# Explainability

## Preparing for Production

Dr Zak Varty

# What are we explaining and to whom?

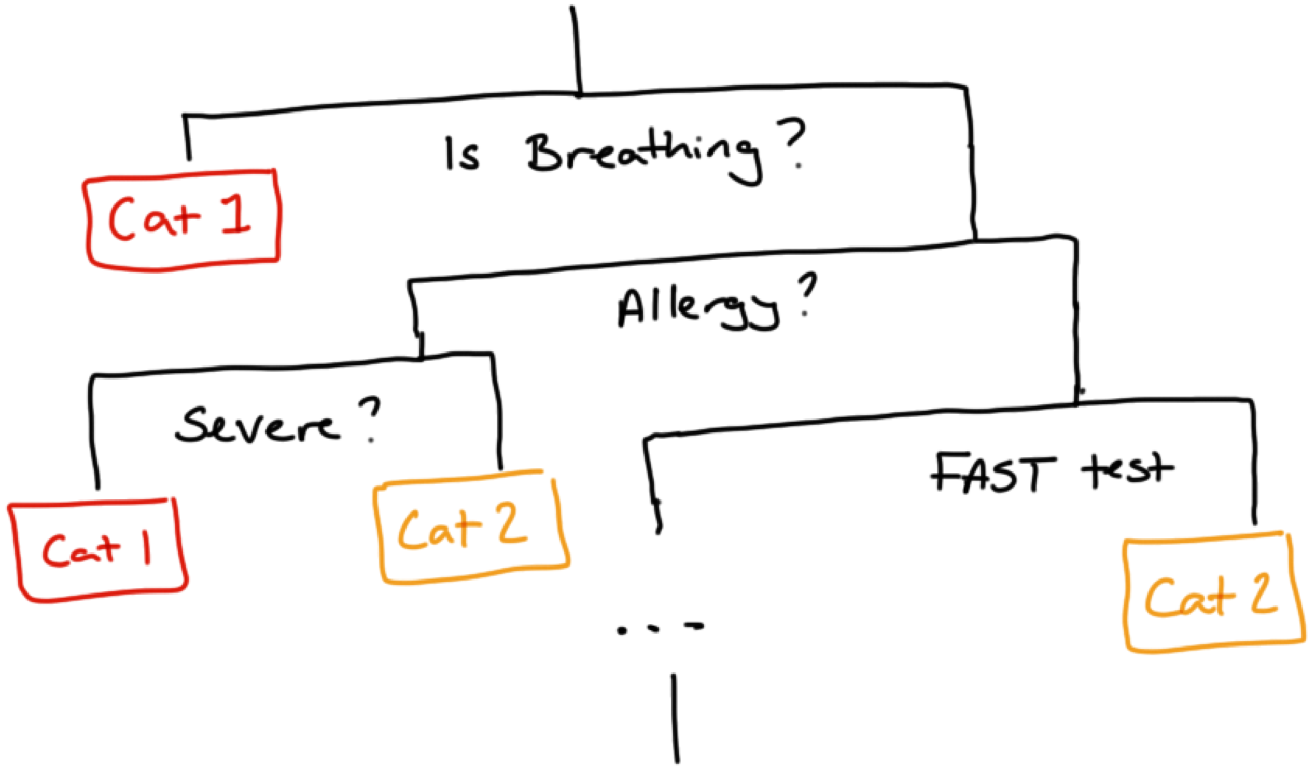
**Working example:** credit scoring system:

Regulatory or legal requirements;

Understand your model and convince stakeholders to use it;

Justify decisions to individual customers.

# Explaining a Decision Tree

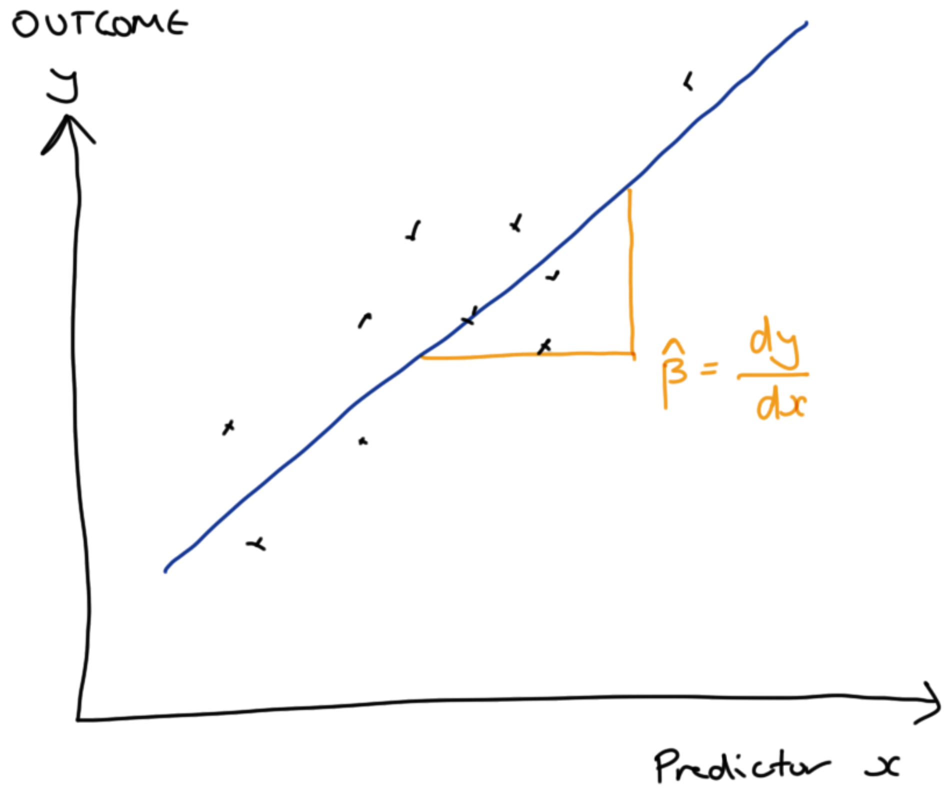


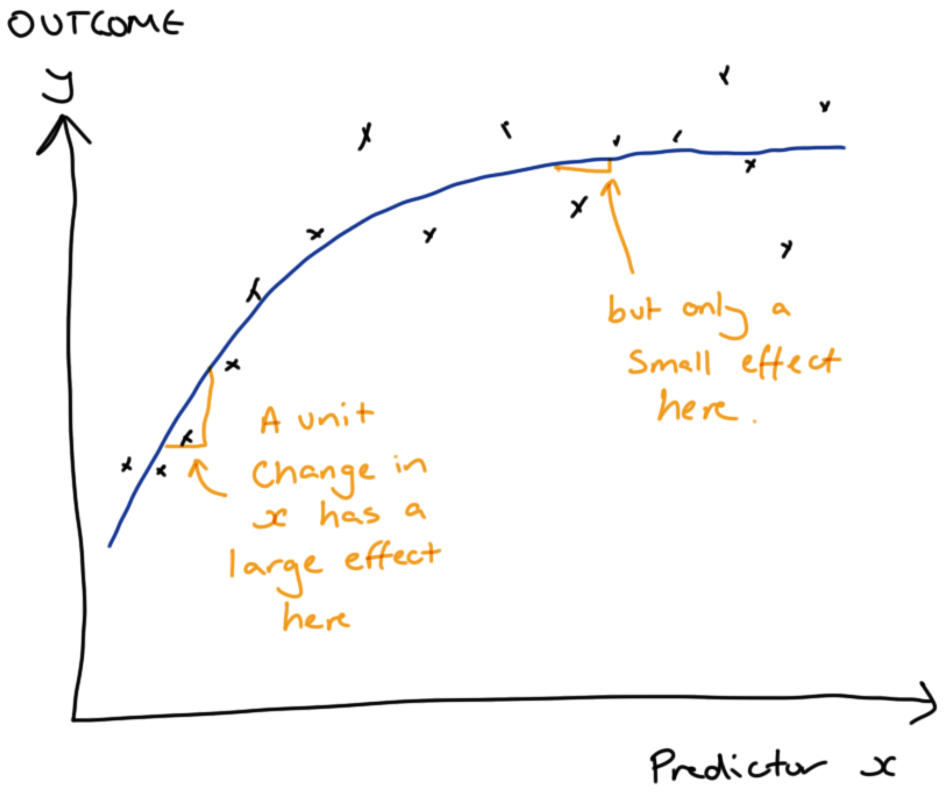

**Example:** Decision Tree

- Repeatedly partition covariate space
- Mimics human decision making
- Medical triage optimises for speed
- Usually optimise for best classifier
- Trade-off flexibility & robustness for an explanation.

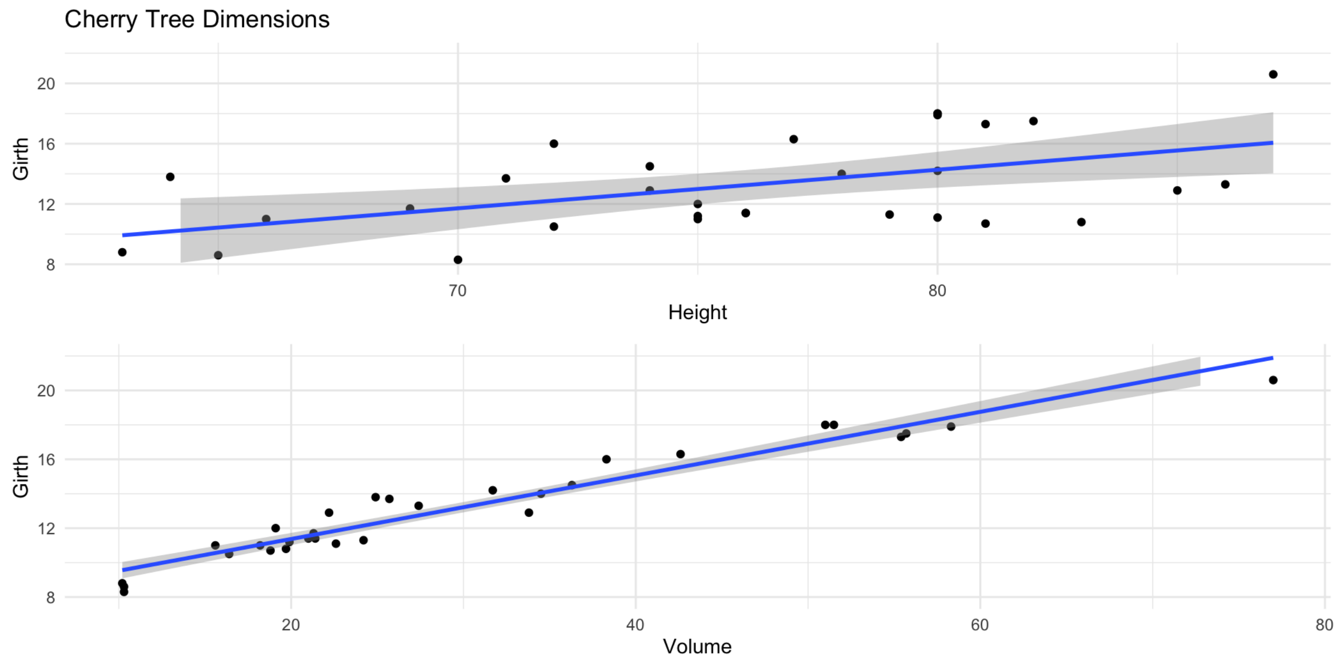
# Explaining Regression Models

- **Conditional explanation**: all other covariates held constant
- **Global/Local explanation** in linear/non-linear regression



# Explaining Regression Models: Cherrywood Example




```r
lm(Girth ~ 1 + Height, data = trees)
```

```
Call:
lm(formula = Girth ~ 1 + Height, data = trees)

Coefficients:
(Intercept)       Height
    -6.1884       0.2557
```

```r
lm(Girth ~ 1 + Volume, data = trees)
```

```
Call:
lm(formula = Girth ~ 1 + Volume, data = trees)

Coefficients:
(Intercept)       Volume
     7.6779       0.1846
```

# Including both terms

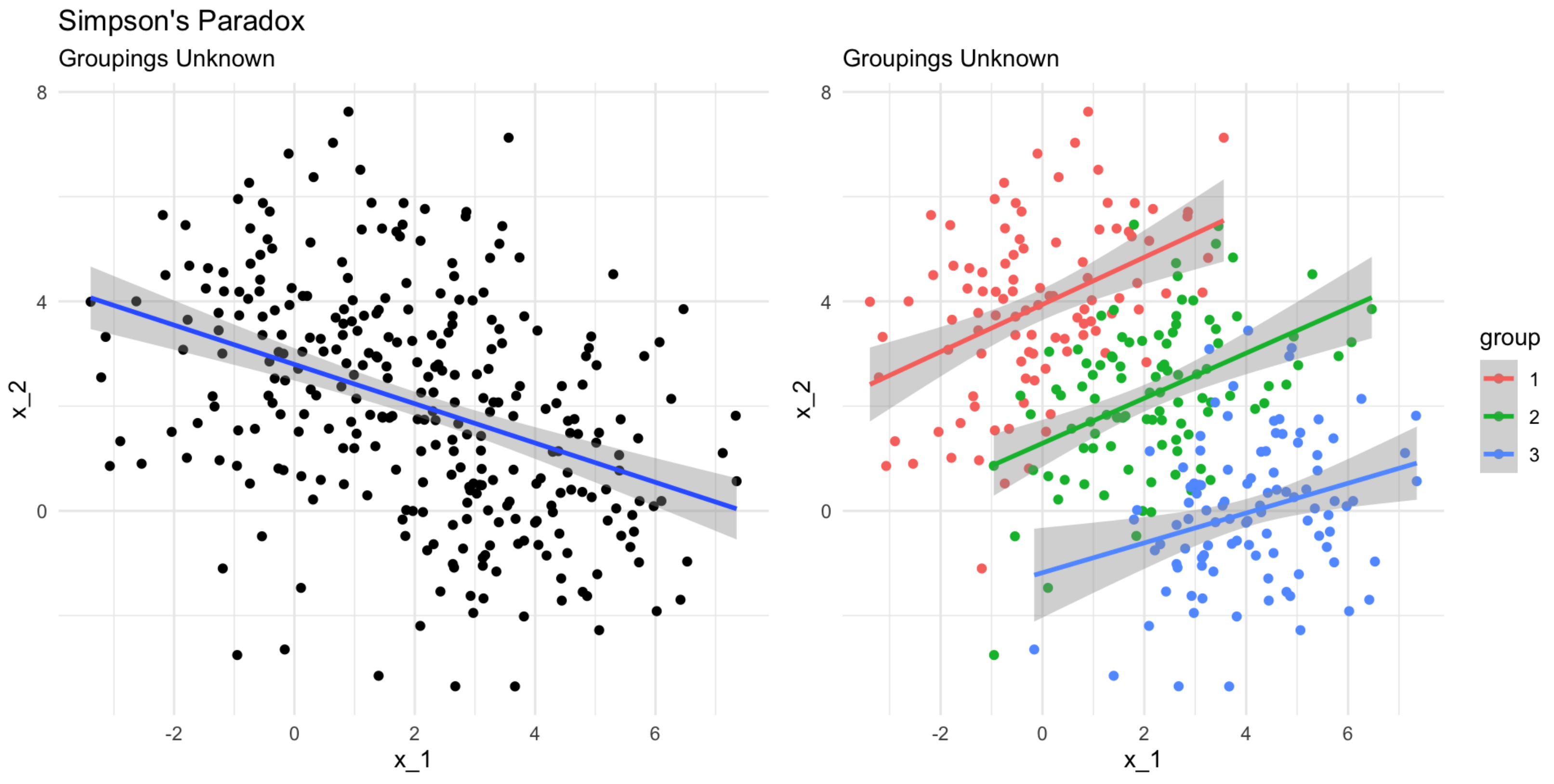
```r
lm(Girth ~ 1 + Height + Volume, data = trees)
```

```
Call:
lm(formula = Girth ~ 1 + Height + Volume, data = trees)

Coefficients:
(Intercept)       Height       Volume
   10.81637     -0.04548      0.19518
```

- Height no longer significant, sign of point estimate changed.
- Effect and interpretation depend on which covariates are included.
- SHAP averages over all combinations.

# Simpson’s Paradox

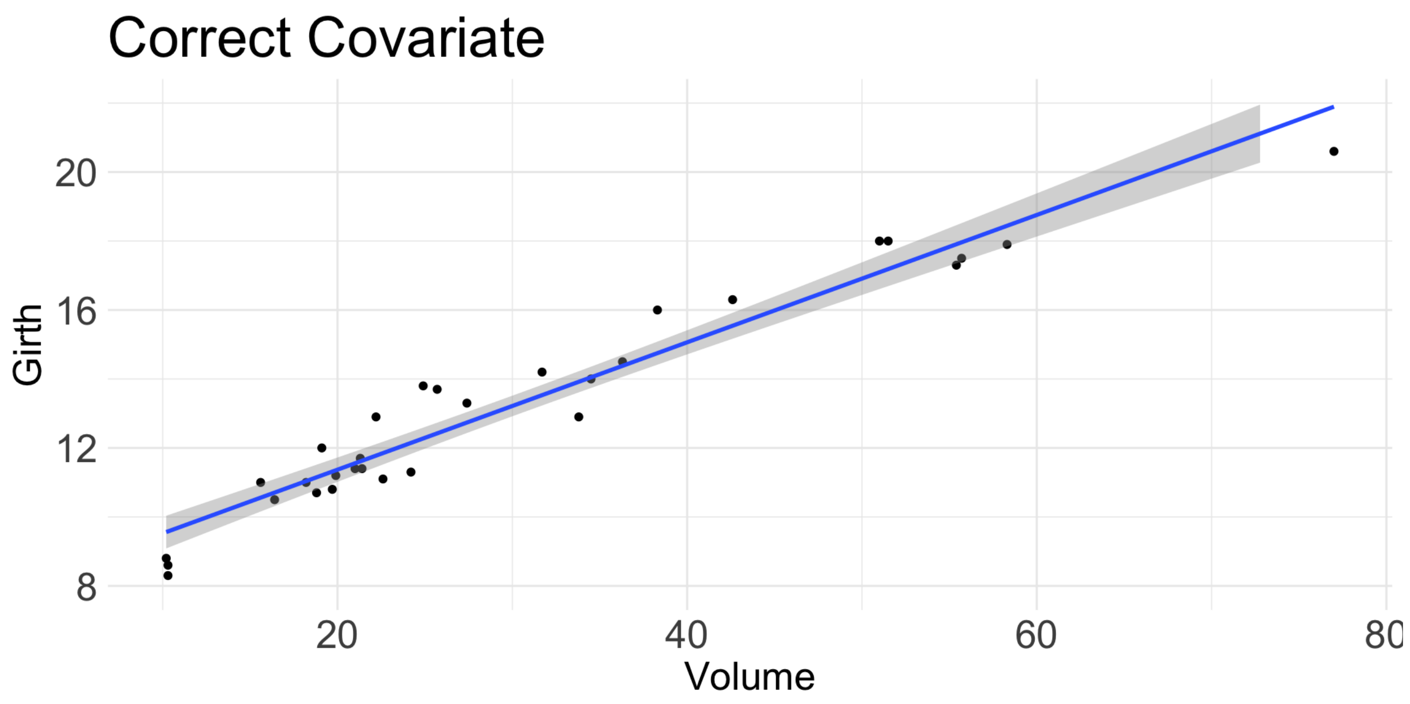


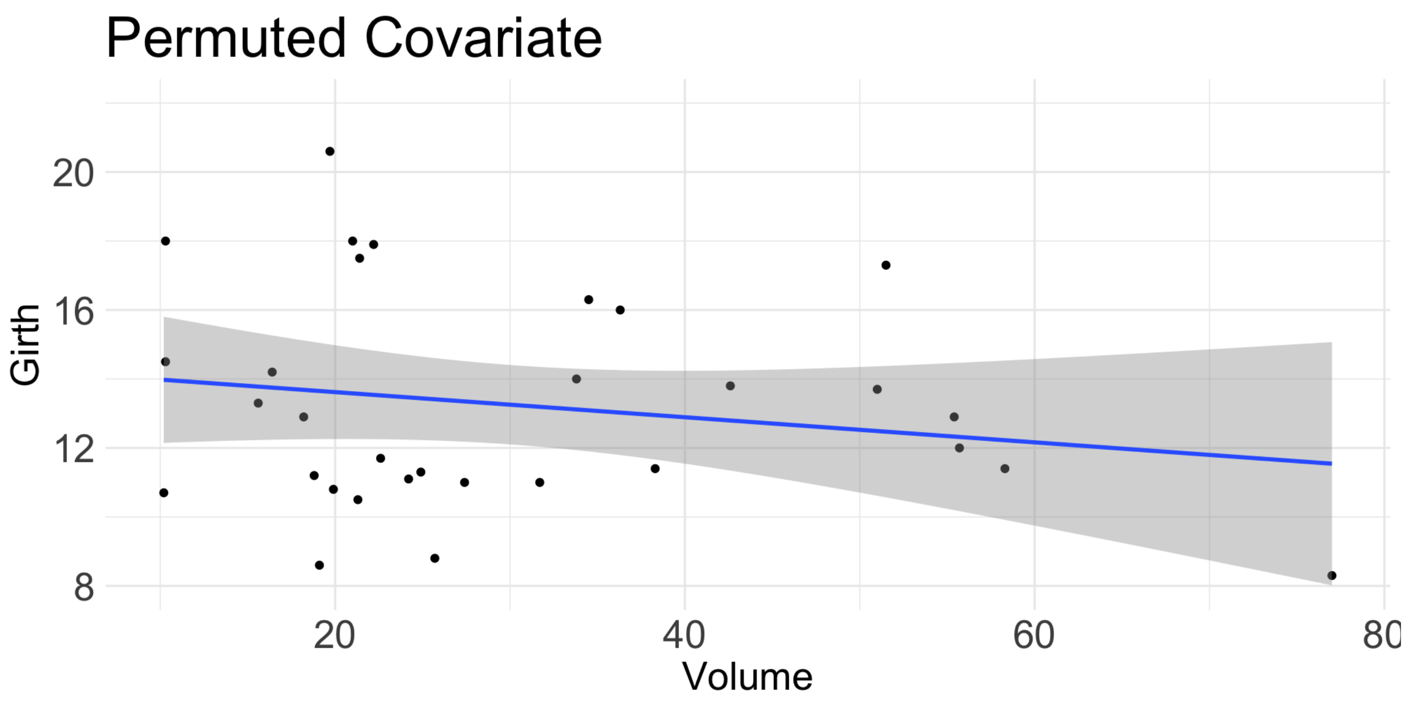

- Trend disappears or reverses when groups are split / combined.
- Lots of other names, including Ecological Fallacy.
- Not actually a paradox at all

# What hope do we have?

# Permutation Testing

Shuffle covariate values to remove any relationship & inspect how predictions change.

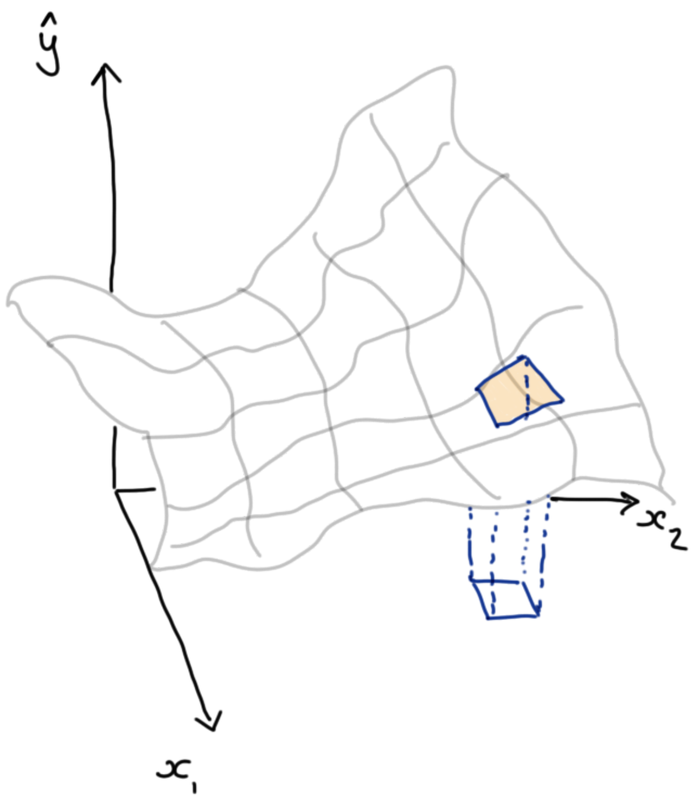


# Meta-modelling

Construct an explainable model to describe the local behaviour of a model that is not explainable.

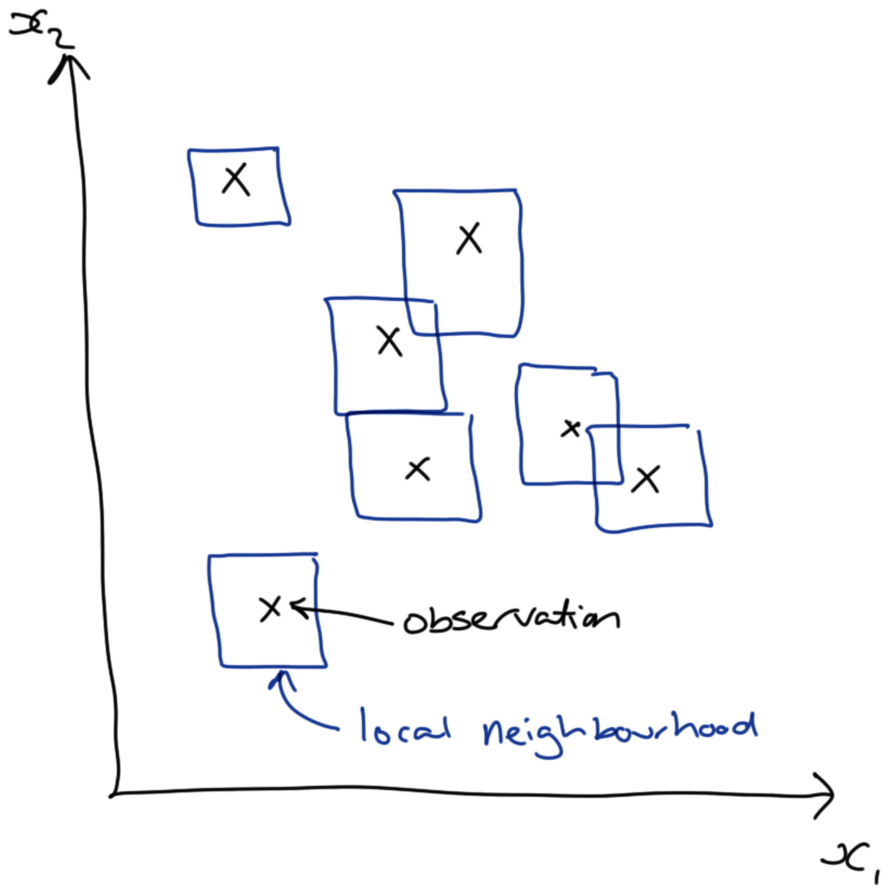
Methods like LIME use a linear meta-model motivated by Taylor's Theorem.




A local linear approximation in two dimensions

# Aggregation

- Can move from conditional to marginal explanations and from local to global explanations.
- Requires **integration** over $f(x)$, which is unknown.
- Use empirical distribution instead and this simplifies to a sum!




Local approximations around each observation can be combined to understand global model behaviour.

# Wrapping Up

- Trade-off between complexity and explainability.
- Conditional effects can be tricky to explain.

- Approximate more complex models to get localised explanation.
- Aggregate local, conditional effects to get global, marginal effects.