# Local Explanations

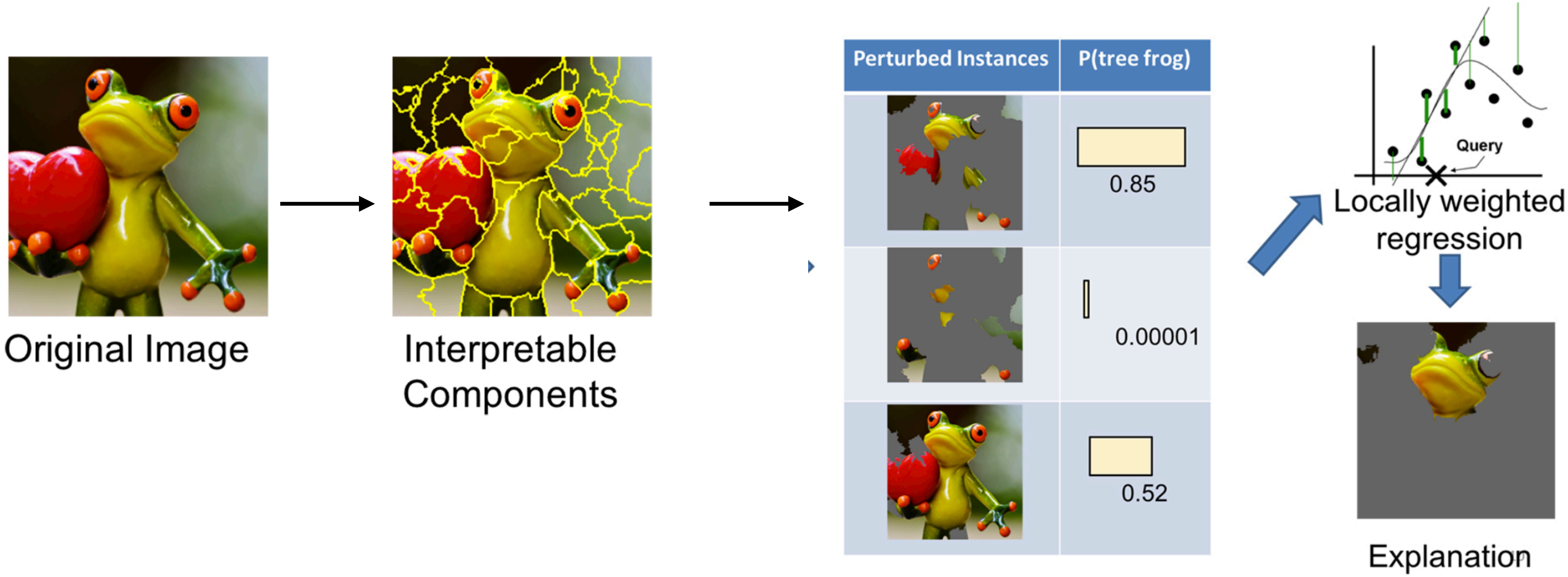
- An explanation should help us answer counterfactual questions: if the input were ***x***' instead of ***x***, what would the output be?

| | Model |
| --- | --- |
| *that movie was not great , in fact it was terrible !* | — |
| *that movie was not ____ , in fact it was terrible !* | — |
| *that movie was not great , in fact it was _____ !* | + |

- Perturb input many times and assess the impact on the model's prediction

# LIME

- LIME: Locally-Interpretable Model-Agnostic Explanations
  - *Local* because we'll focus on this one example
  - *Model-agnostic*: treat model as black box




- Check predictions on subsets of components
- Train a model to explain which components yield the model's preds

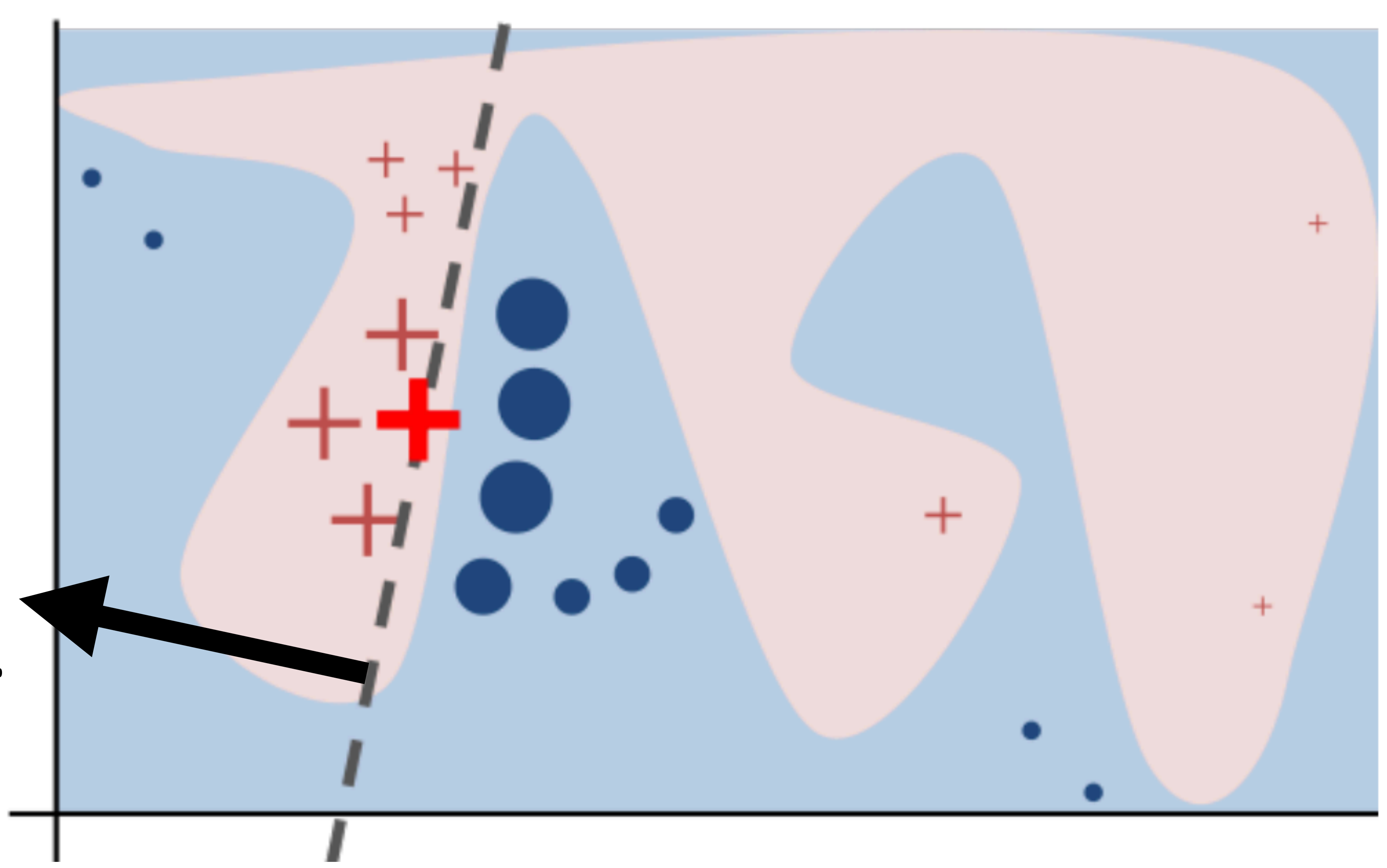
Ribeiro et al. (2016)

https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

# LIME

- Break down input into many small pieces for interpretability
  $x \in \mathbb{R}^d \rightarrow x' \in \{0, 1\}^{d'}$
- Draw samples by using x' as a mask to form a new example x''. Compute f(x'')
- Now learn a model to predict f(x'') based on x'. This model's weights will serve as the explanation for the decision

why it's +



- If the pieces are very coarse, can interpret but can't learn a good model of the boundary. If pieces are too fine-grained, can interpret but not predict

Ribeiro et al. (2016)

https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

# LIME

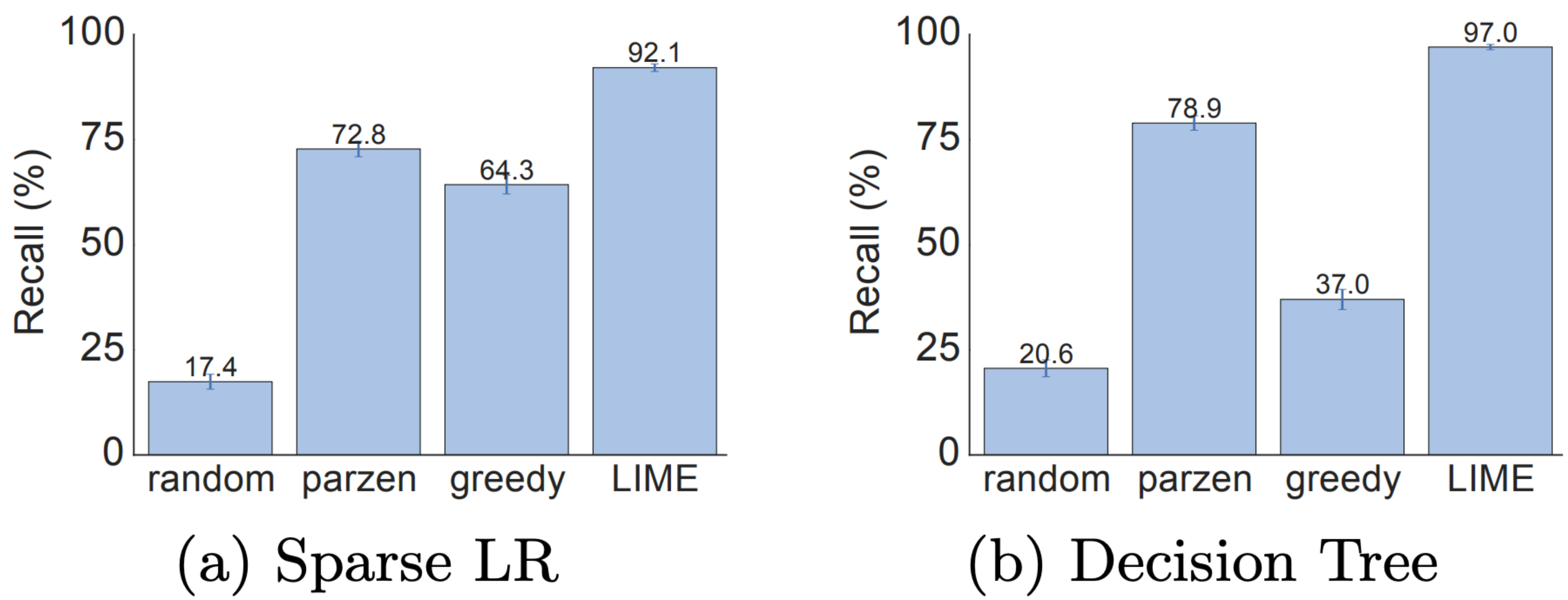



**Figure 6: Recall on truly important features for two interpretable classifiers on the books dataset.**

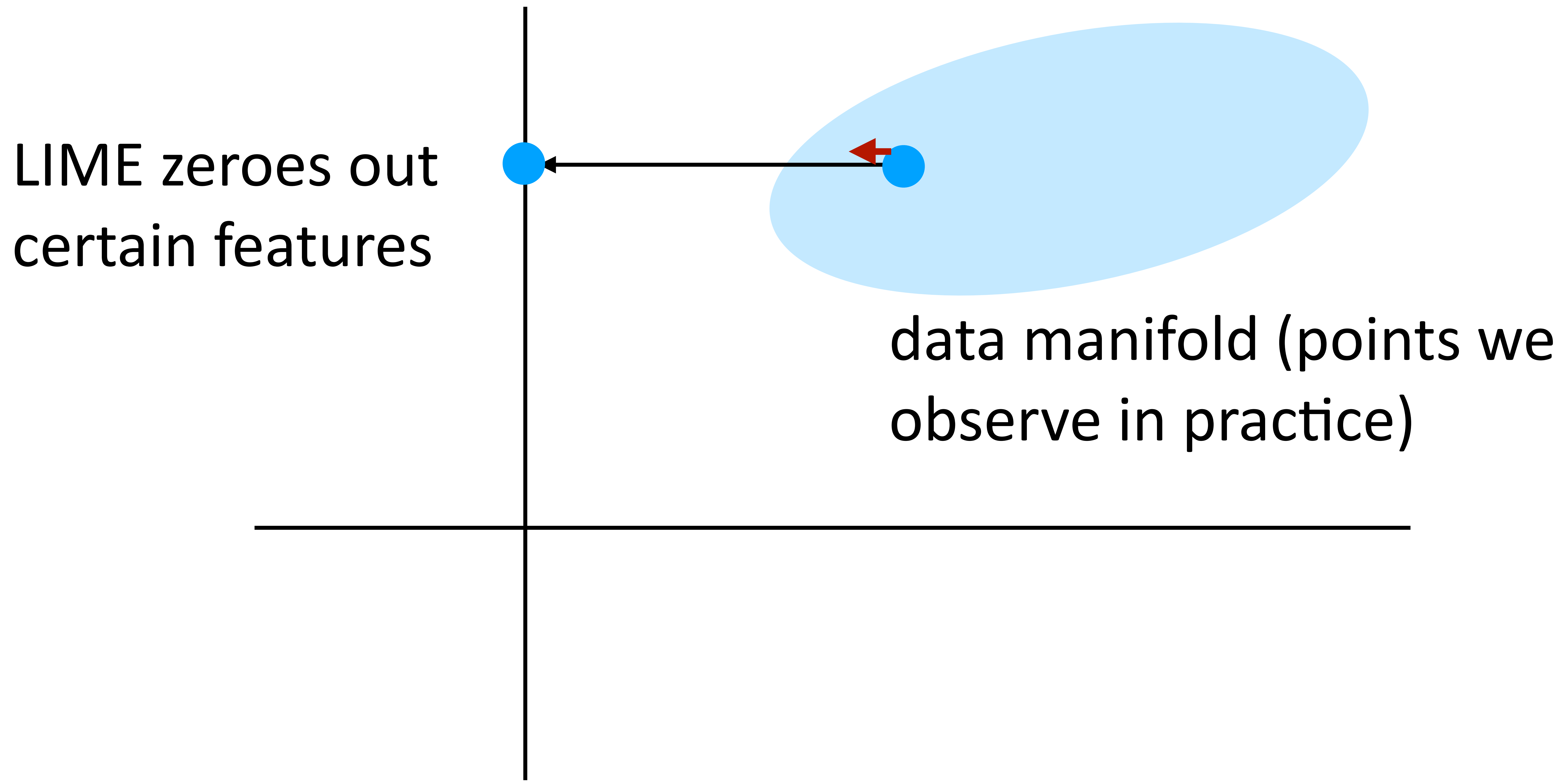
- Evaluation: the authors train a sparse model (only looks at 10 features of each example), then try to use LIME to recover the features. Greedy: remove features to make predicted class prob drop by as much as possible

Ribeiro et al. (2016)

https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

# Gradient-based Methods

- Problem: fully removing pieces of the input may cause it to be very unnatural




- Alternative approach: look at what this perturbation does locally right around the data point using gradients

# Gradient-based Methods

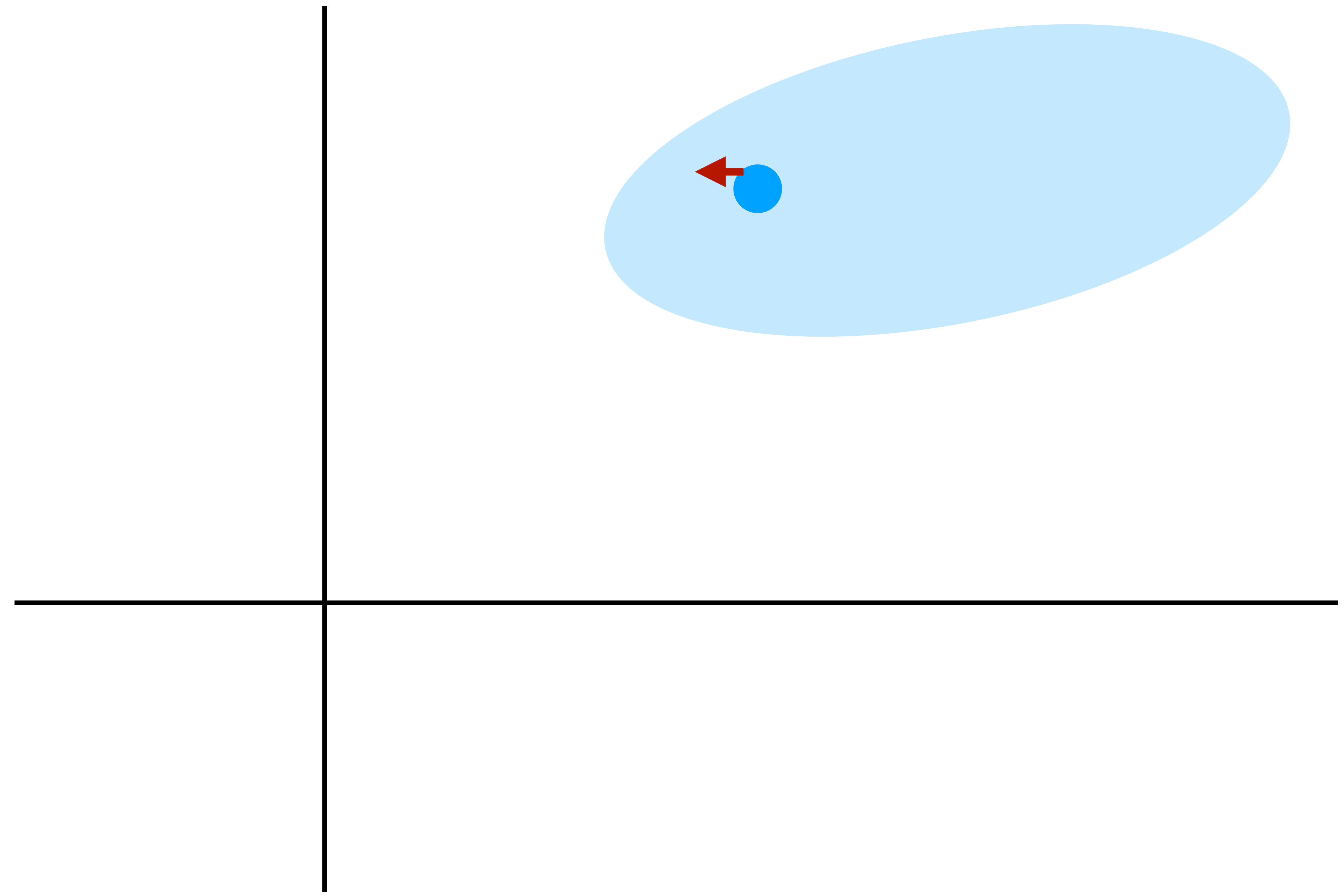
- Originally used for images
- Approximate score with a first-order Taylor series approximation around the current data point

$S_c$ = score of class $c$

$I_0$ = current image

$$S_c(I) \approx w^T I + b$$

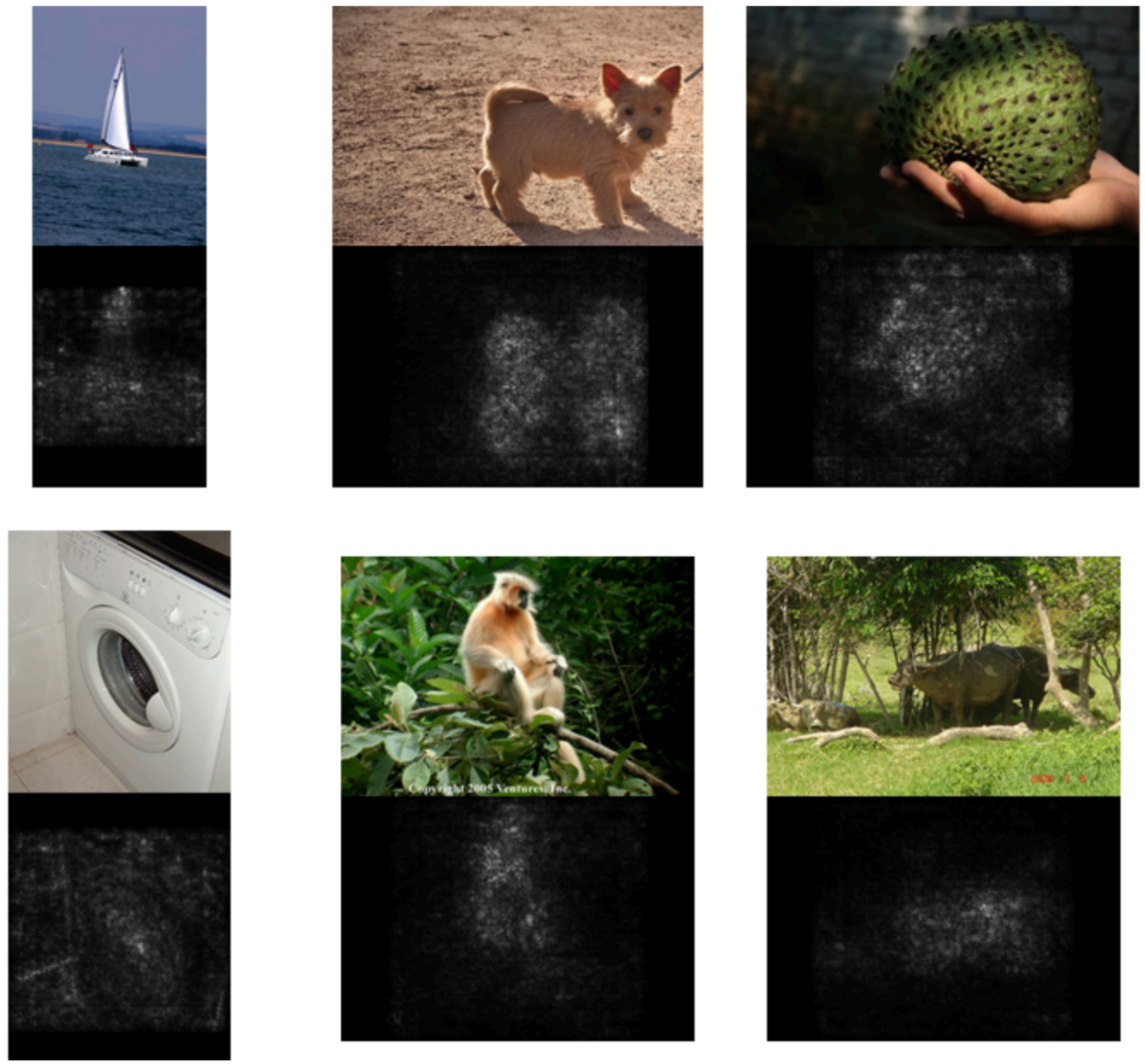
$$w = \left. \frac{\partial S_c}{\partial I} \right|_{I_0}$$



- Higher gradient magnitude = small change in pixels leads to large change in prediction

Simonyan et al. (2013)

# Gradient-based Methods

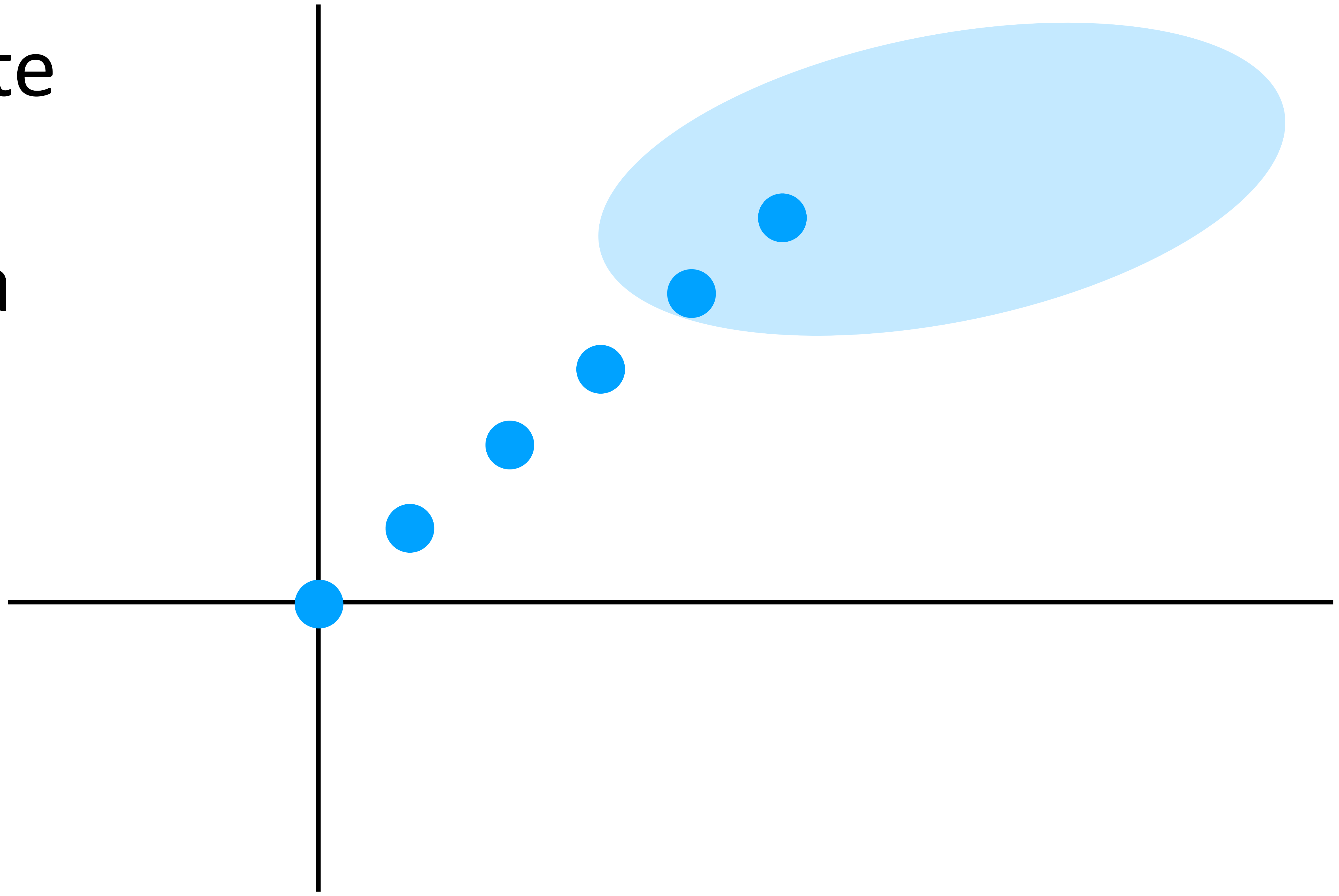


Simonyan et al. (2013)

# Integrated Gradients

- Suppose you have prediction = A OR B for features A and B. Changing either feature doesn't change the prediction, but changing both would. Gradient-based method says neither is important

- Integrated gradients: compute gradients along a path from the origin to the current data point, aggregate these to learn feature importance



- Now at intermediate points, increasing "partial A" or "partial B" reveals the importance of A and B

Sundararajan et al. (2017)