Massachusetts Institute of Technology

# Uncertainty

Stefanie Jegelka
MIT

# Outline

- Conformal Prediction
- Bayesian Models

# Neural networks give confidence scores…

- Recall: sigmoid and softmax convert scores/preactivations to “probabilities”…

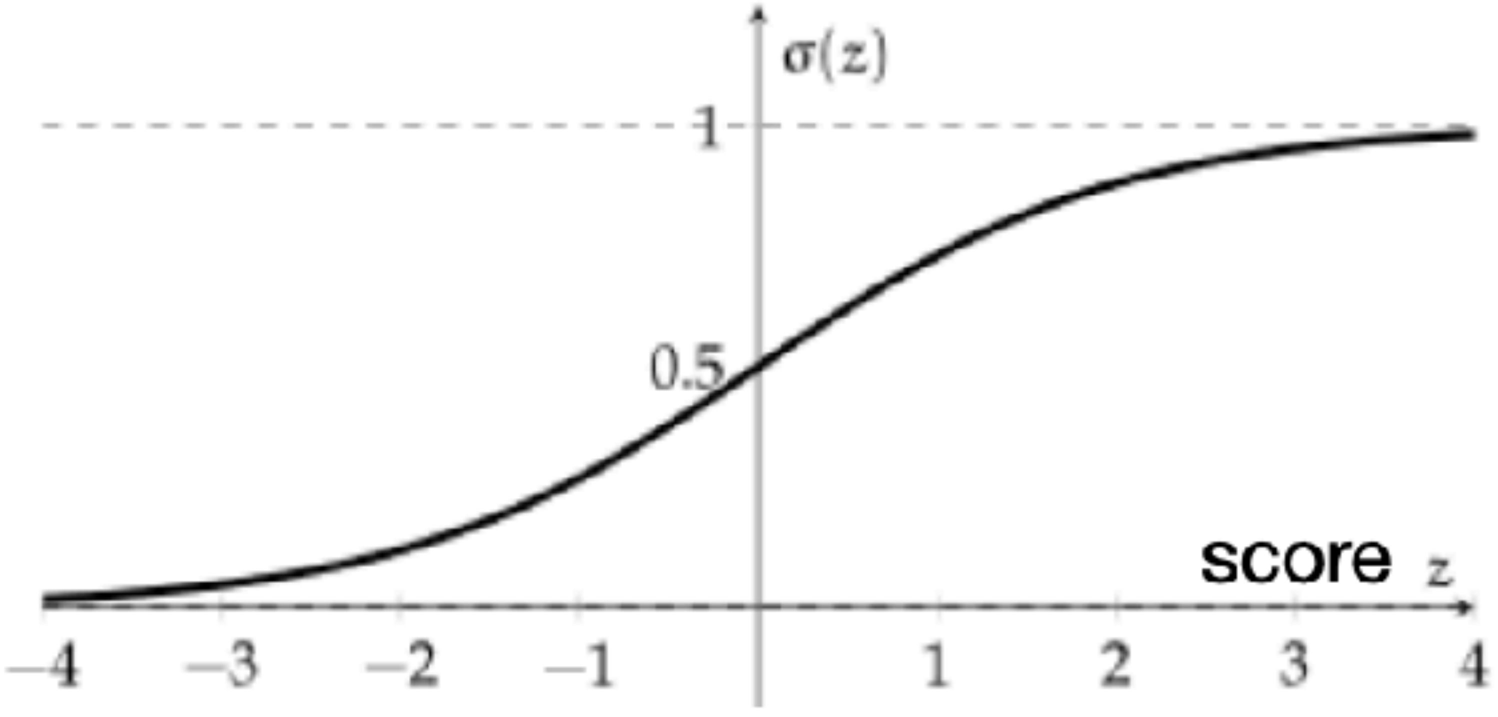
- Use probabilities of +/-1 instead of hard threshold

$$f(x;\theta) = \sigma(\theta \cdot x + \theta_0)$$

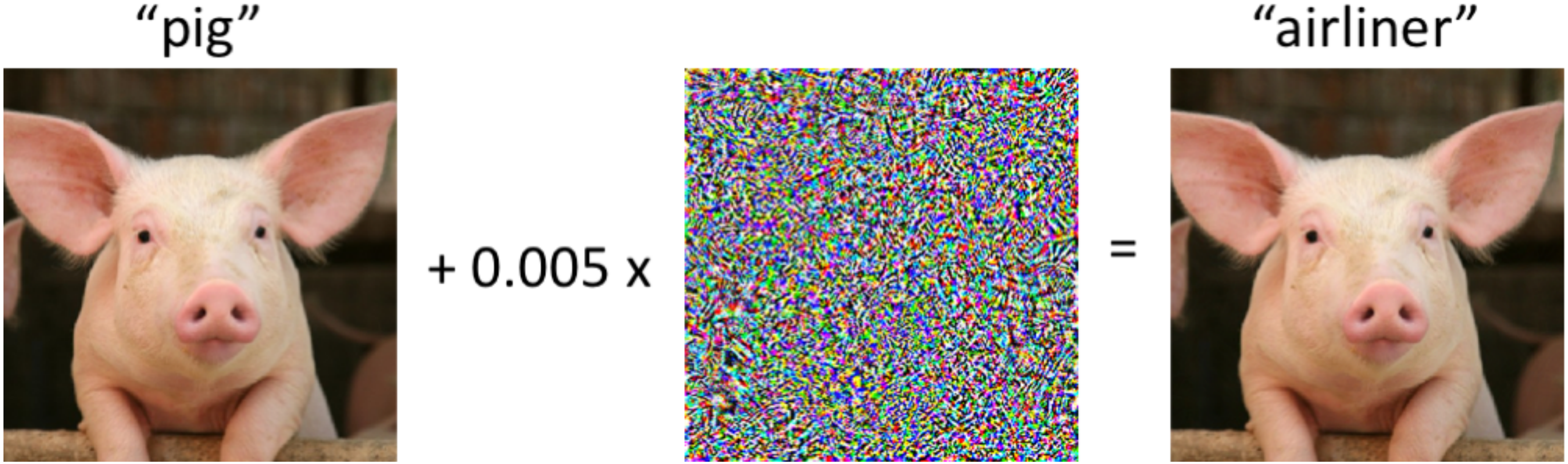
Measures distance from hyperplane

- Sigmoid function transforms score into probability

$$\sigma(z) = \frac{1}{1+e^{-z}}$$



# Are these confidences accurate?




**91% confidence** **99% confidence**

# Are these confidences accurate?

- What does it mean to be “accurate”?

# Are these confidences accurate?

- What does it mean to be “accurate”?
- Well-calibrated: On the points where it says 70%, it should be 70% correct.

# Are these confidences accurate?

- What does it mean to be “accurate”?
- Well-calibrated: On the points where it says 70%, it should be 70% correct.
- Or: output a set of predictions, such that with probability $1 - \epsilon$, the true label is in the set

# Are these confidences accurate?

- What does it mean to be "accurate"?
- Well-calibrated: On the points where it says 70%, it should be 70% correct.
- Or: output a set of predictions, such that with probability $1 - \epsilon$, the true label is in the set

**_Conformal Prediction_ can do this for any ML model.**
**Idea: "re-calibrate" an uncertainty score.**

# Example: classification

- Training data set $S_{train} = \{(x_i, y_i)\}_{i=1}^{n}$
- Calibration data $S_{cali} = \{(x_j, y_j)\}_{j=1}^{m}$ *(25-30% of data, or around 1000)*
- Non-conformity function $f(x, y)$:
  - tells how "unusual" a data point is
  - should be low for true $(x_i, y_i)$, high for wrong labels $(x_i, y \neq y_i)$
  - e.g., $f(x, y) = 1 - \hat{P}_h(y \mid x)$

**output of logistic classifier (classifier "confidence")**

# Non-conformity

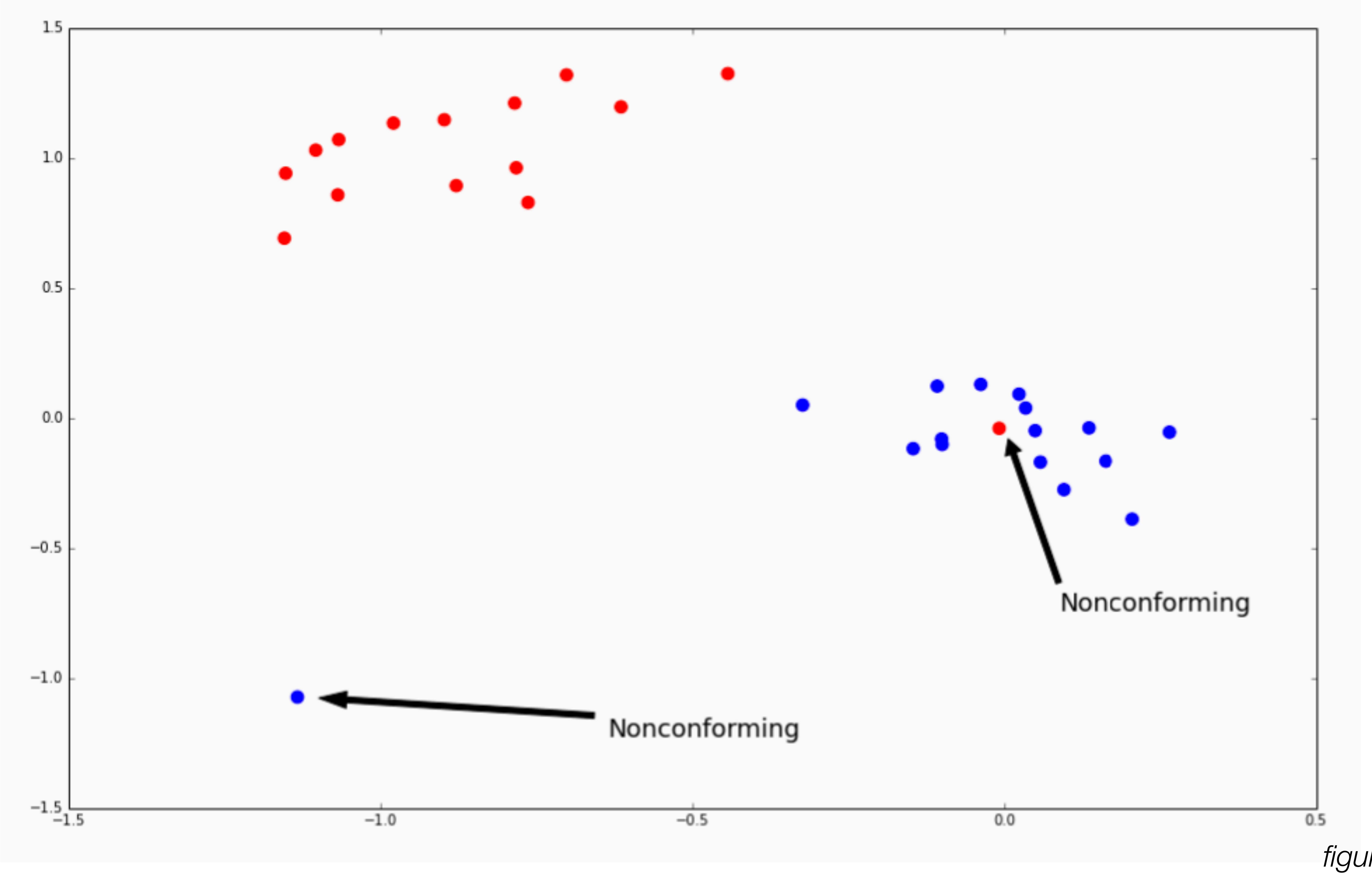



*figure: Henrik Linusson*

# Example: classification

- Training data set $S_{train} = \{(x_i, y_i)\}_{i=1}^{n}$
- Calibration data $S_{cali} = \{(x_j, y_j)\}_{j=1}^{m}$ *(25-30% of data, or around 1000)*
- Non-conformity function $f(x, y)$:
  - tells how "unusual" a data point is
  - should be low for true $(x_i, y_i)$, high for wrong labels $(x_i, y \neq y_i)$
  - e.g., $f(x, y) = 1 - \hat{P}_h(y \mid x)$

**output of logistic classifier (classifier "confidence")**

# Calibration

- Look at the distribution of non-conformity scores for true labels on calibration data:
  Compute $s_j = f(x_j, y_j)$ for all points in the calibration set.

- Sort the $s_j$

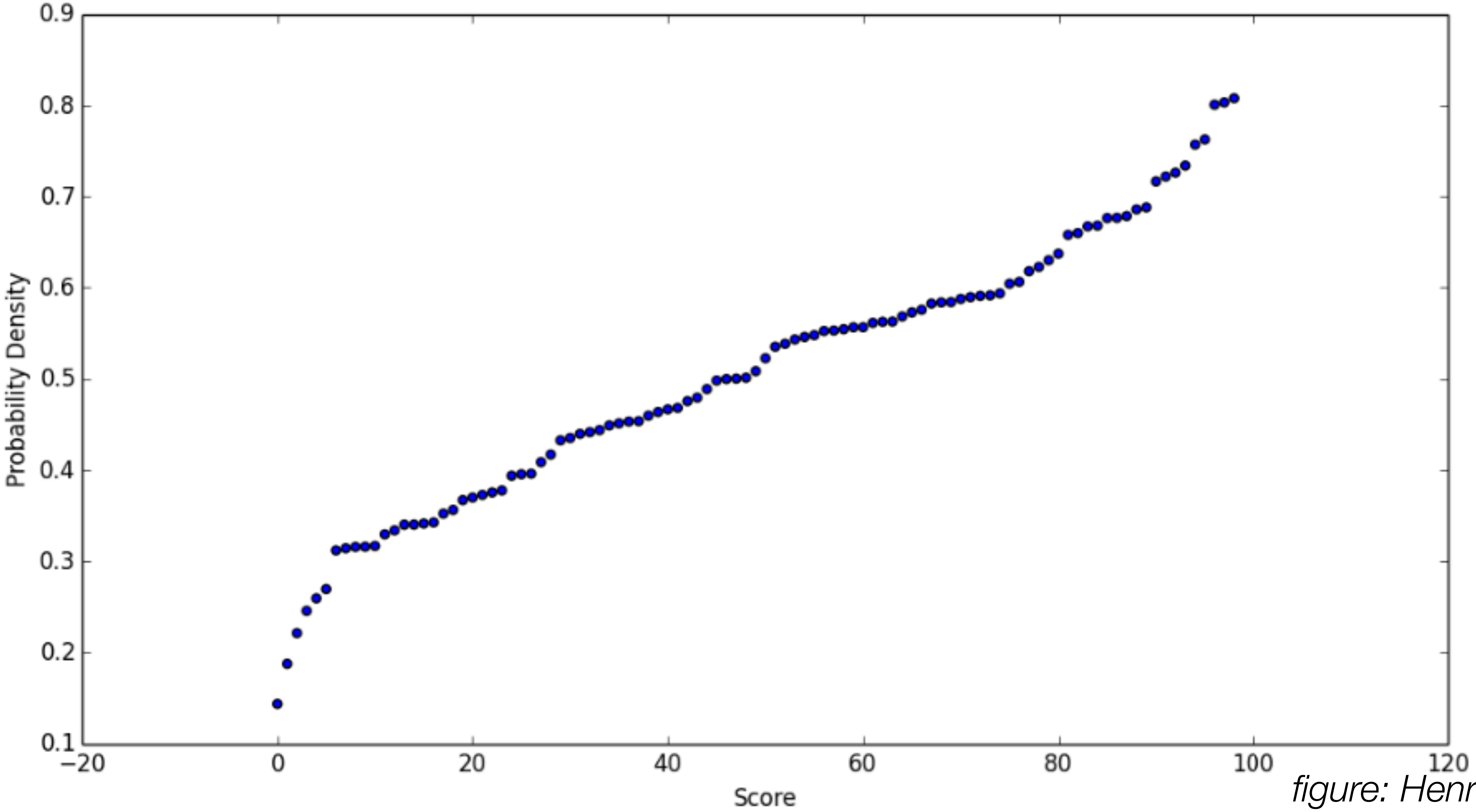
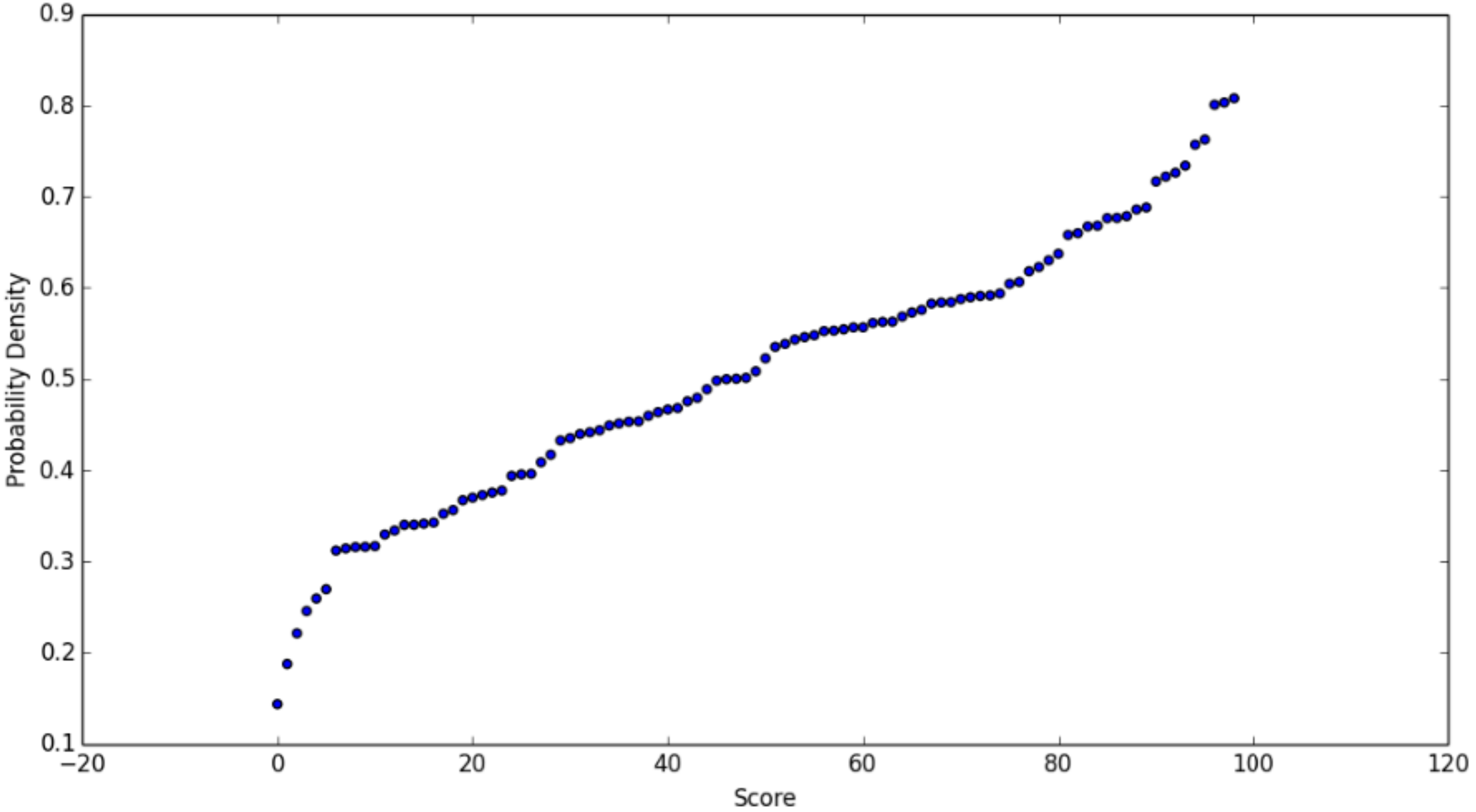



*figure: Henrik Linusson*

# Calibration

- Now for a new prediction, we will see where its score falls in this distribution: is it unlikely?
- For a test point $x_{test}$, compute $f(x_{test}, y') = 1 - \hat{P}(y' | x_{test})$ for all possible labels $y'$
- Prediction set: all labels $y'$ with "low enough" score



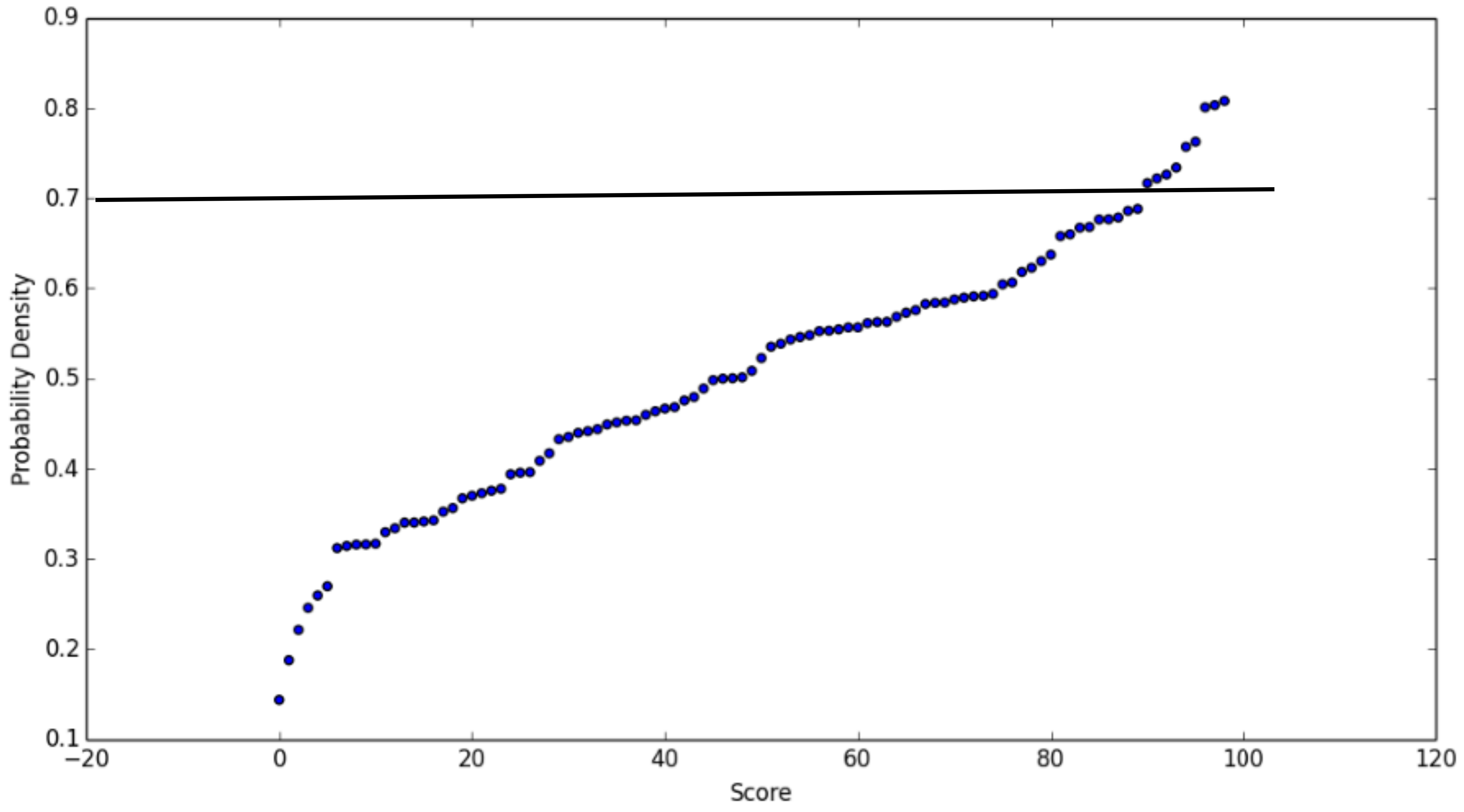
# Predicted set

- All labels $y'$ with "low enough" non-conformity score $f(x_{test}, y')$

$$p(x_{test}, y') = \frac{|\{(x_j, y_j) \in S_{test} : s_j > f(x_{test}, y')\}|}{m+1} + \theta \frac{|\{(x_j, y_j) \in S_{test} : s_j = f(x_{test}, y')\}|}{m+1}$$

**uniform random number between [0,1]**

- Predicted set: all $y'$ with $p(x_{test}, y') > \epsilon$



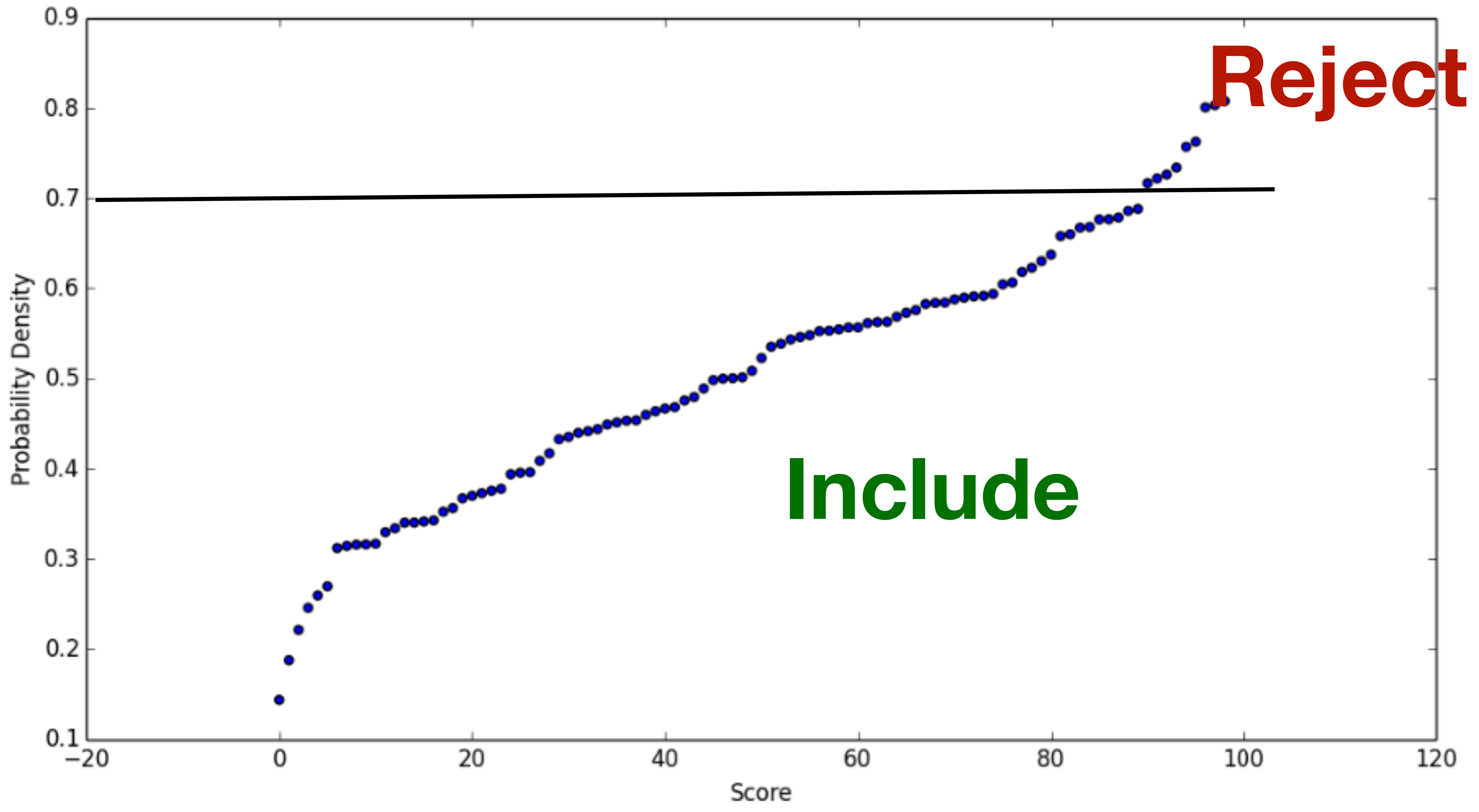
# Predicted set

- All labels $y'$ with "low enough" non-conformity score $f(x_{test}, y')$

$$p(x_{test}, y') = \frac{|\{(x_j, y_j) \in S_{test} : s_j > f(x_{test}, y')\}|}{m+1} + \theta \frac{|\{(x_j, y_j) \in S_{test} : s_j = f(x_{test}, y')\}|}{m+1}$$

**uniform random number between [0,1]**

- Predicted set: all $y'$ with $p(x_{test}, y') > \epsilon$



# Example


***Figure 1: Prediction set examples on Imagenet.*** *We show three progressively more difficult examples of the class* **`fox squirrel`** *and the prediction sets (i.e., $\mathcal{C}(X_{\text{test}})$) generated by conformal prediction.*

- For harder examples, the predicted set will be larger (larger uncertainty)

*figure: Angelopoulos & Bates, 2022*

# General recipe

1. Identify a heuristic notion of uncertainty using the pre-trained model.

2. Define the score function f(x, y). (Larger scores encode worse agreement between x and y.)

3. Compute $\hat{q}$ as the $\frac{\lceil (m+1)(1-\epsilon) \rceil}{m}$ quantile of the calibration scores $s_1 = f(x_1, y_1), \ldots, s_m = f(x_m, y_m)$.

4. Use this quantile to form the prediction sets for new examples: $\mathscr{C}(x_{test}) = \{y : f(x_{test}, y) \leq \hat{q}\}$.

5.

# Conformal regression

- Training data set $S_{train} = \{(x_i, y_i)\}_{i=1}^{n}$
- Calibration data $S_{cali} = \{(x_j, y_j)\}_{j=1}^{m}$ *(25-30% of data, or around 1000)*
- Non-conformity function $f(x, y)$:
  - tells how "unusual" a data point is
  - should be low for true $(x_i, y_i)$, high for wrong labels
  - e.g., $f(x, y) = |y_i - h(x_i)|$

**predicted value of trained model**

- We will predict an interval for a test point

# Conformal Regression

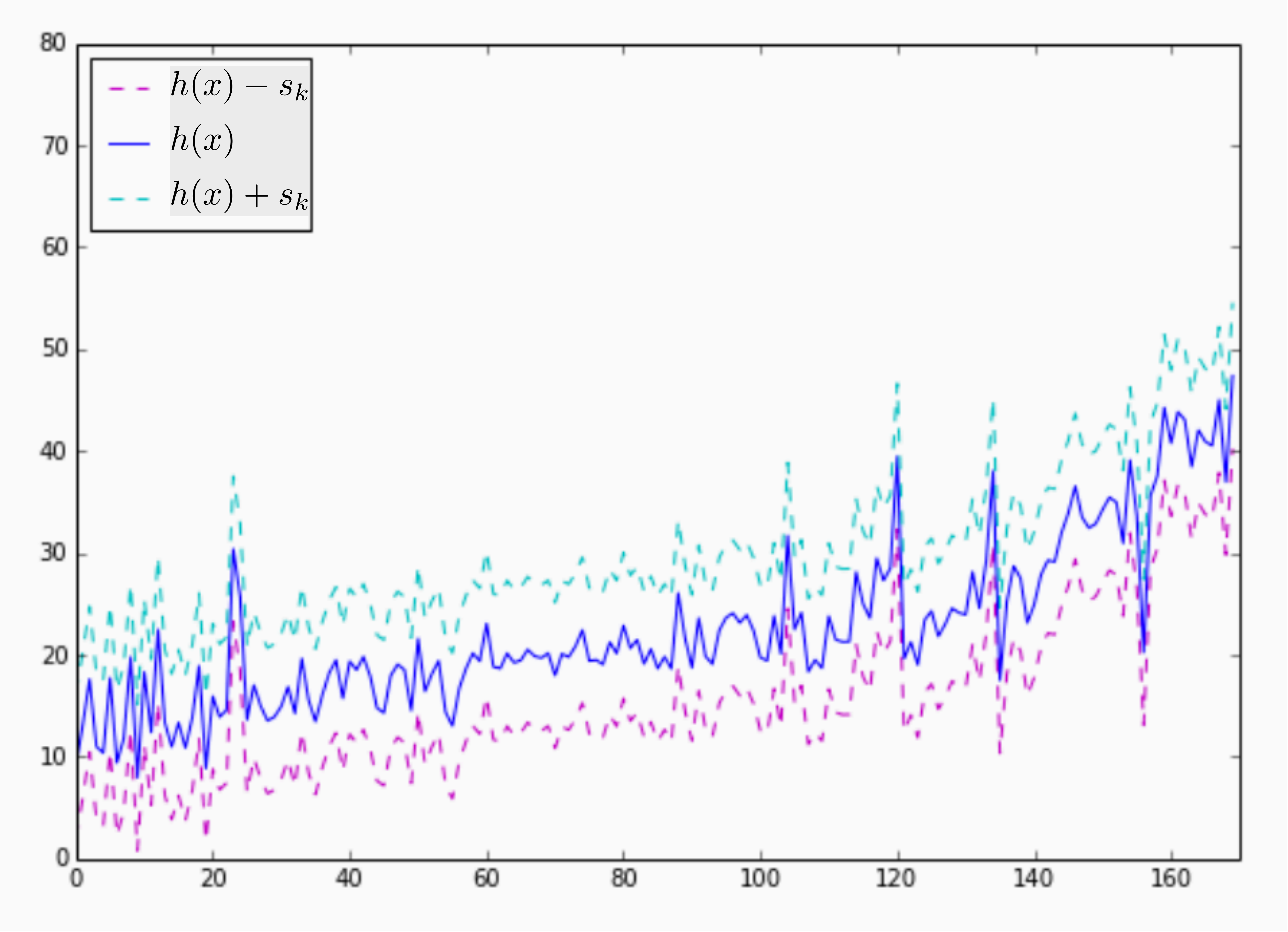
- Non-conformity function: $f(x, y) = |y_i - h(x_i)|$

1. Compute $s_1 = f(x_1, y_1), \ldots, s_m = f(x_m, y_m)$ for points in the calibration set
2. Sort these scores in descending order
3. Get index of $(1 - \epsilon)$-percentile non-conformity score: $k = \lfloor \epsilon(m + 1) \rfloor$
4. Prediction set for $x_{test}$: $\boxed{h(x_{test}) \pm s_k}$

# Conformal Regression

- Non-conformity function: $f(x, y) = |y_i - h(x_i)|$

1. Compute $s_1 = f(x_1, y_1), \ldots, s_m = f(x_m, y_m)$ for points in the calibration set
2. Sort these scores in descending orde
3. Get index of $(1 - \epsilon)$-percentile non-c
4. Prediction set for $x_{test}$: $\boxed{h(x_{test}) \pm s_k}$




**Boston housing dataset, random forest**
*(Henrik Linusson)*

# When can we apply conformal prediction?

- training, calibration, test data come from the same distribution

- data is “exchangeable”: order of observations does not matter (a bit weaker than “statistically independent”)

- Works for any predictive model and any data distribution

# Exchangeable?

**Identically, independently and exchangeably distributed sampling (*iid*)**

- Draw random numbers (with replacement) according to $\mathbb{Z} \sim U[0,3]$
- $P\{1,2,3\} = P\{2,1,3\} = P\{1,1,1\}$

**Identically, non-independently and exchangeably distributed sampling**

- Draw random numbers (without replacement) according to $\mathbb{Z} \sim U[0,3]$
- $P\{1,2,3\} = P\{2,1,3\} \neq P\{1,1,1\}$

**Identically, non-independently and non-exchangeably distributed sampling**

- Draw random numbers (without replacement) according to $\mathbb{Z} \sim U[0,3]$, but skip any number smaller than its predecessor
- $P\{1,2,3\} \neq P\{2,1,3\}$

*slide: Henrik Linusson*

# Many extensions / variations

- conditional conformal prediction: for each test point, true value is in prediction set with probability $1 - \epsilon$ (not only on average)
- group-balanced conformal prediction: equal error rates for different subgroups
- outlier prediction
- covariate shifts: distribution of x changes, but not relationship between x and y (weighted conformal prediction)

*see e.g. Angelopoulos & Bates, A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. https://arxiv.org/abs/2107.07511*

# Outline

- Conformal Prediction
- Bayesian Models

# Probabilistic viewpoint

- Our model M gives us a probability of the data D: $P(D \mid M)$
- Bayesian Inference: we convert this into a probability over models (model parameters)
- *Prior* distribution over models: e.g., regression weights, smoothness of the regression function, sparsity, etc. $P(M)$
- Joint probability of model and data:

$$P(M, D) = P(D \mid M) \cdot P(M) = P(M \mid D) \cdot P(D)$$

- Bayes' rule: $P(M \mid D) = \dfrac{P(D \mid M) P(M)}{P(D)}$

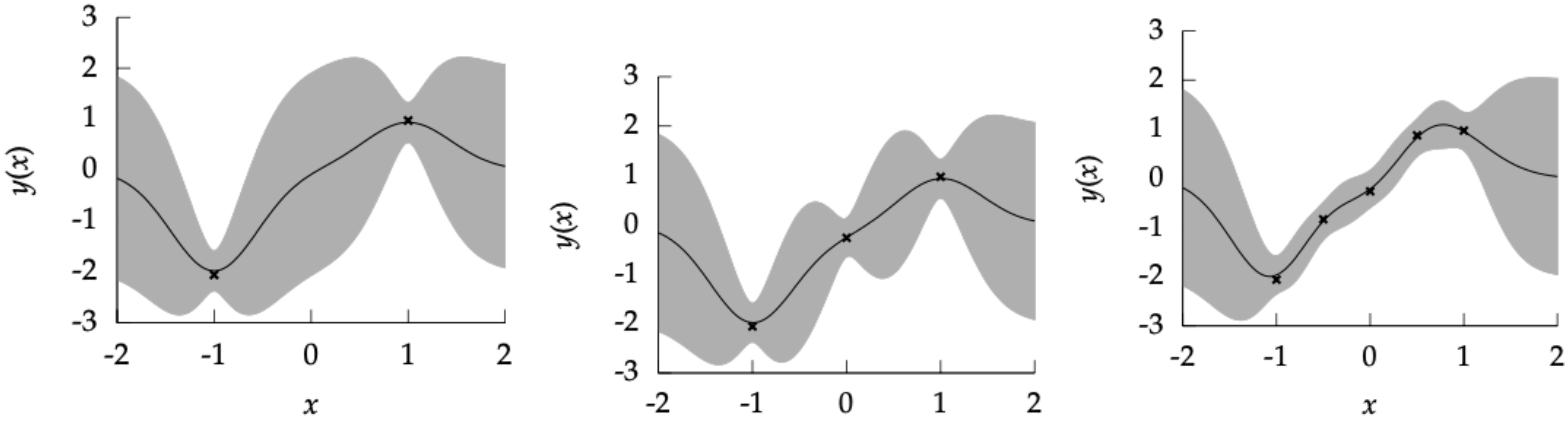
# Bayes' rule

- $$P(M|D) = \frac{P(D|M)P(M)}{P(D)}$$

- We are essentially considering an ensemble of models
- As we observe more data, distribution of the model concentrates (gets sharper): e.g., smaller confidence interval for parameters

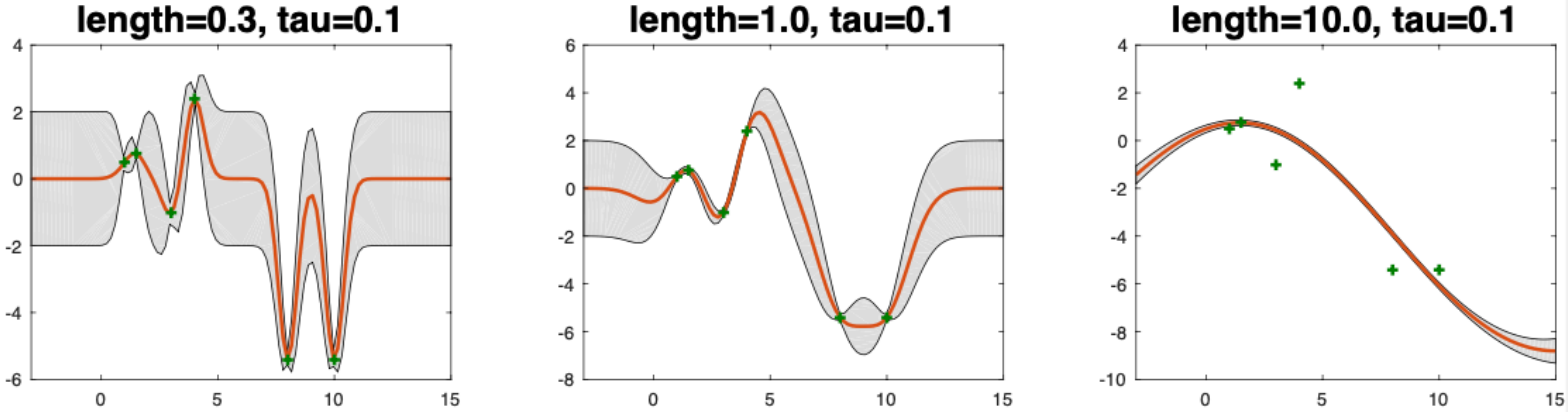
# Example: Gaussian process

- Distribution over regression functions
- $P(M)$ captures smoothness/“stiffness” of the function
- Obtain a Gaussian distribution for each predicted y



# Gaussian Process examples: different priors

$$k(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$



# Other uses of Gaussian Processes

- Active learning / sensing / collecting measurements in uncertain regions
- Bayesian Black-box Optimization

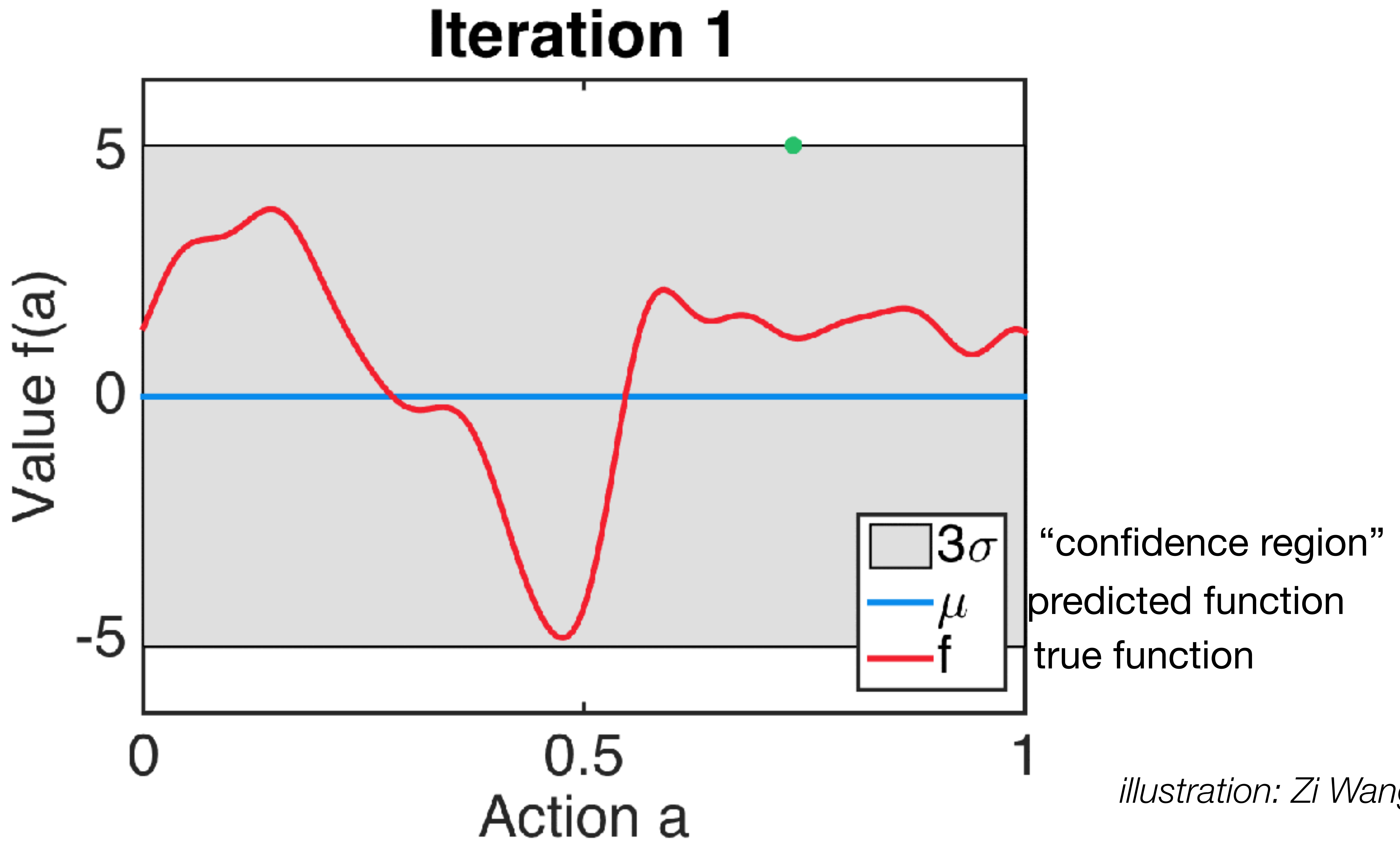



*illustration: Zi Wang*

# Some references


- *Angelopoulos & Bates, A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification.* *https://arxiv.org/abs/2107.07511*
- *Henrik Linusson. An introduction to conformal prediction. https://cml.rhul.ac.uk/copa2017/presentations/CP_Tutorial_2017.pdf*