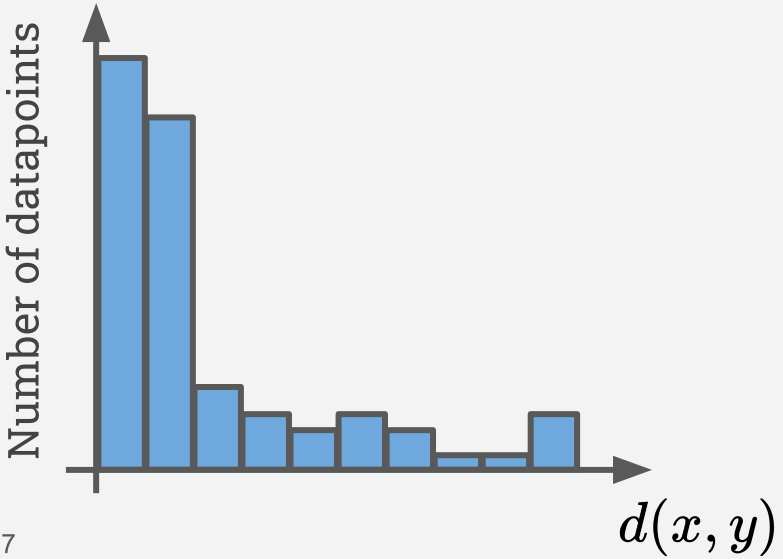
Larger scores mean the model prediction was far from the true value.

In fact any score works! The art is that some are more useful than others.

The Conformal Recipe

2. Compute the disagreement scores on the calibration dataset.
3. Take the ~$1 - \alpha$ quantile of scores.

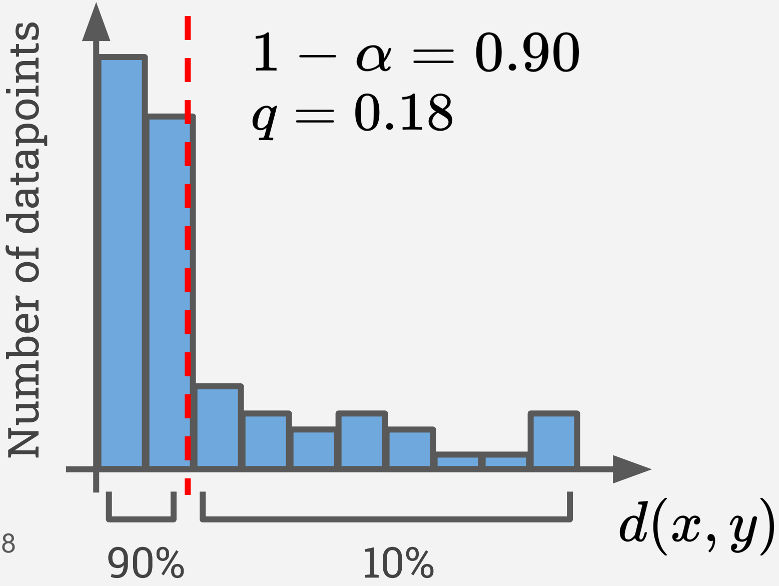
Histogram on calibration data



The Conformal Recipe

2. Compute the disagreement scores on the calibration dataset.
3. Take the ~$1 - \alpha$ quantile of scores.

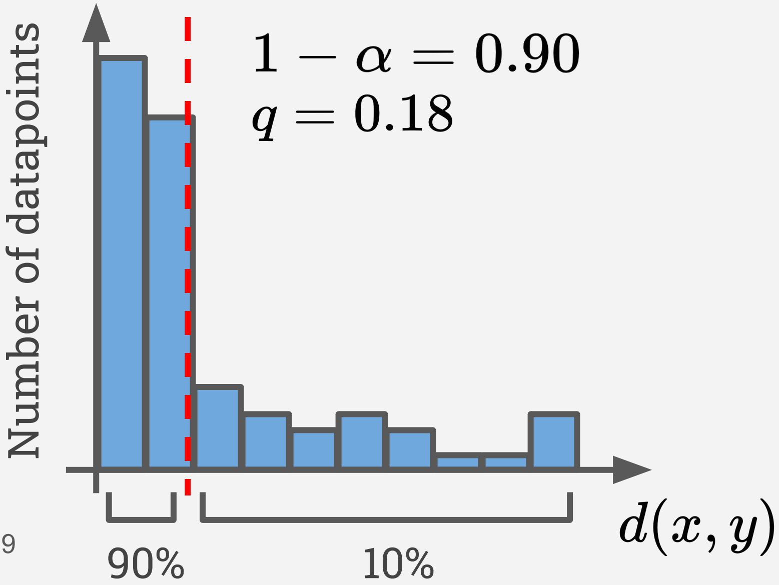
Histogram on calibration data



The Conformal Recipe

2. Compute the disagreement scores on the calibration dataset.
3. Take the ~$1 - \alpha$ quantile of scores.

Histogram on calibration data



If we were to take a new datapoint, there would be a 90% chance that it's disagreement score would be less than the threshold *q*.

layer 6
AI at TD

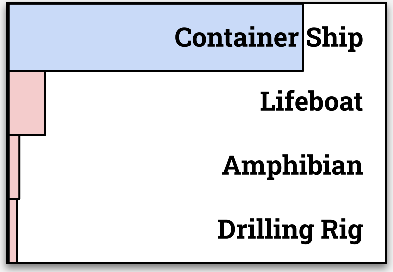
The Conformal Recipe

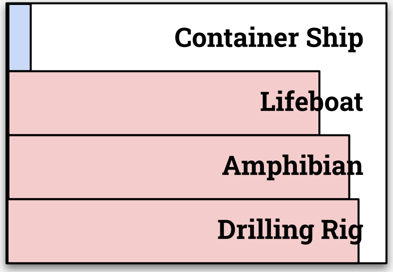
4. Form prediction sets containing all values y that have disagreement less than q.

$$S(x) = \{y : d(x, y) \leq q\}$$

Classification:

$f(x)_k$ $\qquad$ $d(x,k) = 1 - f(x)_k$





Regression:

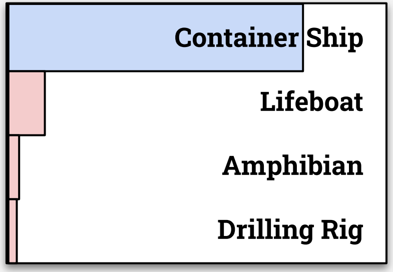
The Conformal Recipe

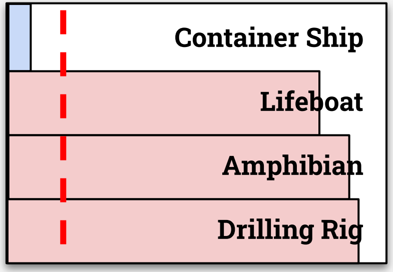
4. Form prediction sets containing all values y that have disagreement less than q.

$$S(x) = \{y : d(x,y) \leq q\}$$

Classification:

$f(x)_k$ $\qquad$ $d(x,k) = 1 - f(x)_k$





$$S(x) = \{\text{Container Ship}\}$$

Regression:

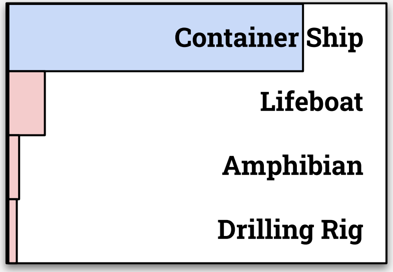
The Conformal Recipe

4. Form prediction sets containing all values y that have disagreement less than q.

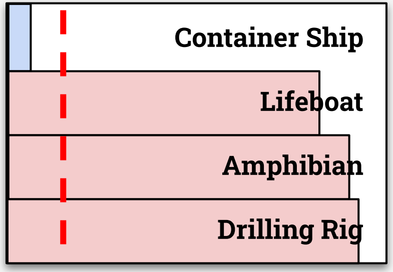
$$S(x) = \{y : d(x,y) \leq q\}$$

Classification:

$f(x)_k$



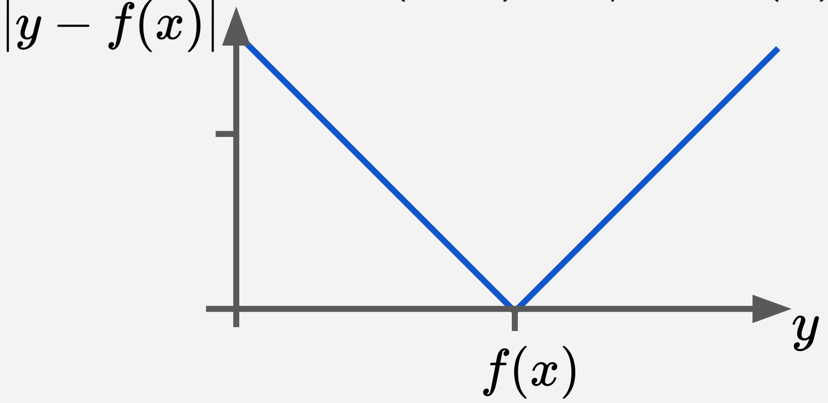
$d(x,k) = 1 - f(x)_k$



$S(x) = \{\text{Container Ship}\}$

Regression:

$d(x,y) = |y - f(x)|$



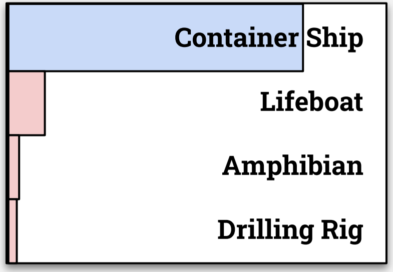
The Conformal Recipe

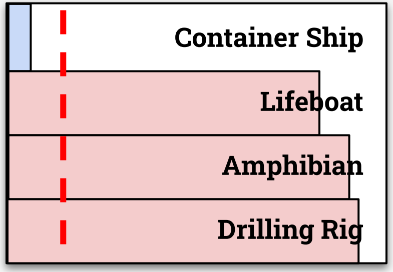
4. Form prediction sets containing all values y that have disagreement less than q.

$$S(x) = \{y : d(x,y) \leq q\}$$

Classification:

$f(x)_k$ $\qquad$ $d(x,k) = 1 - f(x)_k$

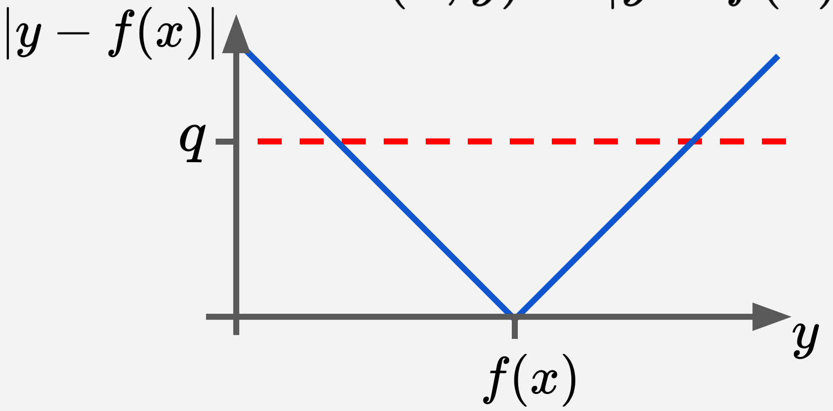




$$S(x) = \{\text{Container Ship}\}$$

Regression:

$$d(x,y) = |y - f(x)|$$



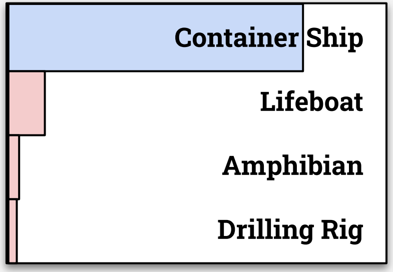
The Conformal Recipe

4. Form prediction sets containing all values y that have disagreement less than q.

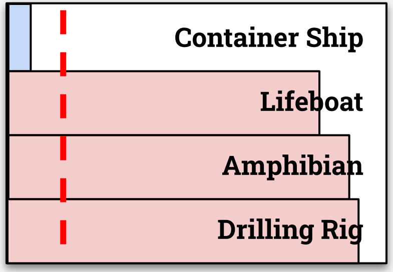
$$S(x) = \{y : d(x,y) \le q\}$$

Classification:

$f(x)_k$



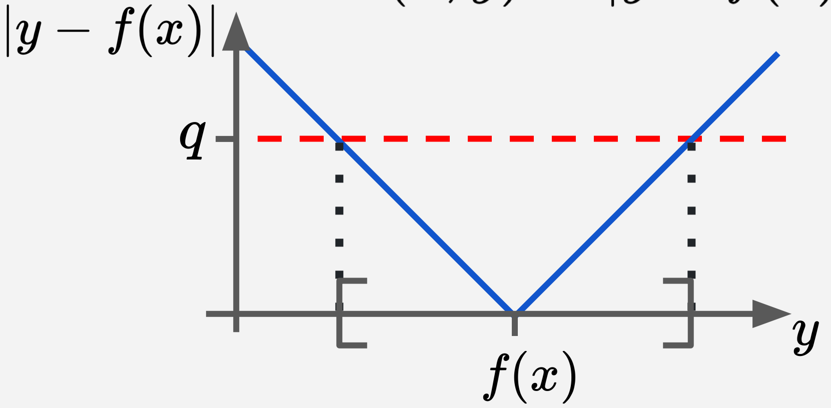
$d(x,k) = 1 - f(x)_k$



$$S(x) = \{\text{Container Ship}\}$$

Regression:

$$d(x,y) = |y - f(x)|$$



$$S(x) = [f(x) - q, f(x) + q]$$

The Conformal Recipe - Recap

1. Define disagreement score as level of disagreement between *x* and *y*
2. Compute disagreement scores on calibration set
3. Find score value at the ~1-**α** quantile
4. Form prediction set of all labels with score less than quantile threshold

$$S(x) = \{y : d(x, y) \leq q\}$$

For a new datapoint, we are guaranteed that the correct answer is in the prediction set with probability at least 1-**α**.

$$P(y_i \in S(x_i)) \geq 1 - \alpha$$

This works for any disagreement score.

Applications

What to do with a set?

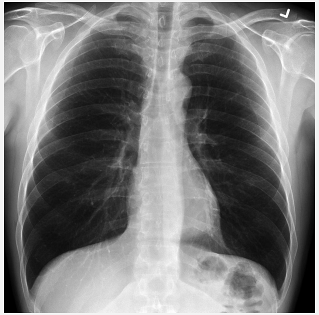
Achieving statistical guarantees is great, but what can we do with a prediction set?

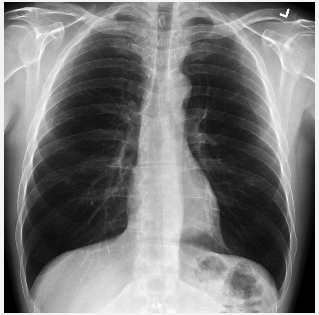
Unlike a point estimate, a set of possible decisions is not a decision. We need a further strategy to convert a set into a decision.

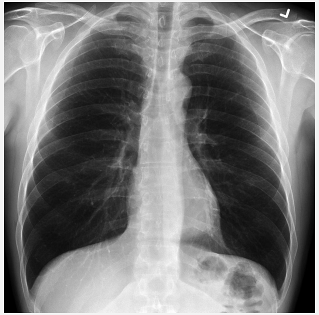
Prediction sets are perfect for human-in-the-loop decision pipelines:

- Fully in-the-loop - Prediction set informs person who makes decision.
- Partially in-the-loop - Decision deferred to person when model is uncertain.

Fully in-the-loop







Model information:

None

Top-1 Class
Pulmonary Haemorrhage

Prediction Set
{Haemorrhage, Edema}
with 95% confidence

A doctor tries to diagnose a patient with breathing difficulty. They order a chest X-ray.

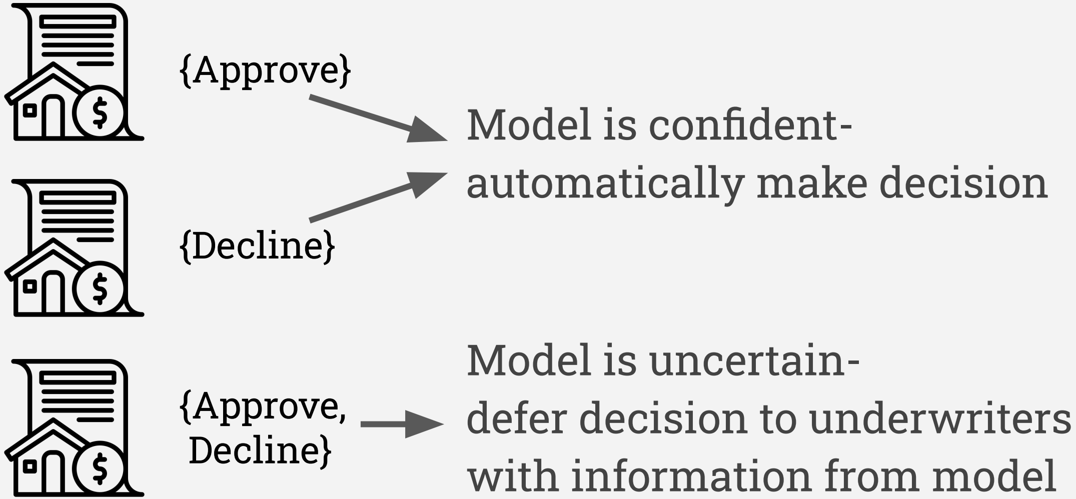
Prediction sets narrow the focus of the investigation, and speed up decision making.
A set is more useful than a prediction, especially if multiple conditions are present.

The doctor makes the diagnosis among all possibilities, and assigns likelihoods.

Partially in-the-loop

Some tasks vary in difficulty, where simpler instances can be fully automated with little risk, but trickier problems are best left to humans.

Mortgage underwriting



{Approve}

{Decline}

Model is confident- automatically make decision

{Approve, Decline}

Model is uncertain- defer decision to underwriters with information from model

Through the conformal guarantee we can tune the acceptable error rate.

Conclusions

Conclusions

Quantifying uncertainty is crucial for trustworthy AI applications, whether we use them in automated decisioning with human impact, or are searching for information through a chatbot.

Modern ML methods lack uncertainty quantification, and are poorly calibrated.

Conformal prediction converts any base model into one with statistical guarantees by creating prediction sets.

Prediction set size quantifies uncertainty and offers alternatives to a point estimate.

Prediction sets can be integrated in ML pipelines where humans are partially, or fully in-the-loop.

Thank you!

jesse@layer6.ai

jesse.cresswell@td.com