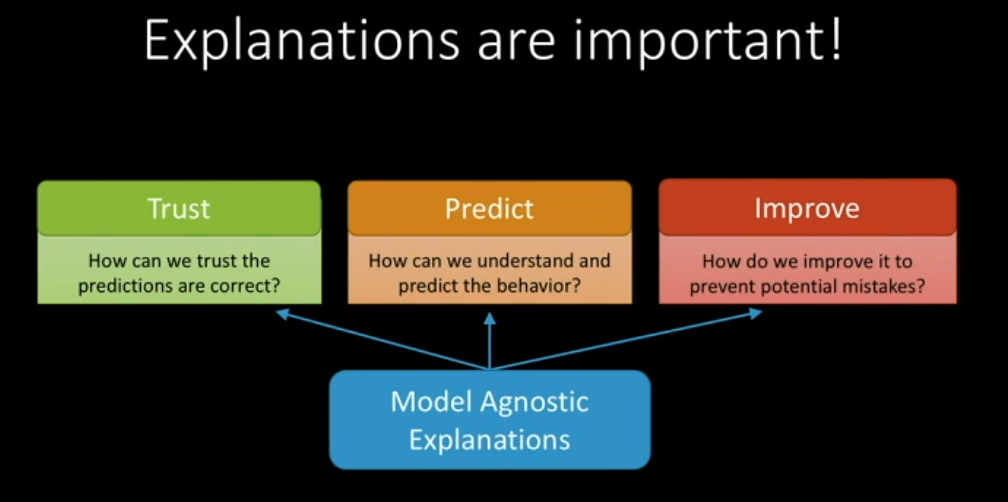
How can we predict what our algorithm is going to do.

Reduce the risk / instability in our predictions. Aim:



Try to move away from the black box concept

ML started working on simple data eg. 2 dimensional data, very easy to classify and say say one dimension has higher effect than other by looking at coeffiecients

From here we use more lines to split groups : if else relationships : can still explain structure

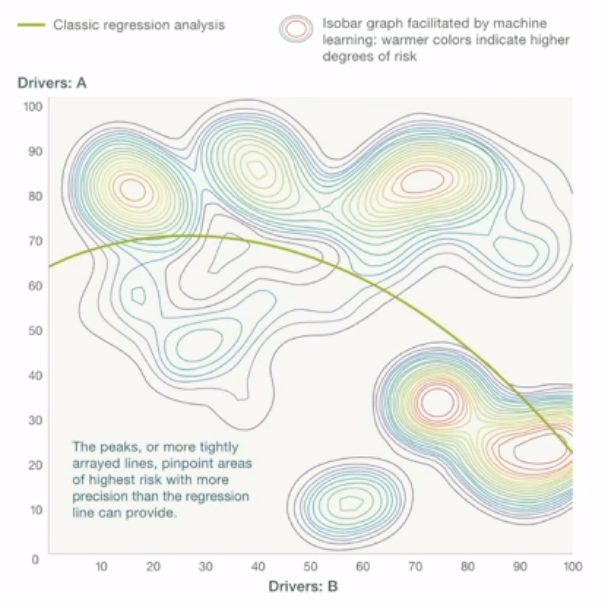
1. Trust preds are correct: We can trust a model or not if its structure matches our intuitions
2. Predict behaviour and understand: Structure tells us exactly what will happen on any data
3. Import: Structure tells us where the error is and thus how to fix it and prevent potential mistakes

Big data changed the applications of ML

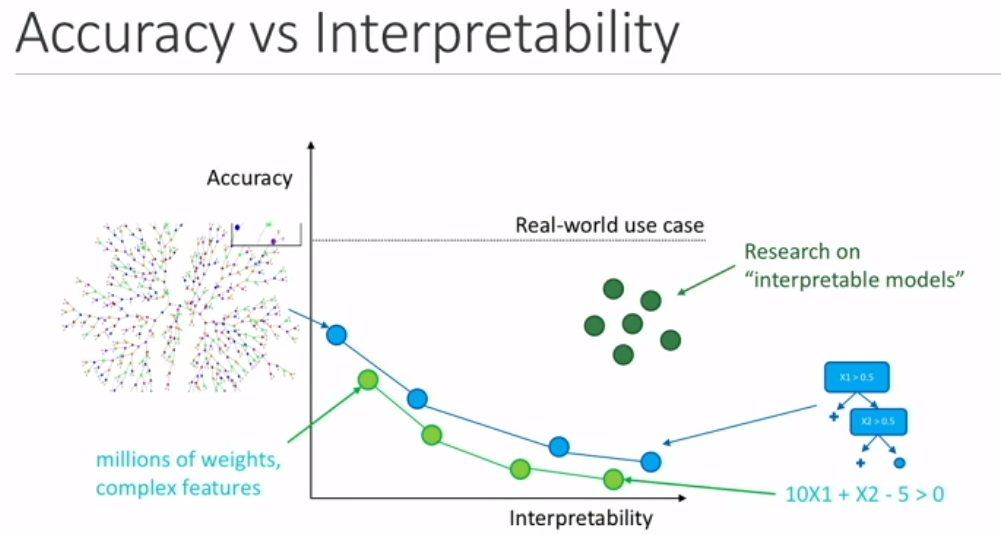
Language, economics, biomedicine, banking, retail, biology etc.

Underlying data became more complex: equations became very high polynomial curves

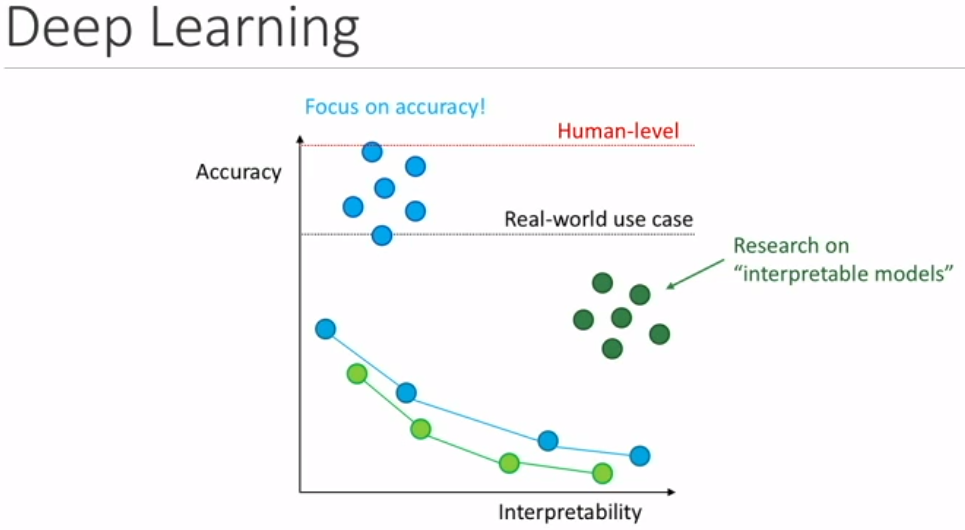
Surface becomes more complex as well as the number of dimensions eg. Thousands/millions of features.



We can’t even visualize 4d let alone more -> black box learning

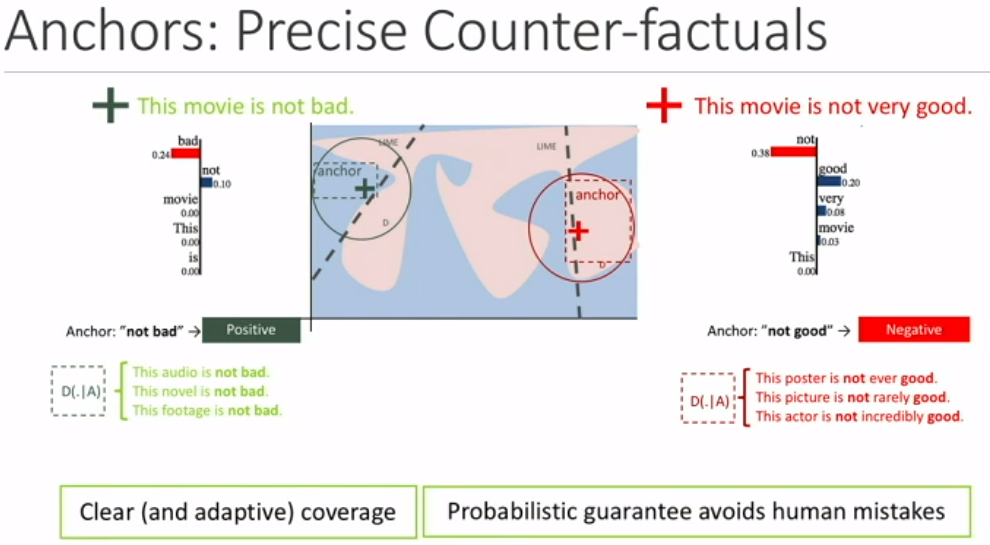


Deep learning came along and focused on accuracy (sometimes beating humans!) YET with zero interpretability and structure (like before)



**LIME algorithm to simplify and explain predictions:**

Forget structure, replace with f(x), treat as black box (can substitute any algorithm into it) so to make everything interpretable without any sacrifice in accuracy.



How do we choose which definition to use on unseen data.

Use an anchor of test sentence sets – if holds for majority then good to go.

Anchor explanations is a rule (square) now rather than a set of coefficients (circle).

Eg.

