
Simplicity

Linear Functions

- Simple function

$$f(x) = w^\top x + b = \sum_{i=1}^{n} w_i x_i + b$$

- Easy to interpret everything
  - Weights w_i tell us how important feature i is
  - Covariates x_i tell us how strong this is for specific x
- Problems if we have too many coordinates
  (e.g. images, microarray data, time series)

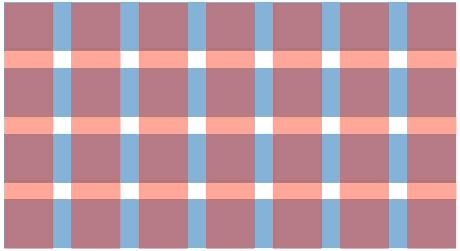
Linear Functions

- Simple function

$$f(x) = w^\top x + b = \sum_{i=1}^{n} w_i x_i + b$$

- Sparsity regularization for extra simplicity
  - Vanilla penalty $\lambda \|w\|_1$ (with lots of theory for it)
  - Structured penalty (with even more theory for it)

$$\sum_i \|W_{[\cdot,i]}\|_1 + \|W_{[i,\cdot]}\|_1$$



Group penalty

Linear Functions

- For high-dimensional data this is still too complex
- Linearity is not always a good model
- **Often not suitable for humans**
  - Too tedious to evaluate
  - Difficult to understand
  - Difficult to operationalize **in a stressful situation**

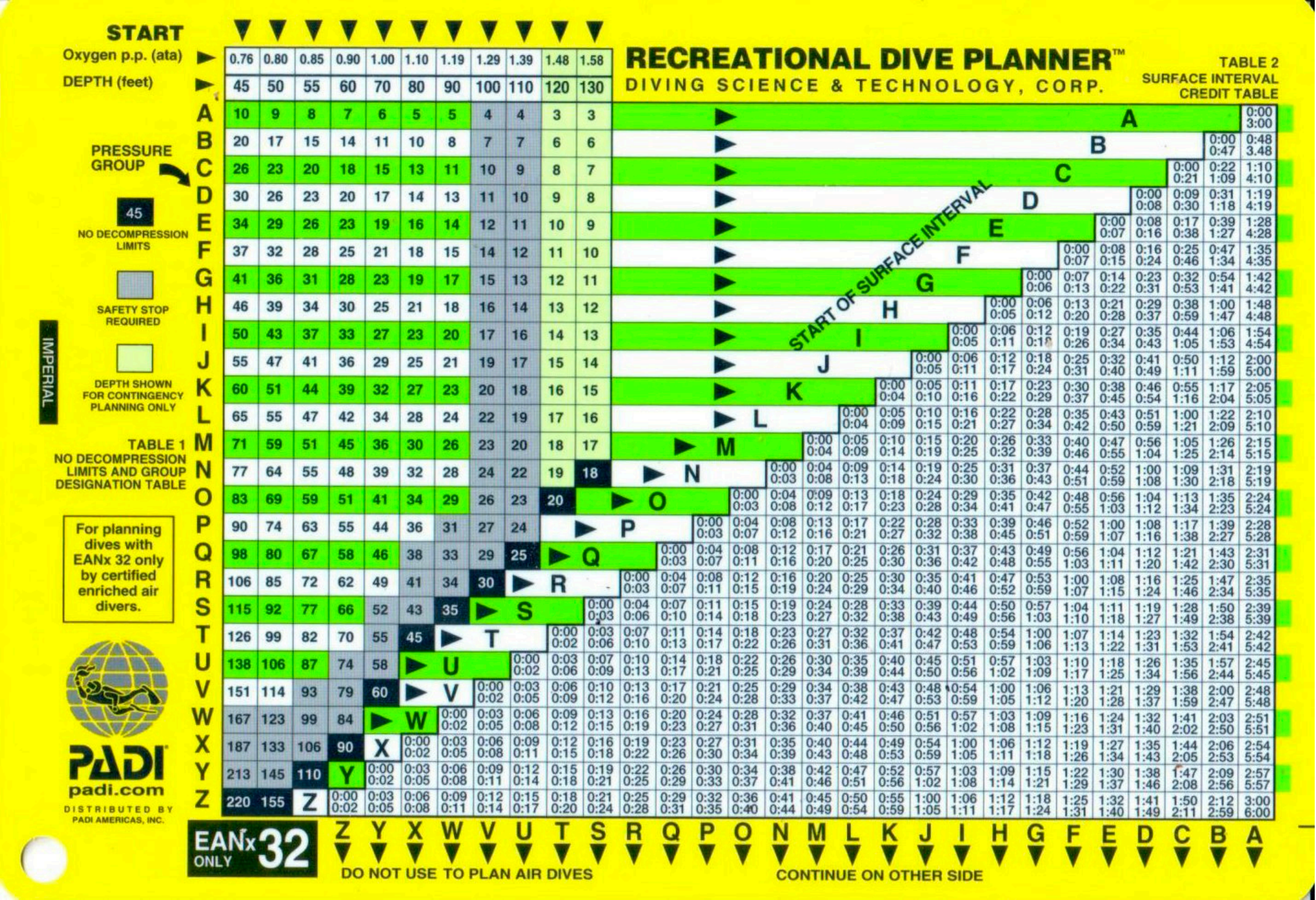
PADI

Dive

Chart

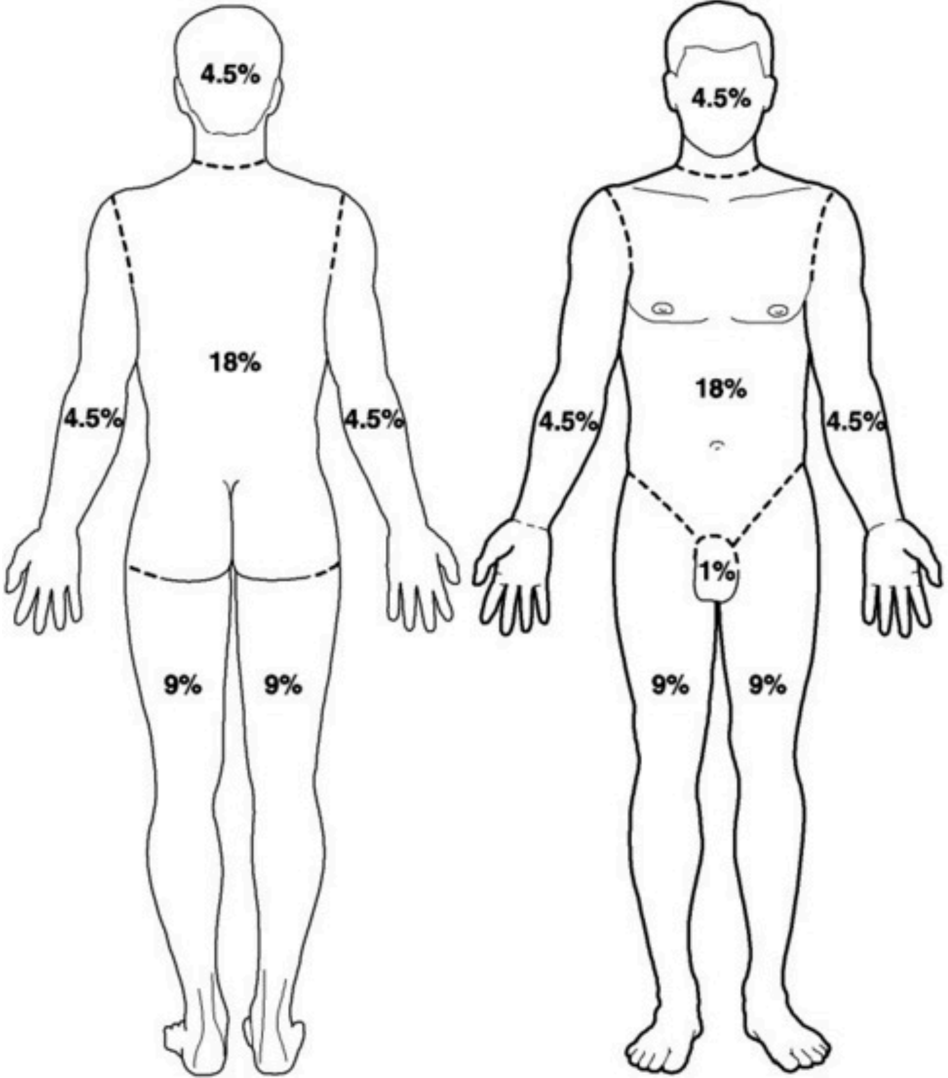
Solves PDE.

Use while Diving.



Rule of Nines

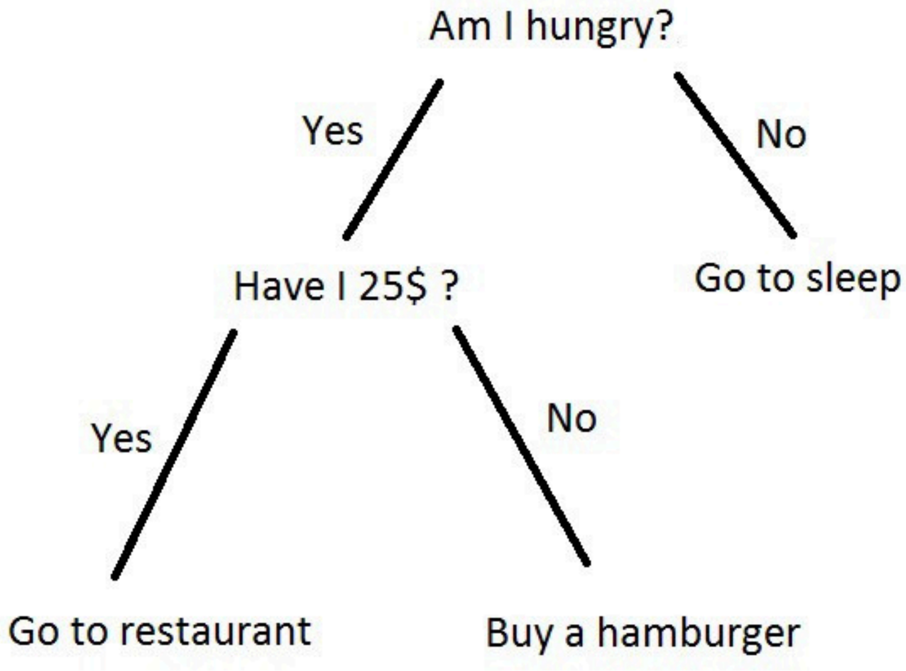
- Trauma Center
  - Assess whether patient will survive their burns.
  - Calculate burned area
  - Use to guide treatment and survival probability.
- Easy to remember (simple percentages)



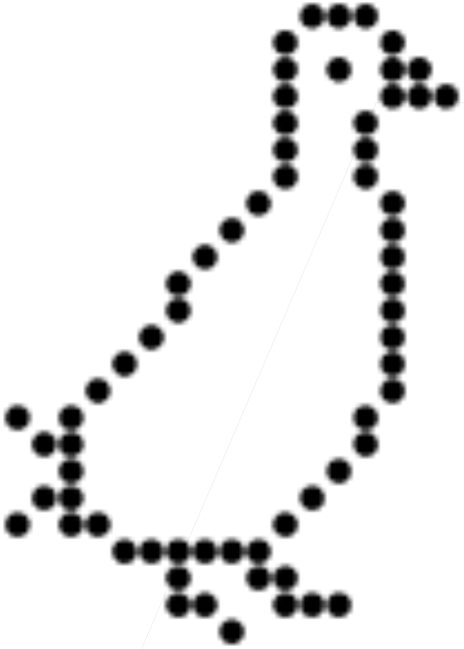
Trees and Lists (see e.g. Rudin & coauthors)

- Estimate simple function
  - Decision tree
    (only simple if small)
  - Decision list
    (if then elsif then elsif then)

Type/probability	# bits	code
$P(\text{Trek}) = 0.5$	1	1
$P(\text{Specialized}) = 0.25$	2	01
$P(\text{Cervelo}) = 0.125$	3	001
$P(\text{Serrota}) = 0.125$	3	000



Approximate Simplicity

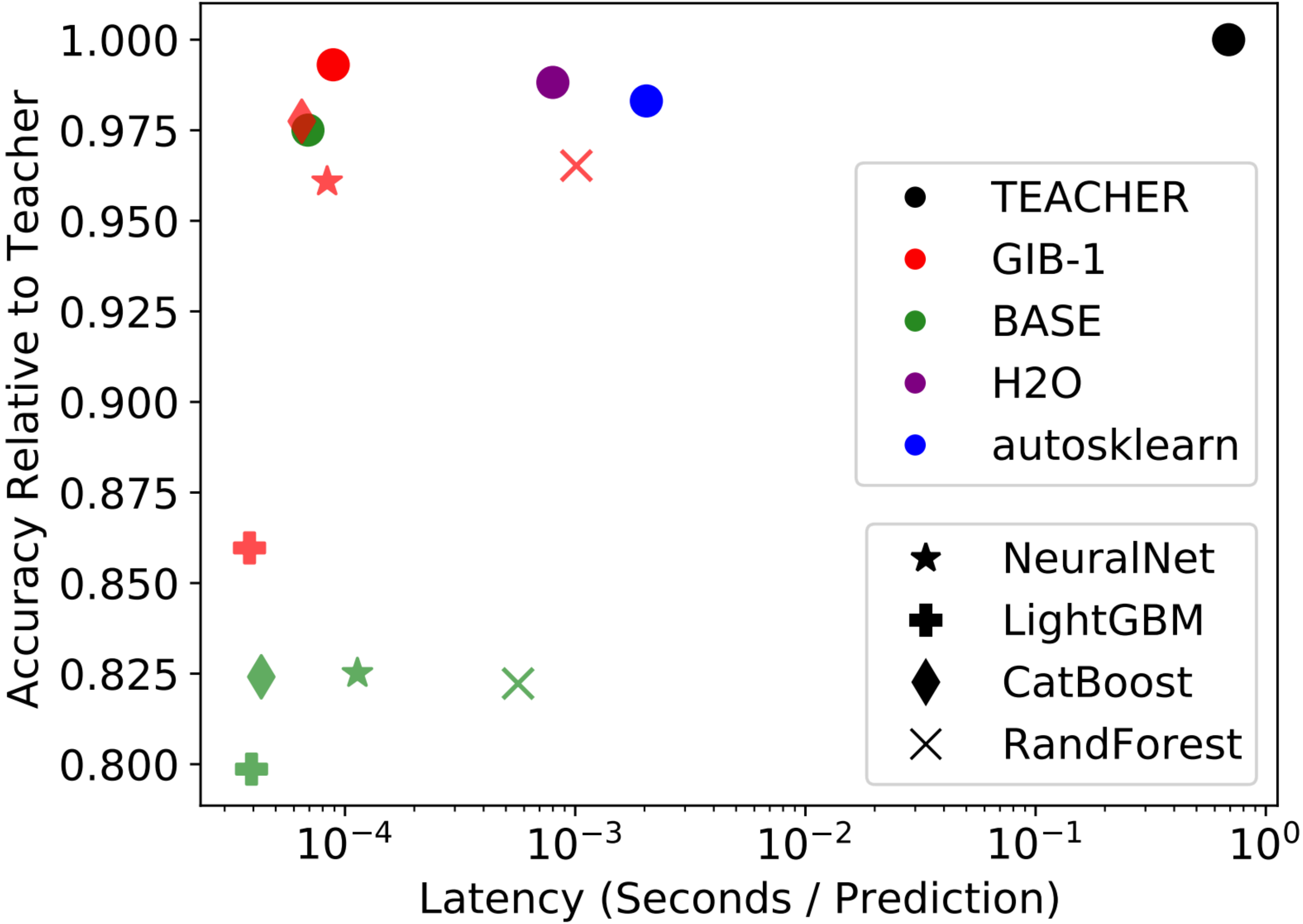


Alternatives

- Very difficult to estimate efficiently
  - Difficult to incorporate prior knowledge
  - Bagging / stacking / model combinations
  - Large amounts of data
- Distillation to the rescue (train with AutoML then distill)

$$\underset{g}{\text{minimize}} \sum_i l(g(x_i), f(x_i))$$

Fakoor et al., 2020, Fast, Accurate, and Simple Models for Tabular Data via Augmented Distillation, https://arxiv.org/abs/2006.14284


Accuracy Relative to Teacher
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0.800
Latency (Seconds / Prediction)
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
TEACHER
GIB-1
BASE
H2O
autosklearn
NeuralNet
LightGBM
CatBoost
RandForest

Generating Data for Distillation

- **Training data**
  - Might be small
  - Already used it to train fancy model
    Labels probably not a lot better than true labels, if at all.
- **Auxiliary data**
  - If unlabeled data available, it's perfect
  - If not, learn generative model on training data (e.g. via Transformer Density Estimation)
  - Gibbs sampler explores neighborhood $x_i \sim p(x_i \mid x_{-i})$

Local Approximation (LIME)

Generic Classifier

- Black Box classifier can be very complex
- Impossible to approximate well in general
- Maybe can be linearized **locally**

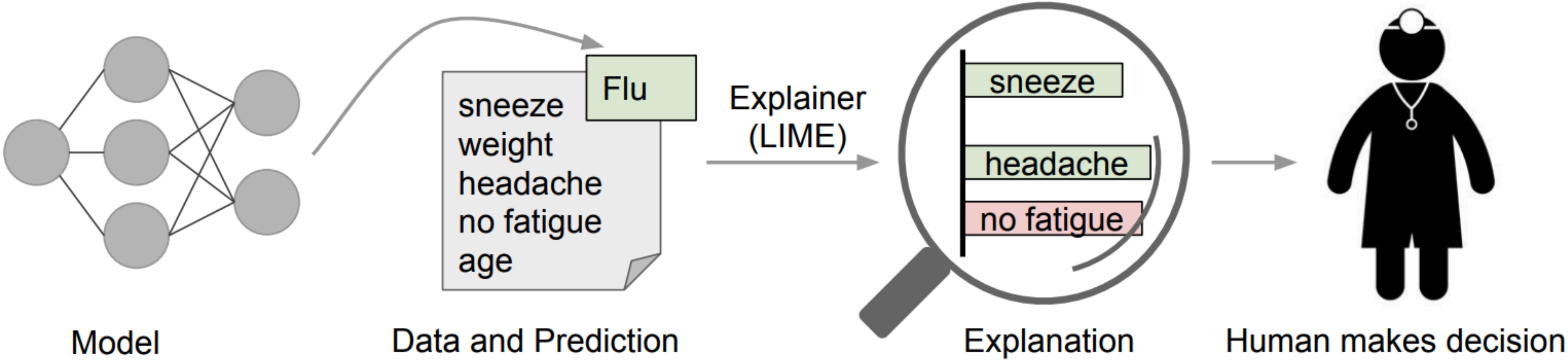
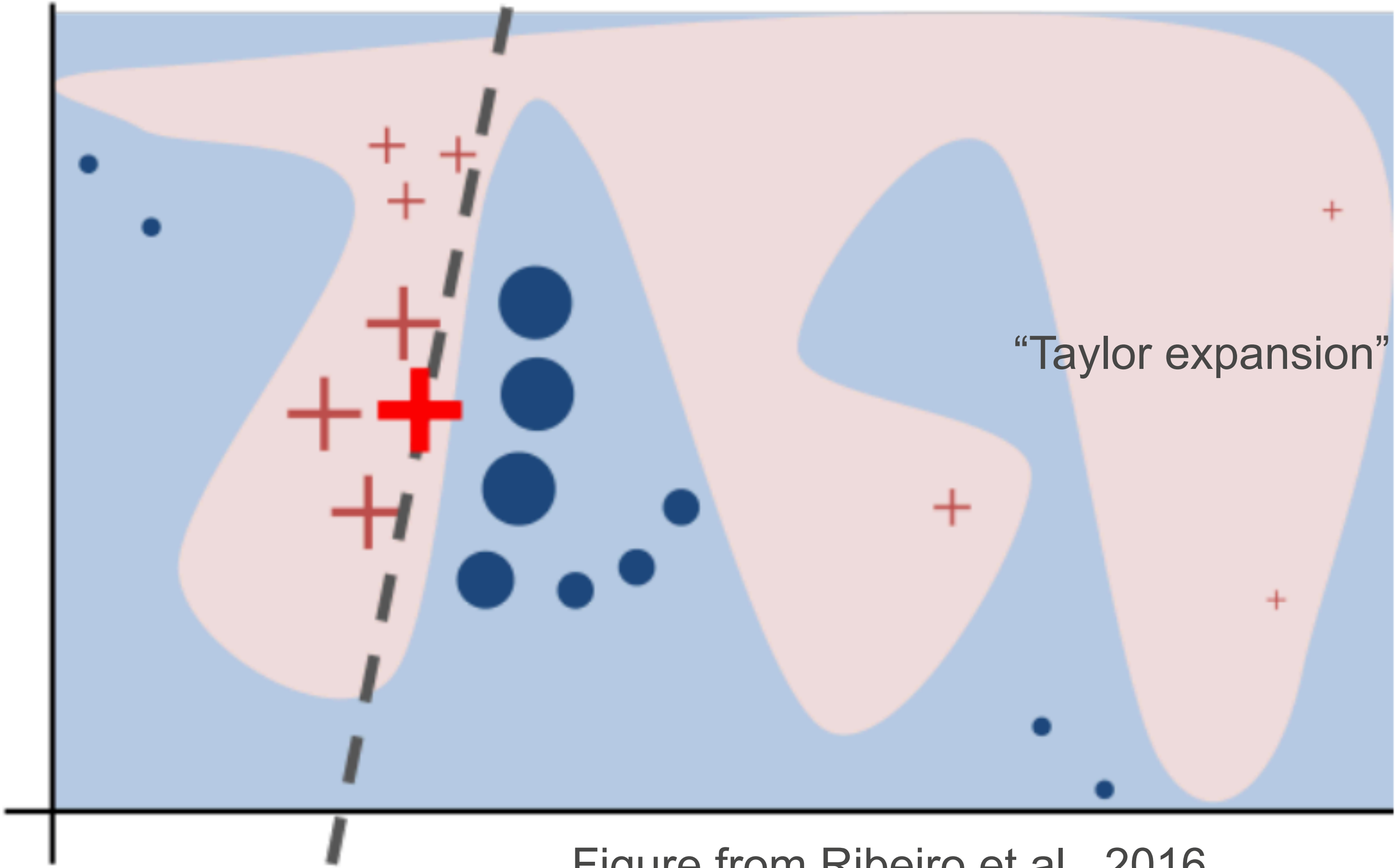


Figure from Ribeiro et al., 2016

Local Approximation



“Taylor expansion” by regression

Figure from Ribeiro et al., 2016

Locally Linear Approximation

- Get data x_j around query x
- Approximate $(x_j, f(x_j))$ pairs with function g
- Can also select across all datapoints for global features (sub modular feature selection)

Figure from Ribeiro et al., 2016

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*