

Visualising ML Models

Amit Kapoor
@amitkaps

Story

“We don’t see things as they are, we see them as we are.”

— *Anais Nin*

The Blind Men & the Elephant

“And so these men of Indostan
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong,
Though each was partly in the right,
And all were in the wrong.”

— *John Godfrey Saxe*

The Elephant: Data

“Data is just a clue to the end
truth”

— *Josh Smith*

The Men: Building Models

"All models are wrong, but some
are useful"

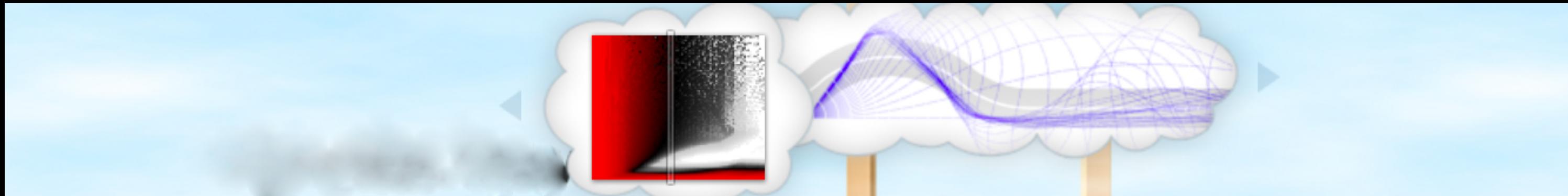
— George Box

Ladder of Abstraction

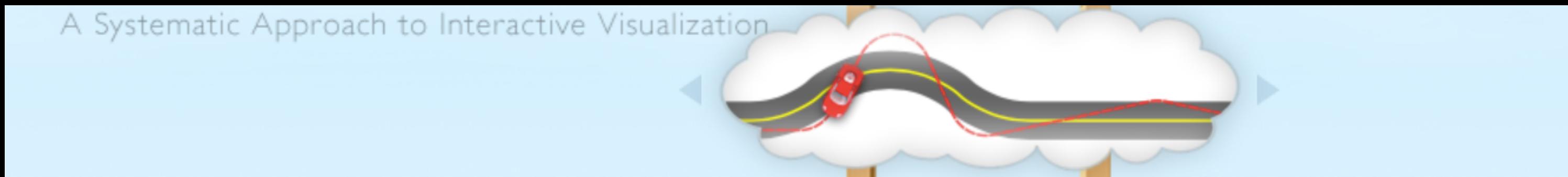
Data Abstraction

Visual Abstraction

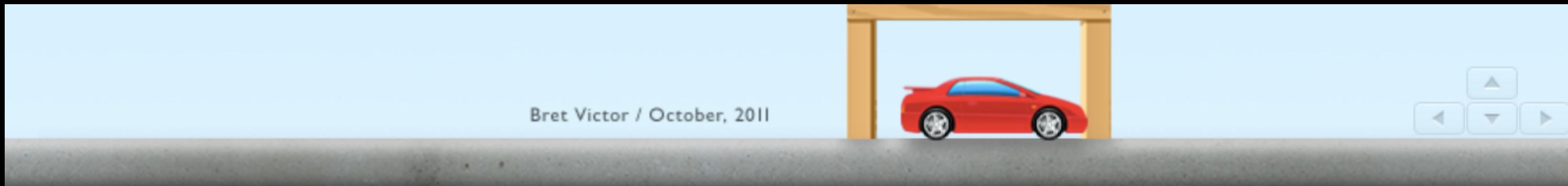
Model Abstraction



Up and Down the Ladder of Abstraction



A Systematic Approach to Interactive Visualization



Bret Victor / October, 2011



Why Build Models?

First Level of Ignorance

"I know, what I don't know"

Why Visualise Models?

Second Level of Ignorance

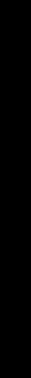
"I don't know, what I don't know"

Machine Learning (ML) Speak

Data Transformation
Visual Exploration
Model Building

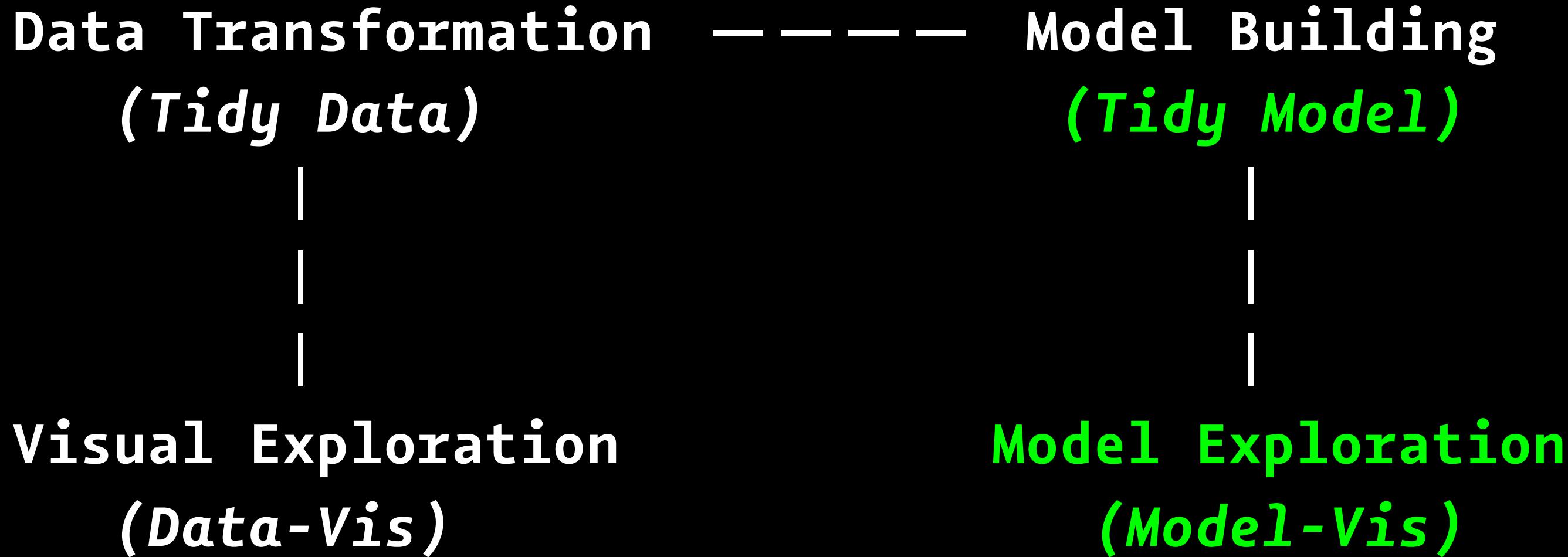
ML Pipeline

Data Transformation ----- Model Building
(Tidy Data)



Visual Exploration
(Data-Vis)

ML Pipeline++



Model-Vis Key Concept

Use visualisation to aid the transition of **implicit knowledge** in the data and your head to **explicit knowledge** in the model.

Model-Viz Approach

- [0] Visualise the **data space**
- [1] Visualise the **predictions in the data space**
- [2] Visualise the **errors in model fitting**
- [3] Visualise with **different model parameters**
- [4] Visualise with **different input datasets**
- [5] Visualise the **entire model space**
- [6] Visualise the **entire feature space**
- [7] Visualise the **many models together**

Model-Vis Examples

Regression: Small

Classification: Large p

Regression: Large n

Model-Vis Examples

Cars ($n < 50$, $p = 4$)

Digits ($n \sim 5K$, $p = 785$)

Taxi ($n \sim 10M$, $p = 20$)

Regression: Small

Cars dataset - **price** vs **kmpl**

Scraped from **comparison** website

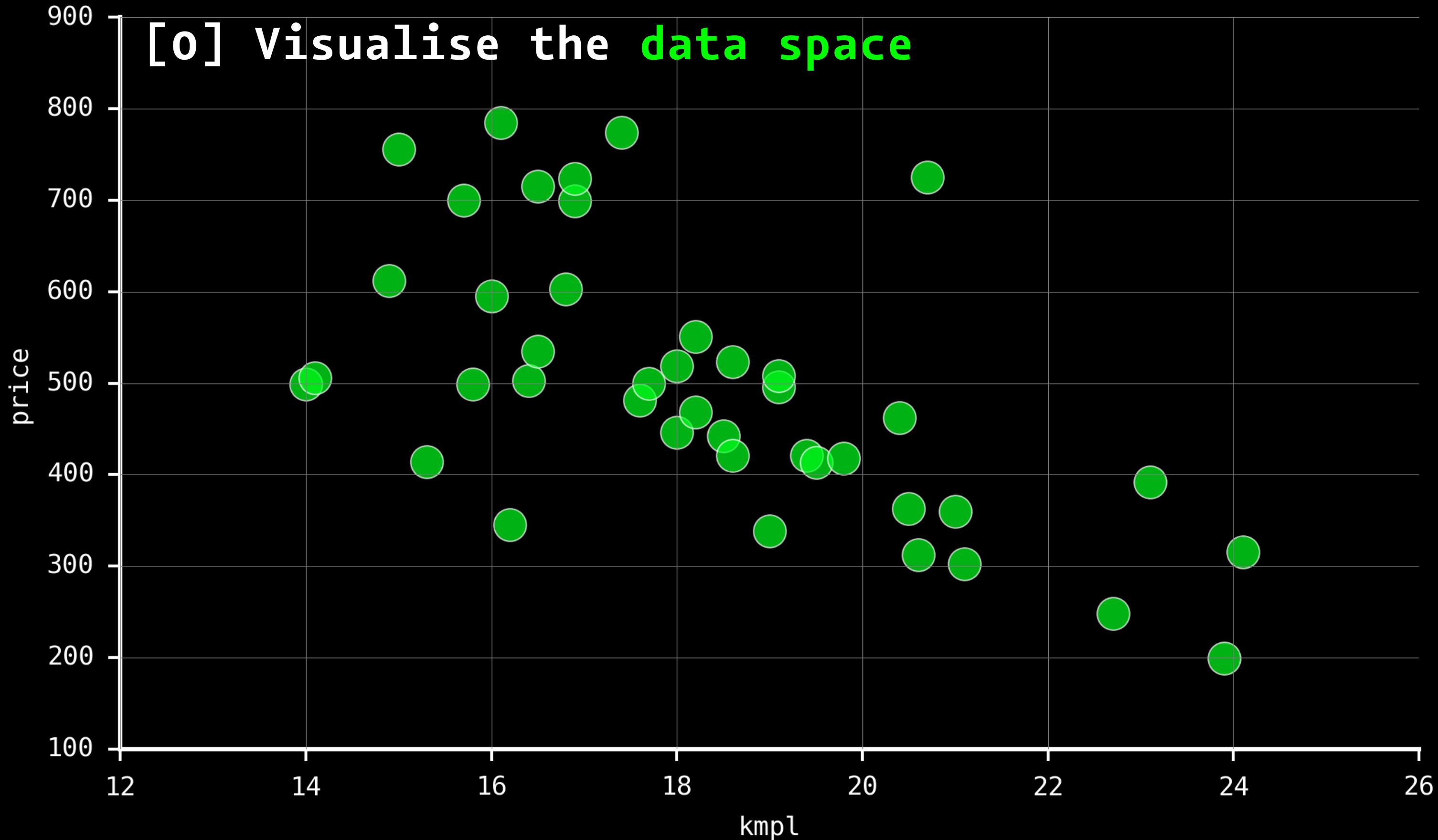
Refined & tidied up

Base version for **petrol** cars

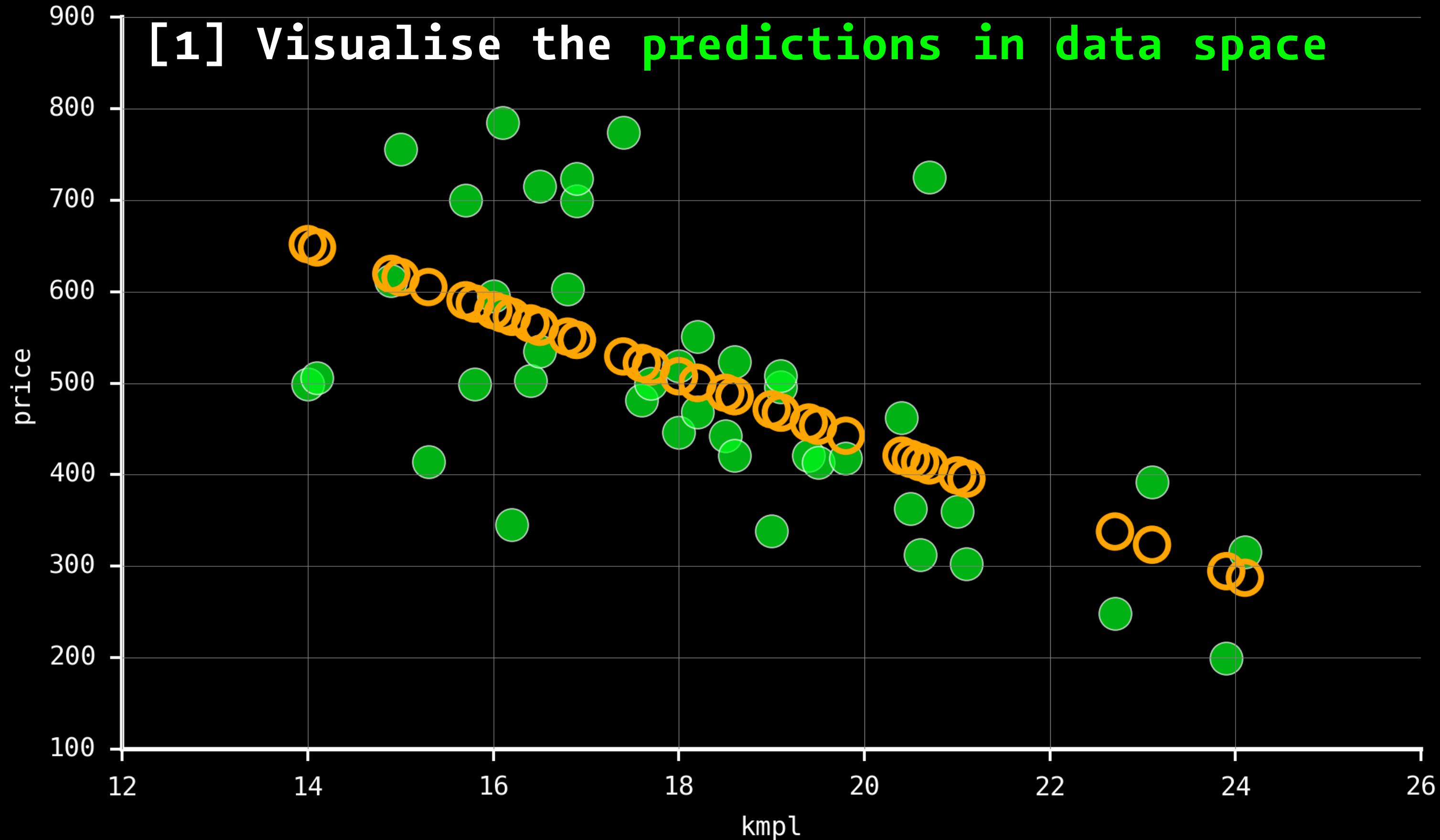
Price < ₹ 1,000K, n = 42

brand	model	price	kmpl	type	bhp
Tata	Nano	199	23.9	Hatchback	38
Suzuki	Alto800	248	22.7	Hatchback	47
Hyundai	EON	302	21.1	Hatchback	55
Nissan	Datsun	312	20.6	Hatchback	67
...
Suzuki	Ciaz	725	20.7	Sedan	91
Skoda	Rapid	756	15.0	Sedan	104
Hyundai	Verna	774	17.4	Sedan	106
VW	Vento	785	16.1	Sedan	104

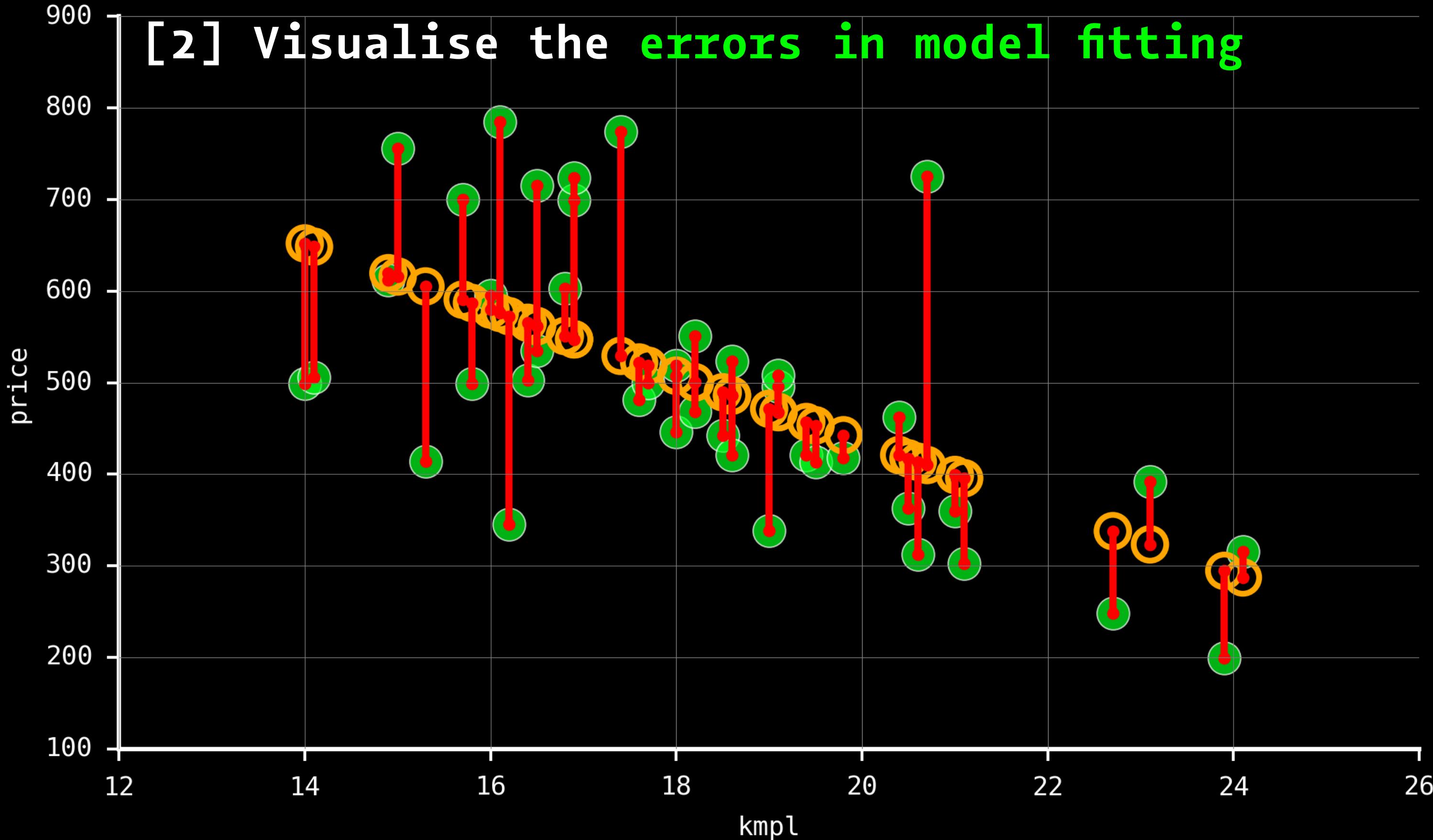
[o] Visualise the data space



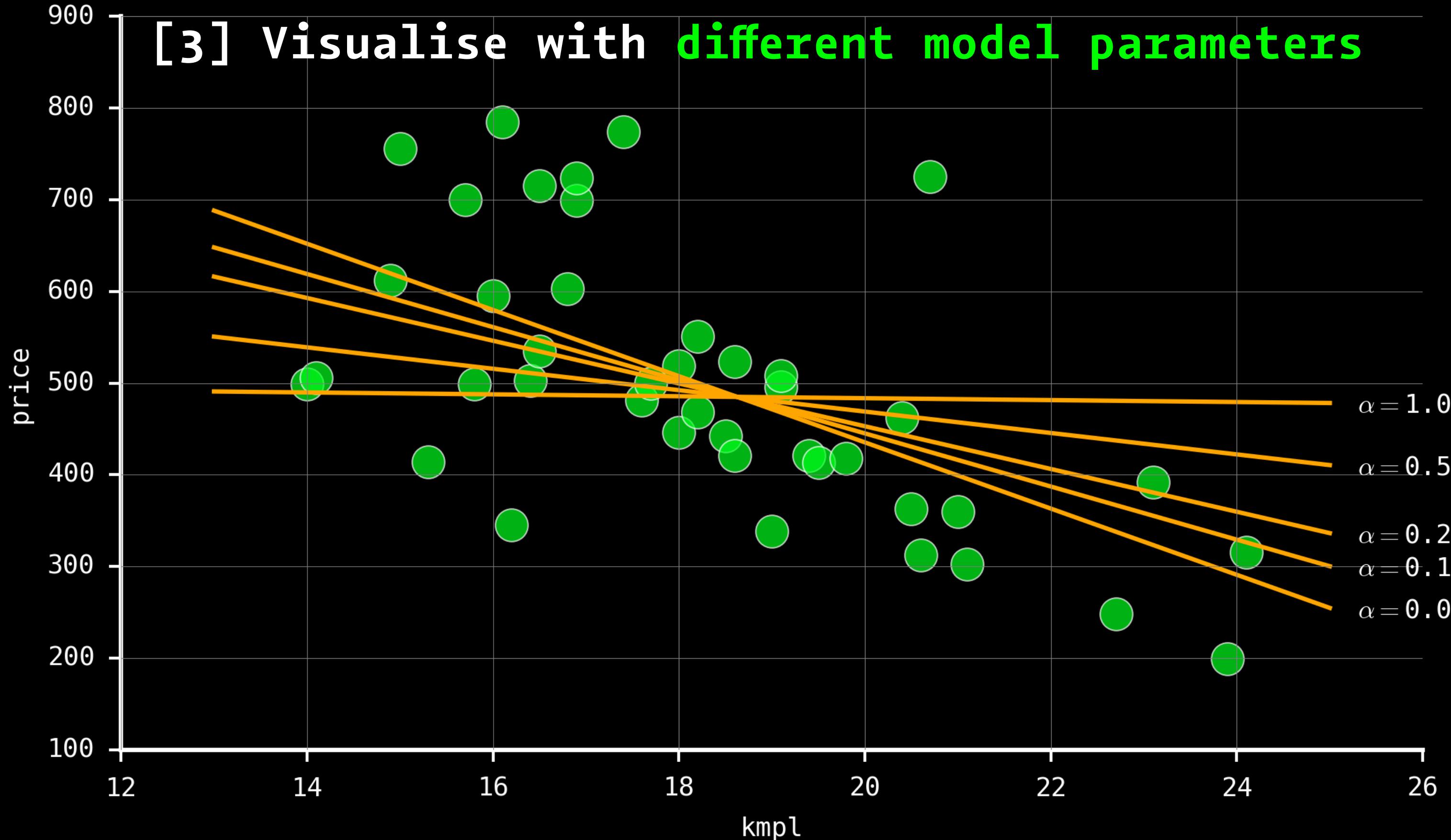
[1] Visualise the predictions in data space



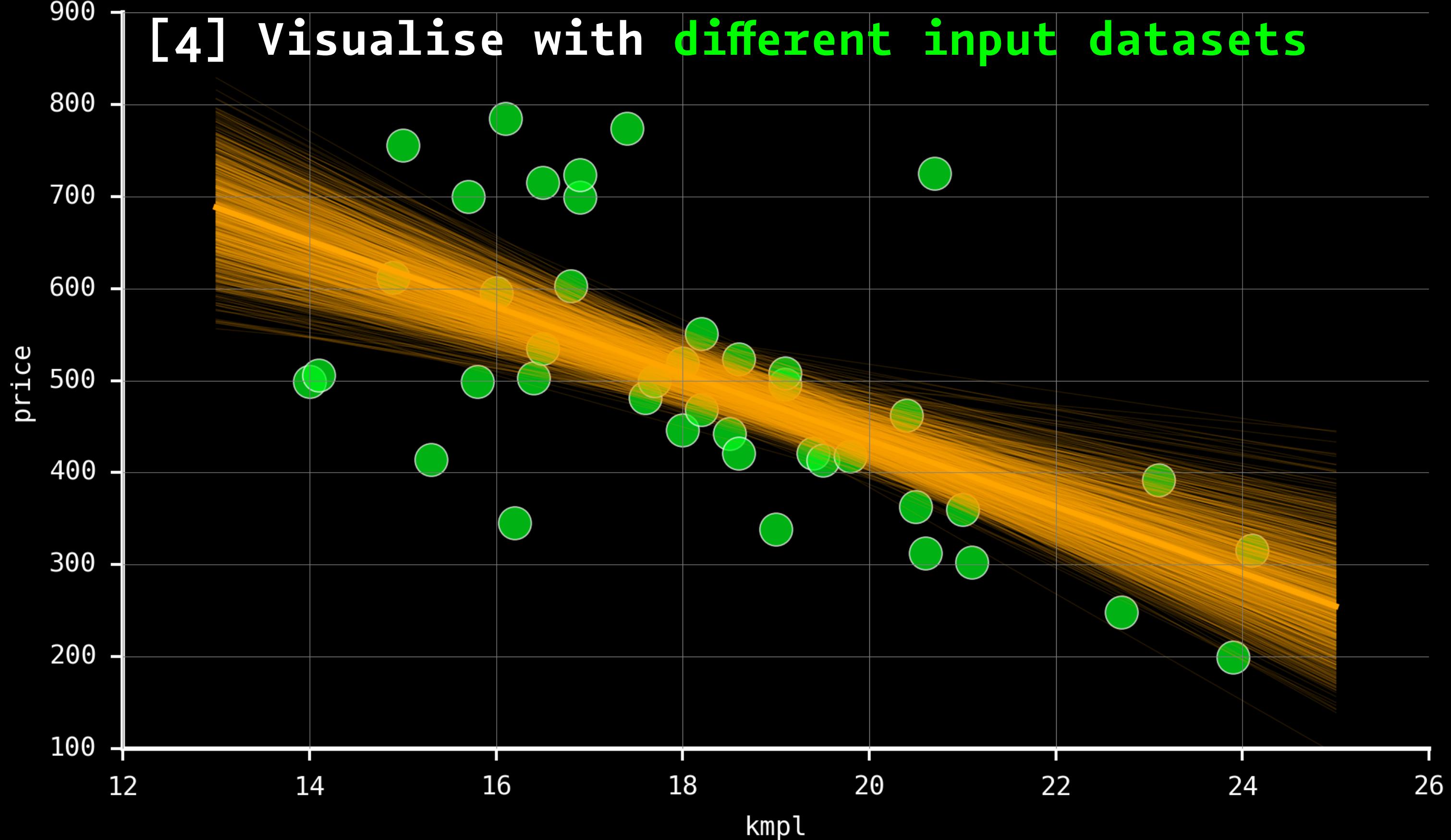
[2] Visualise the errors in model fitting



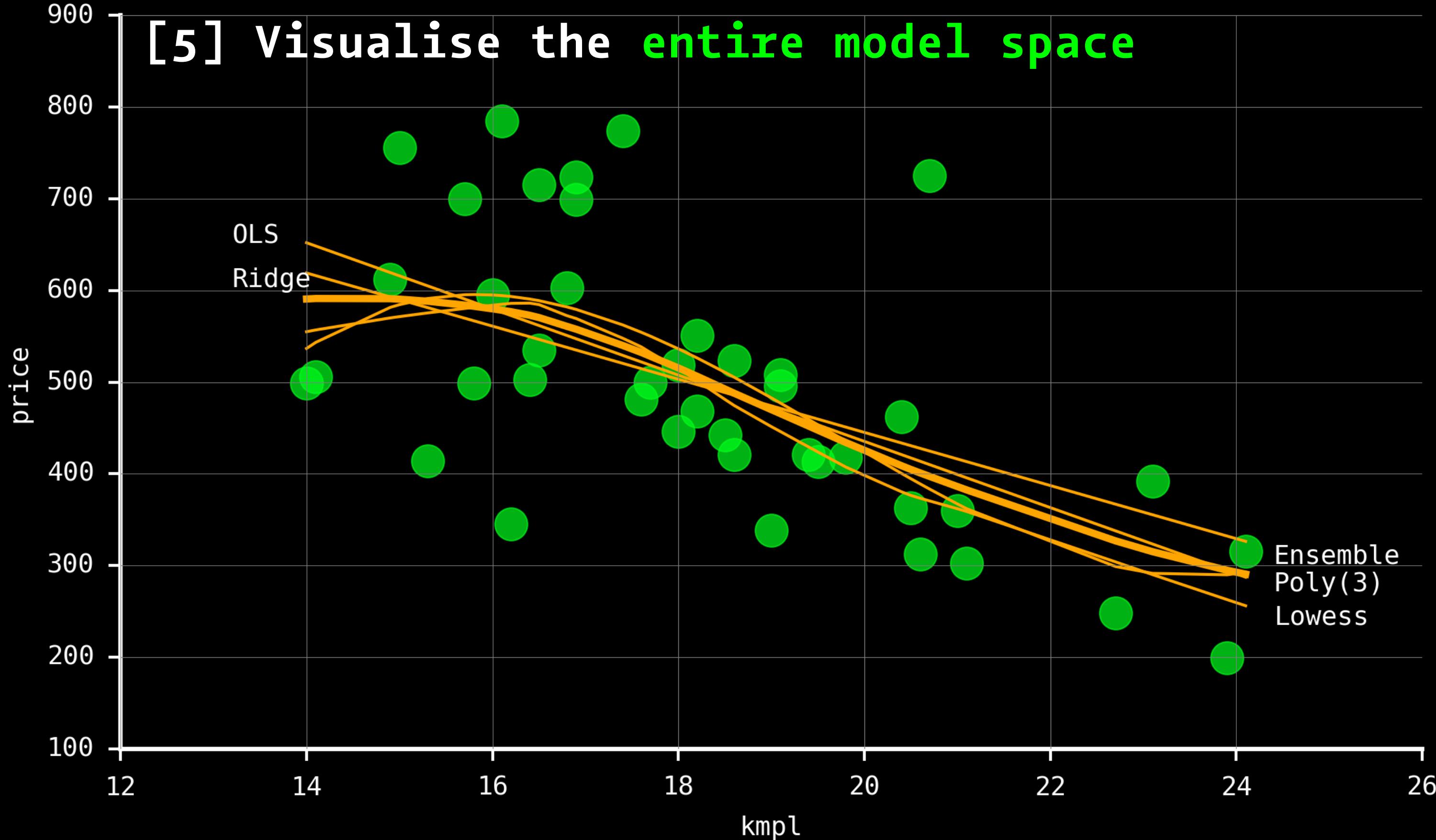
[3] Visualise with different model parameters



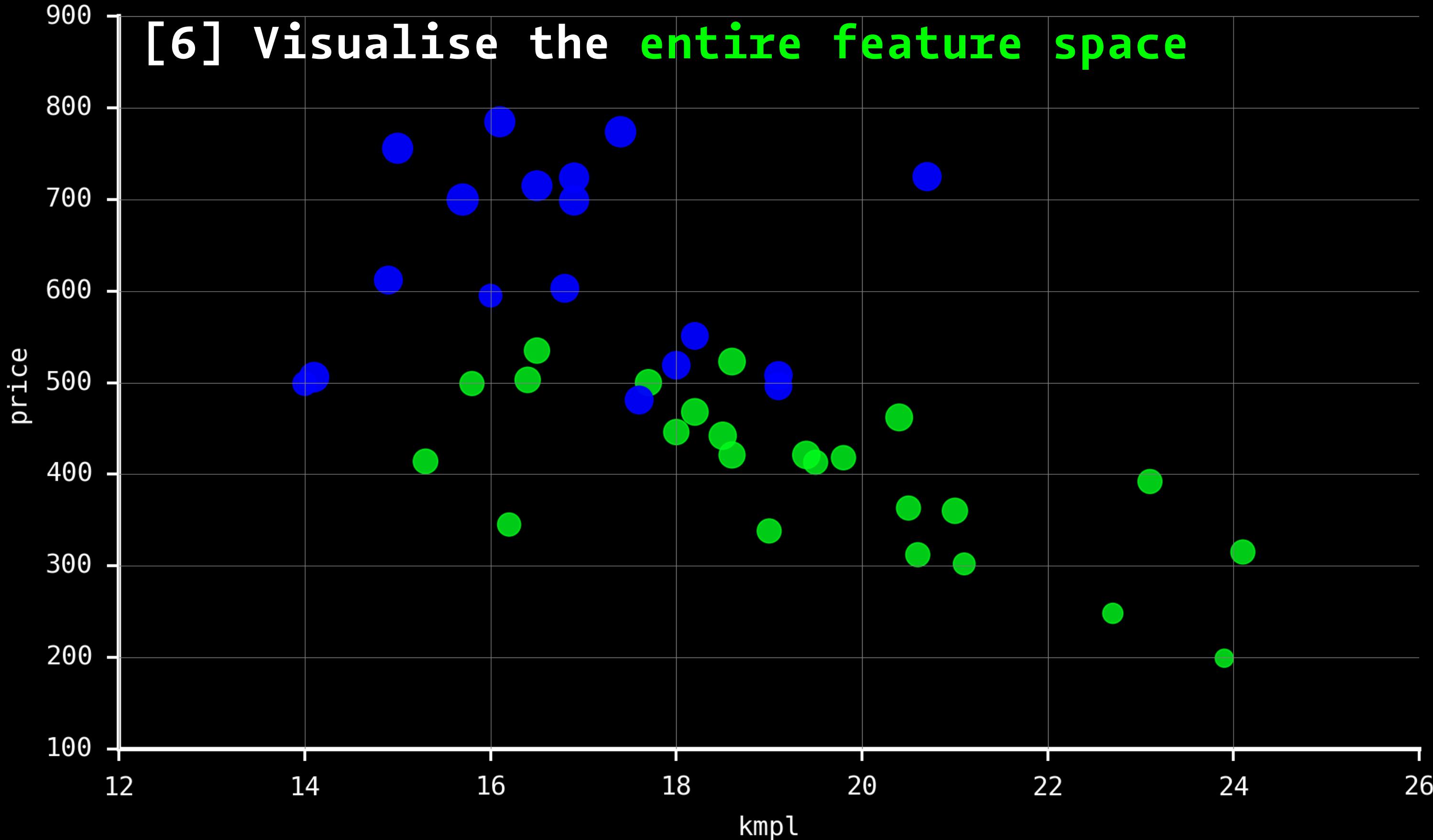
[4] Visualise with different input datasets



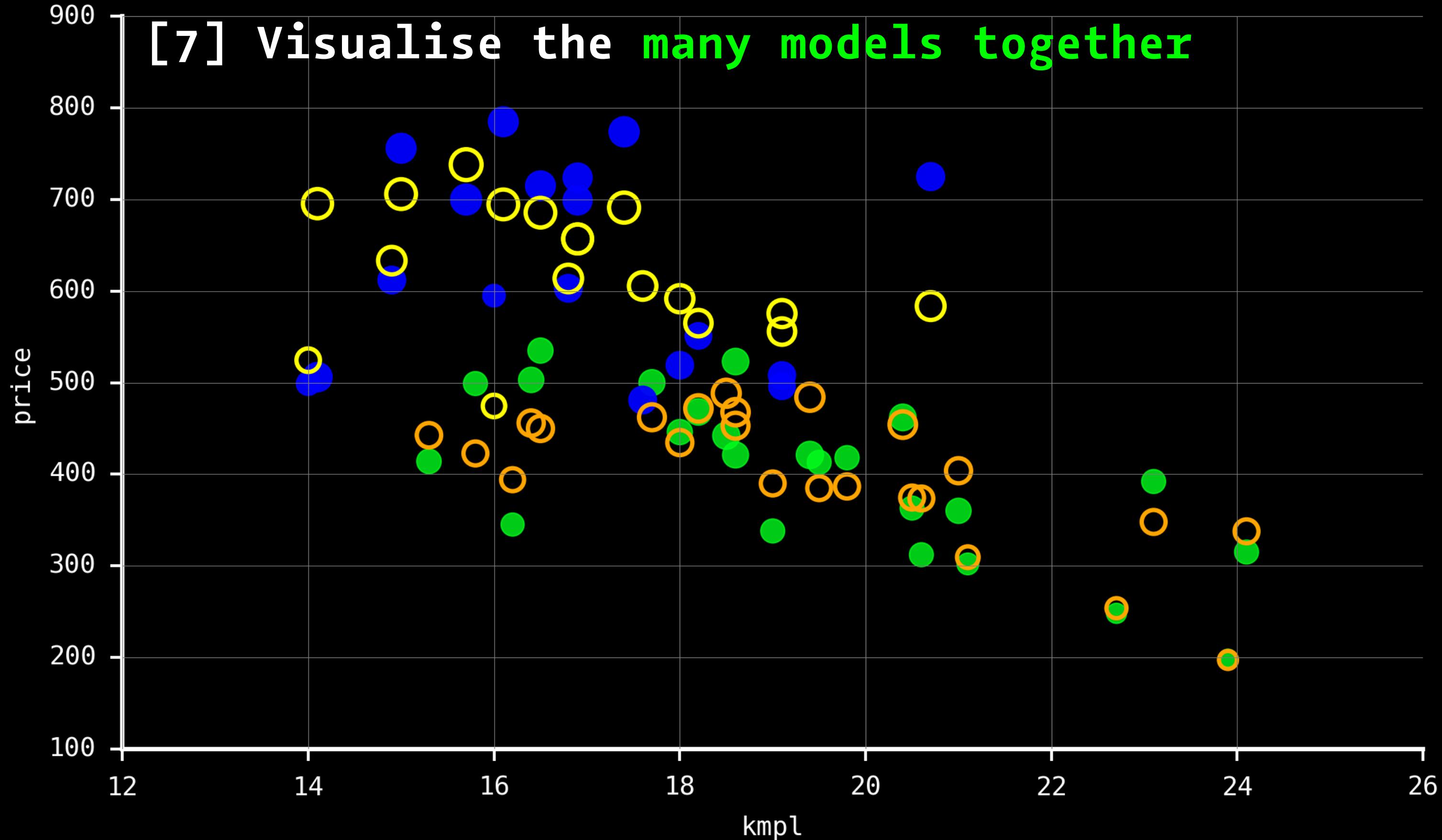
[5] Visualise the entire model space



[6] Visualise the entire feature space



[7] Visualise the many models together



Model-Viz Approach

- [0] Visualise the **data space**
- [1] Visualise the **predictions in the data space**
- [2] Visualise the **errors in model fitting**
- [3] Visualise with **different model parameters**
- [4] Visualise with **different input datasets**
- [5] Visualise the **entire model space**
- [6] Visualise the **entire feature space**
- [7] Visualise the **many models together**

Model-Vis & ML Approach

- [0] **DATA VIS**: the data space
- [1] **PREDICTION**: the predictions in the data space
- [2] **VALIDATION**: the errors in model fitting
- [3] **TUNING**: with different model parameters
- [4] **BOOTSTRAP**: with different input datasets
- [5] **ENSEMBLE**: the entire model space
- [6] **FEATURE ENGG**: the entire feature space
- [7] **N-MODELS**: the many models together

Move through Layers

Iterative, not linear

Up and Down, not lateral

Complementary, not exclusive

p/n/N Model-Vis challenge

p -- High dimensional data

n -- Large and big data

N -- Multiple models

Classification: 2 Class

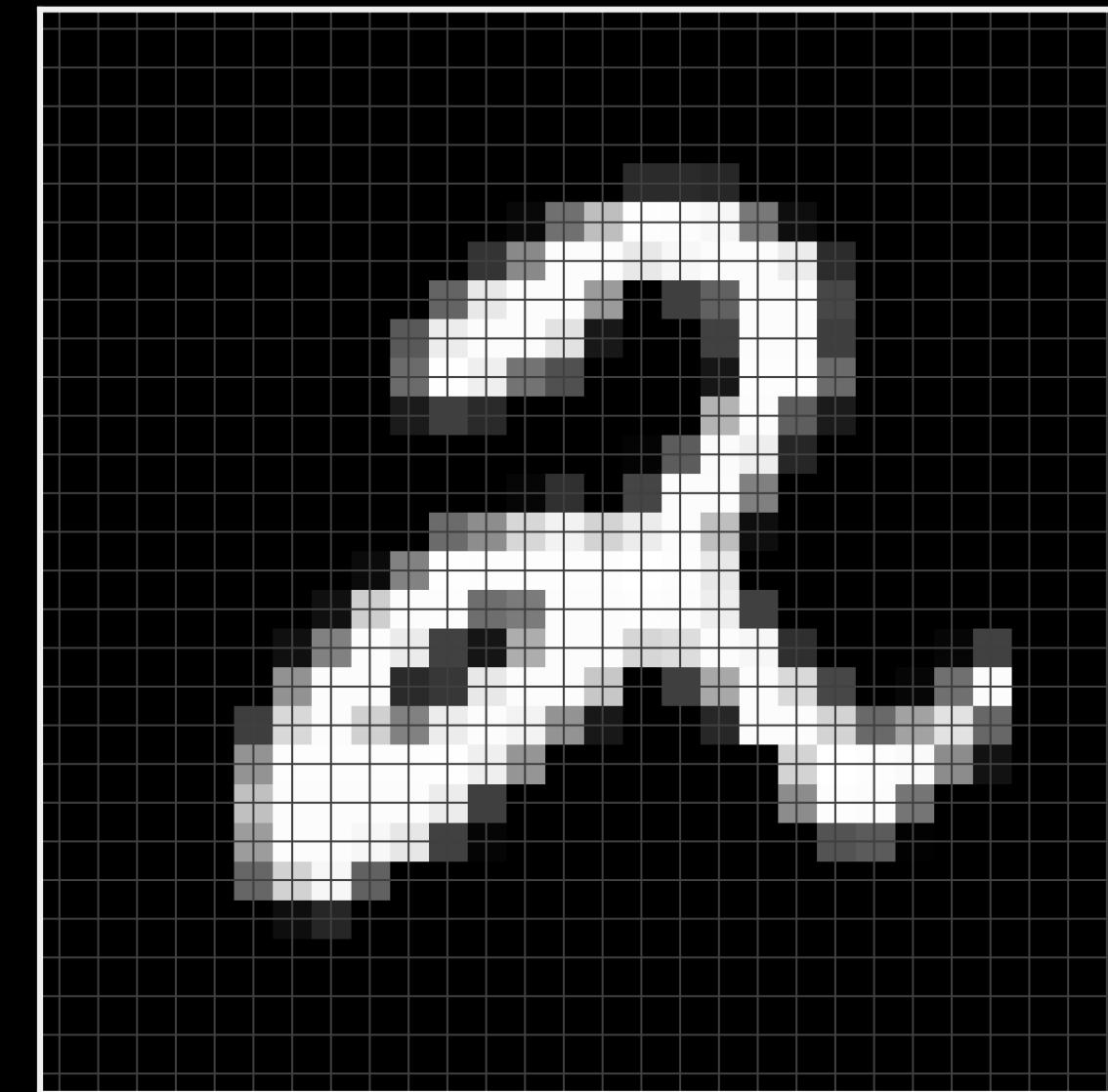
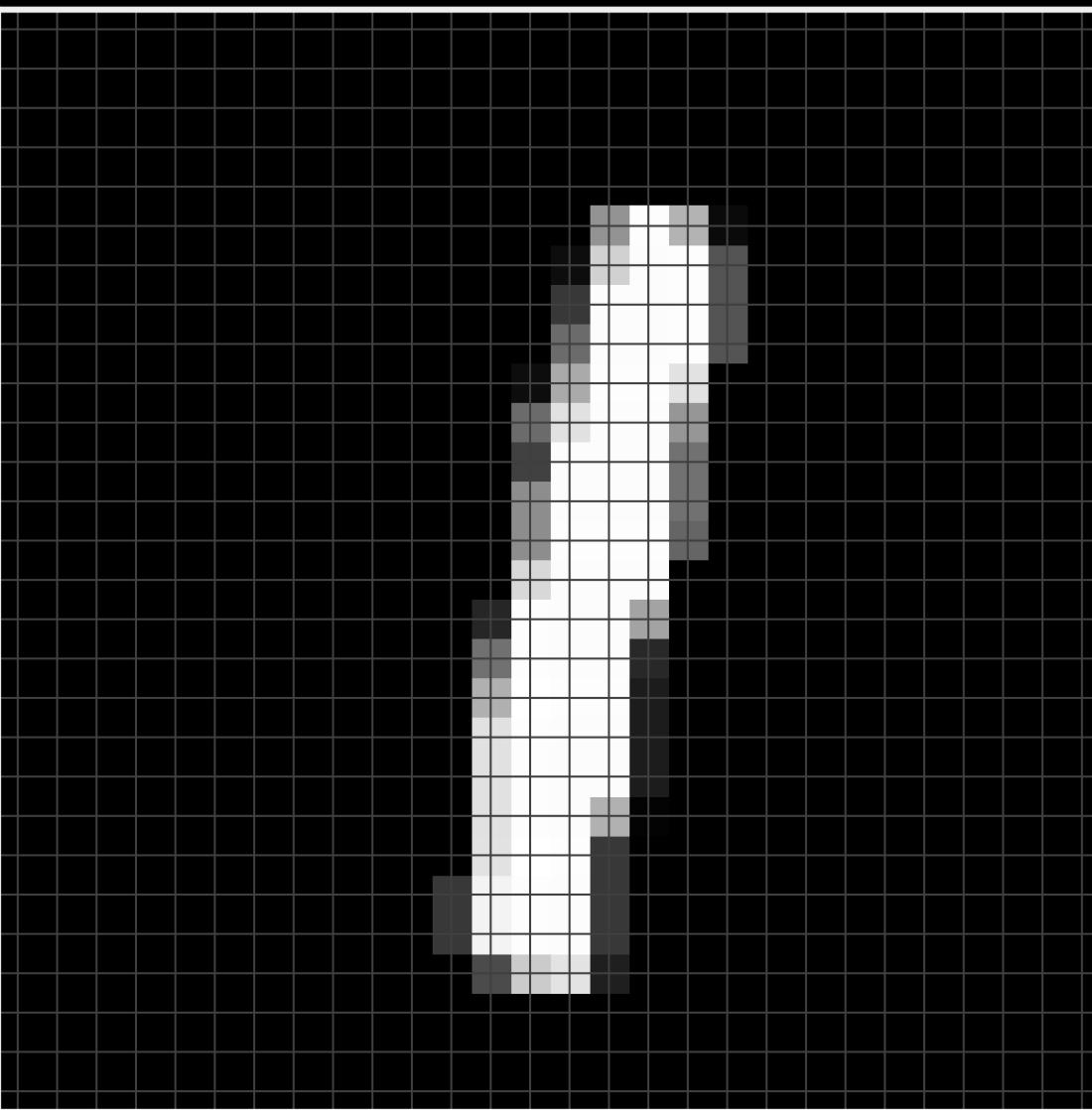
MNIST - digit recognition

Reduced to 2-class: 1 and 2

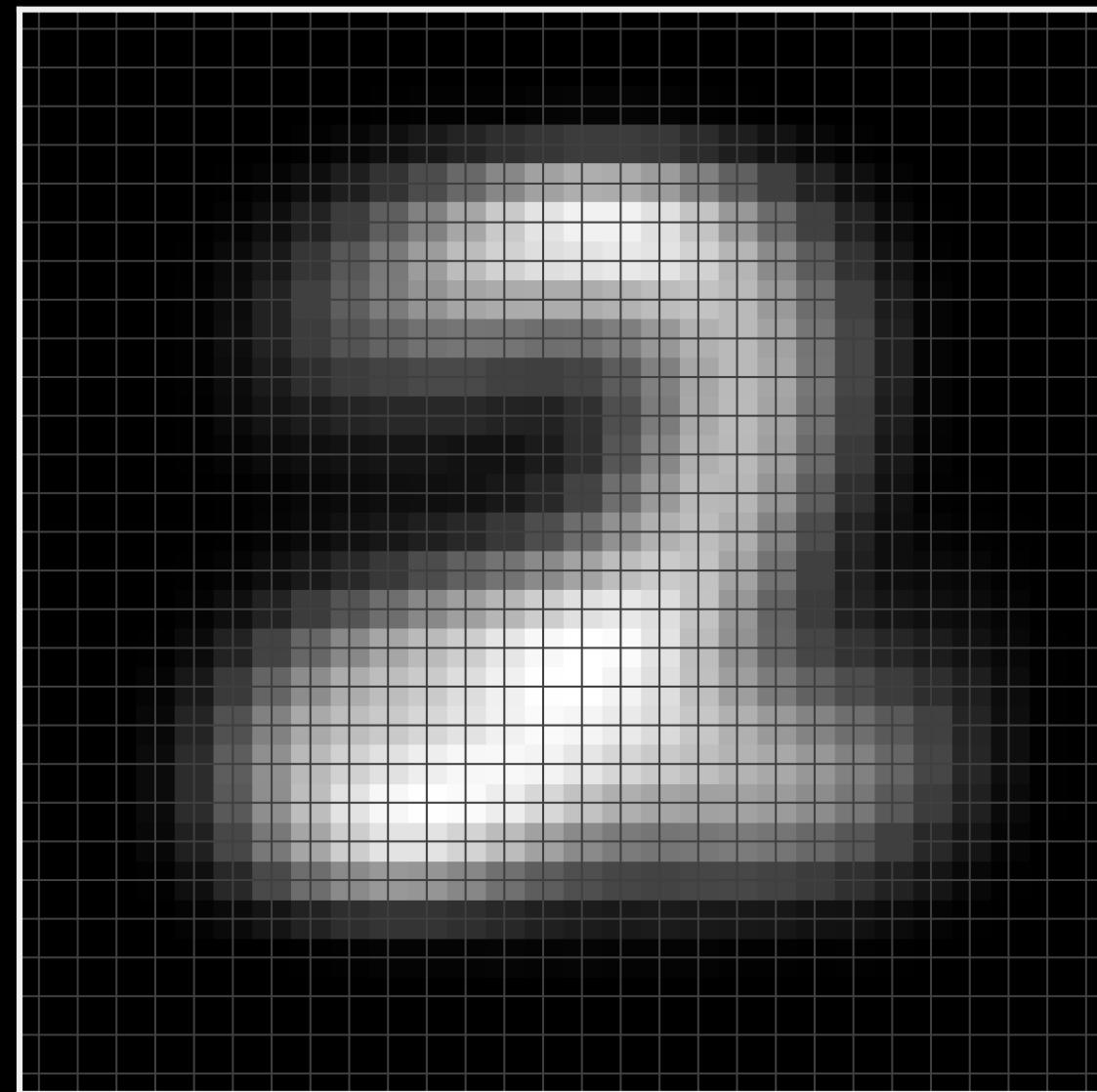
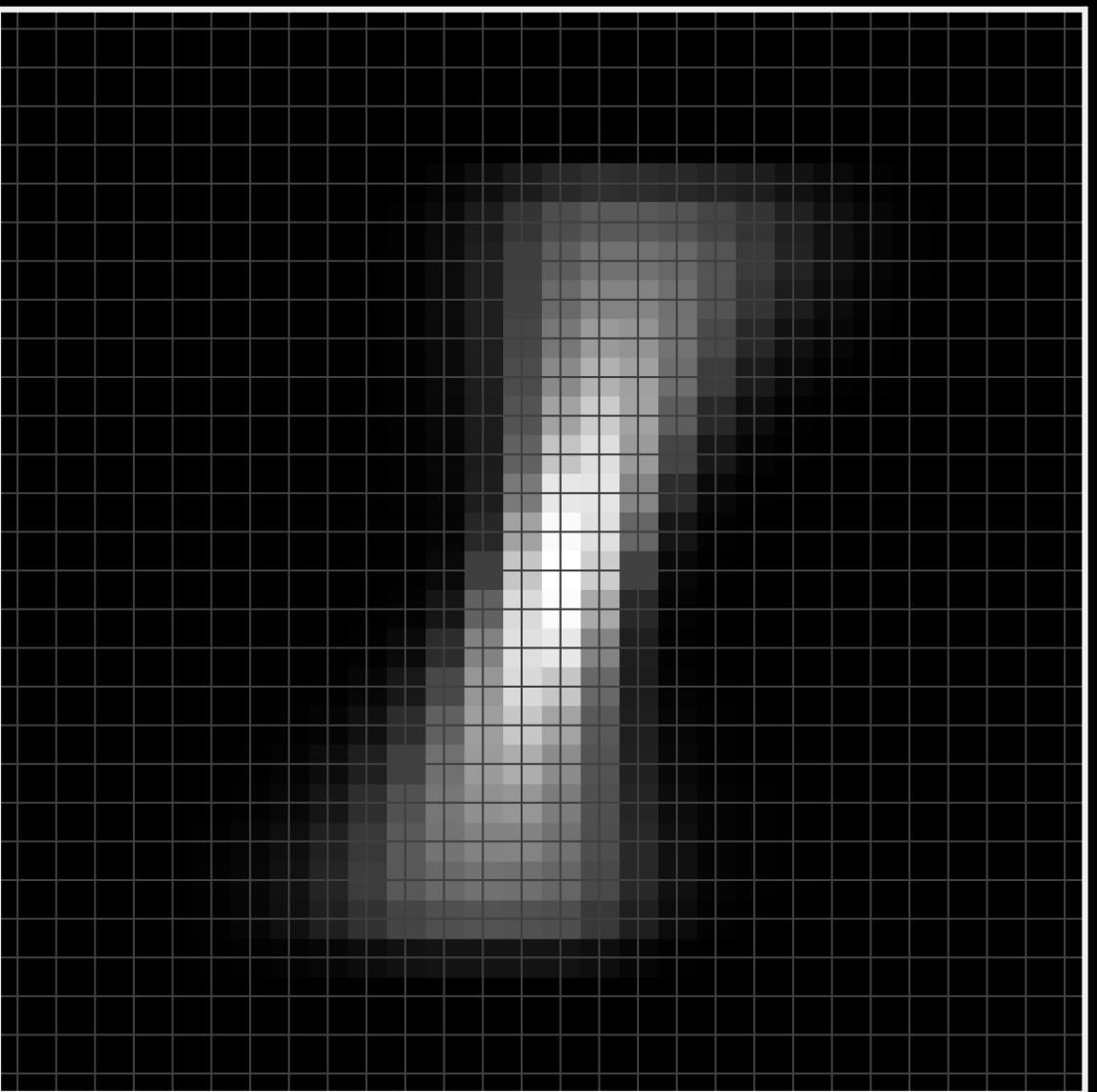
$p = 784$, 28 x 28 gray pixel map

$n > 5000$

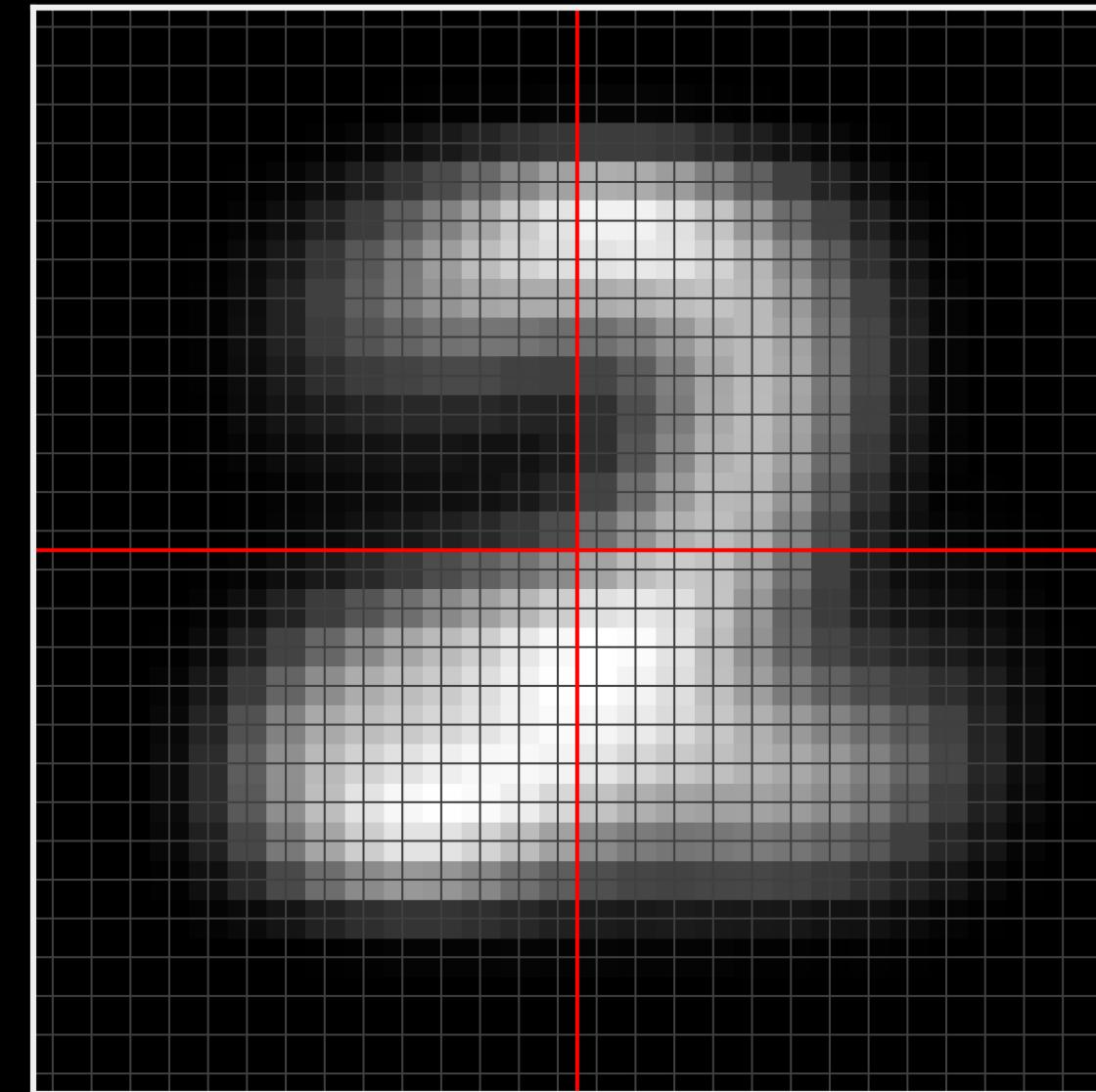
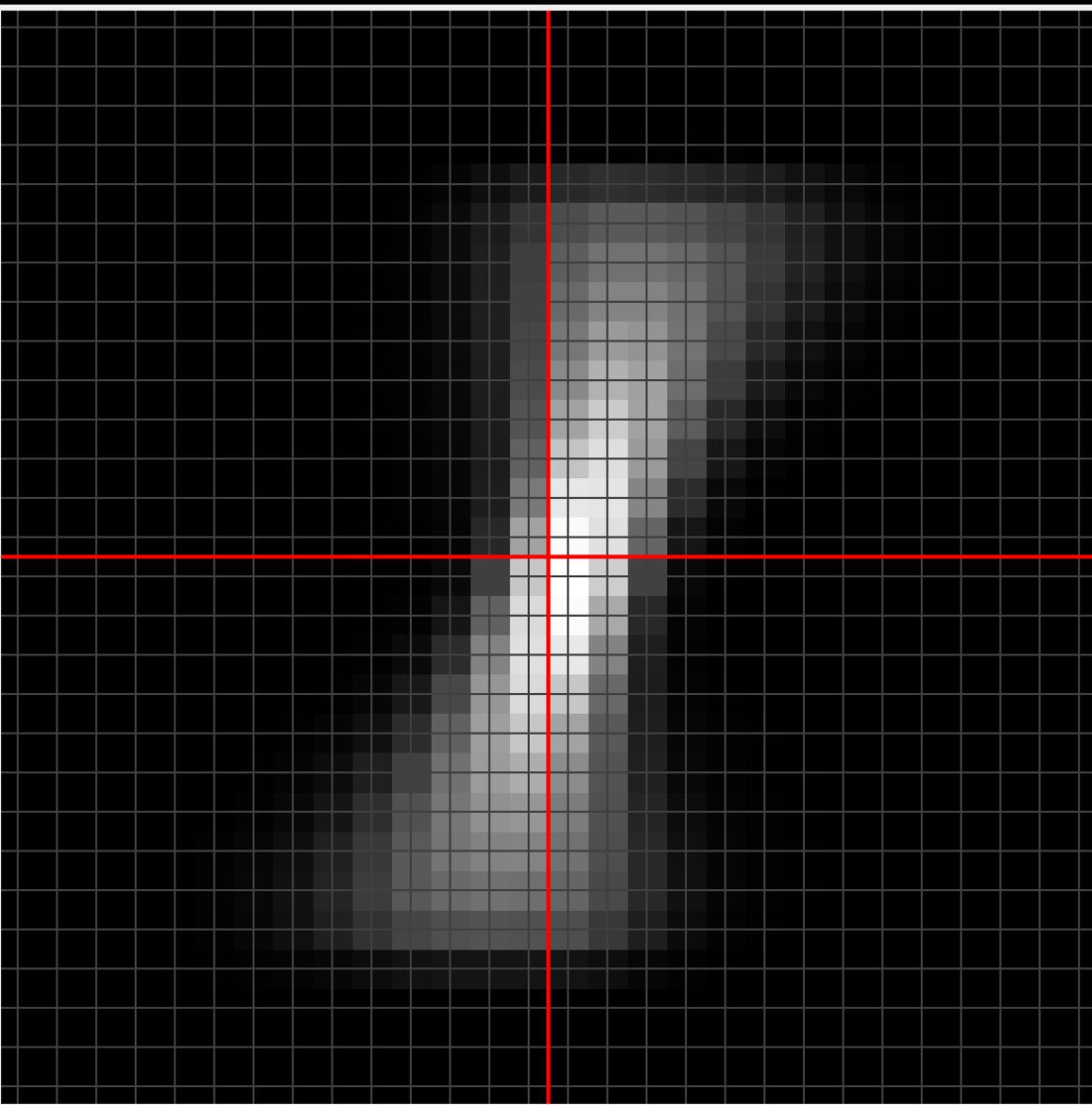
MNIST dataset: Examples of number 1 and 2



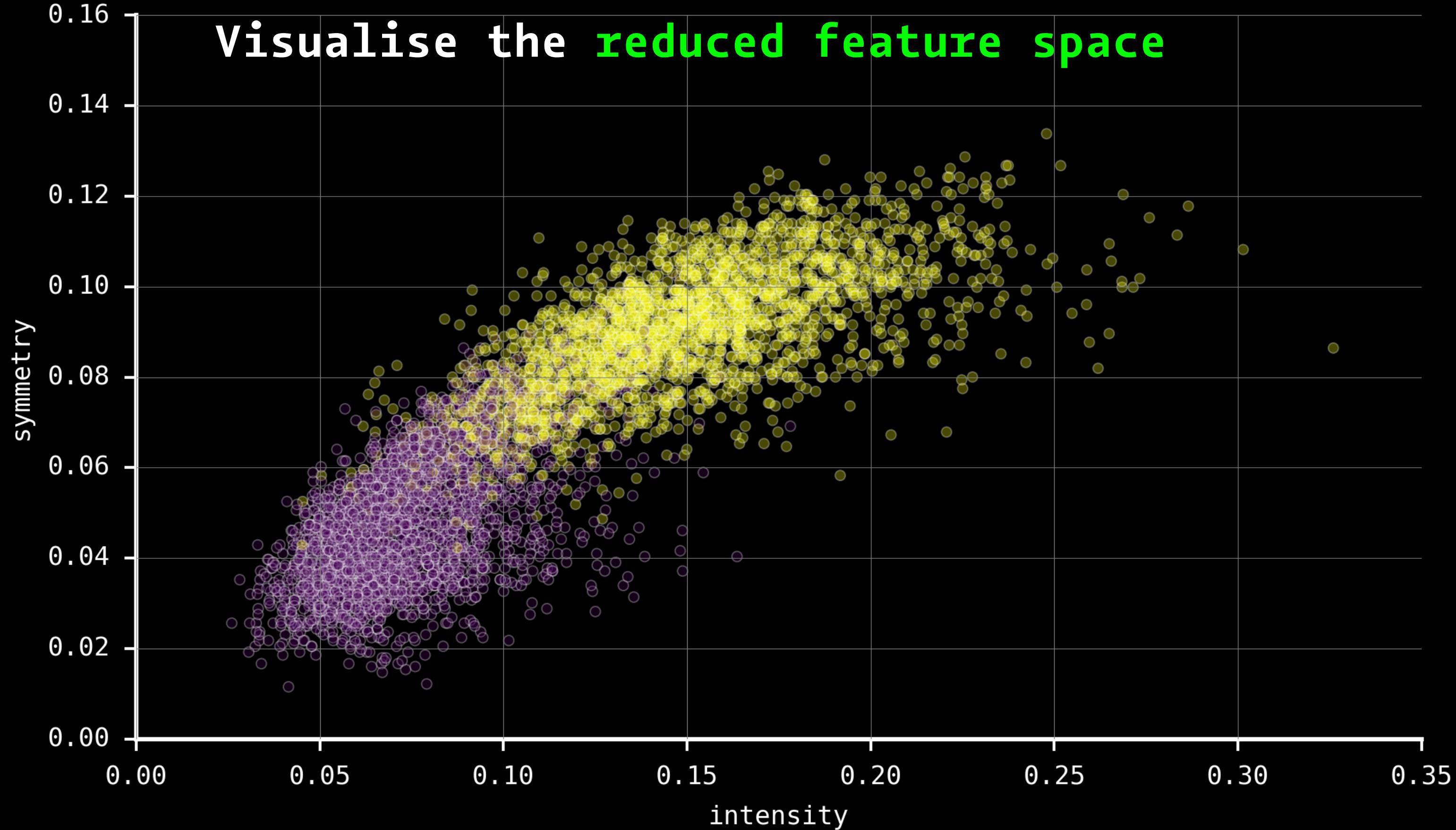
Visualise the data space



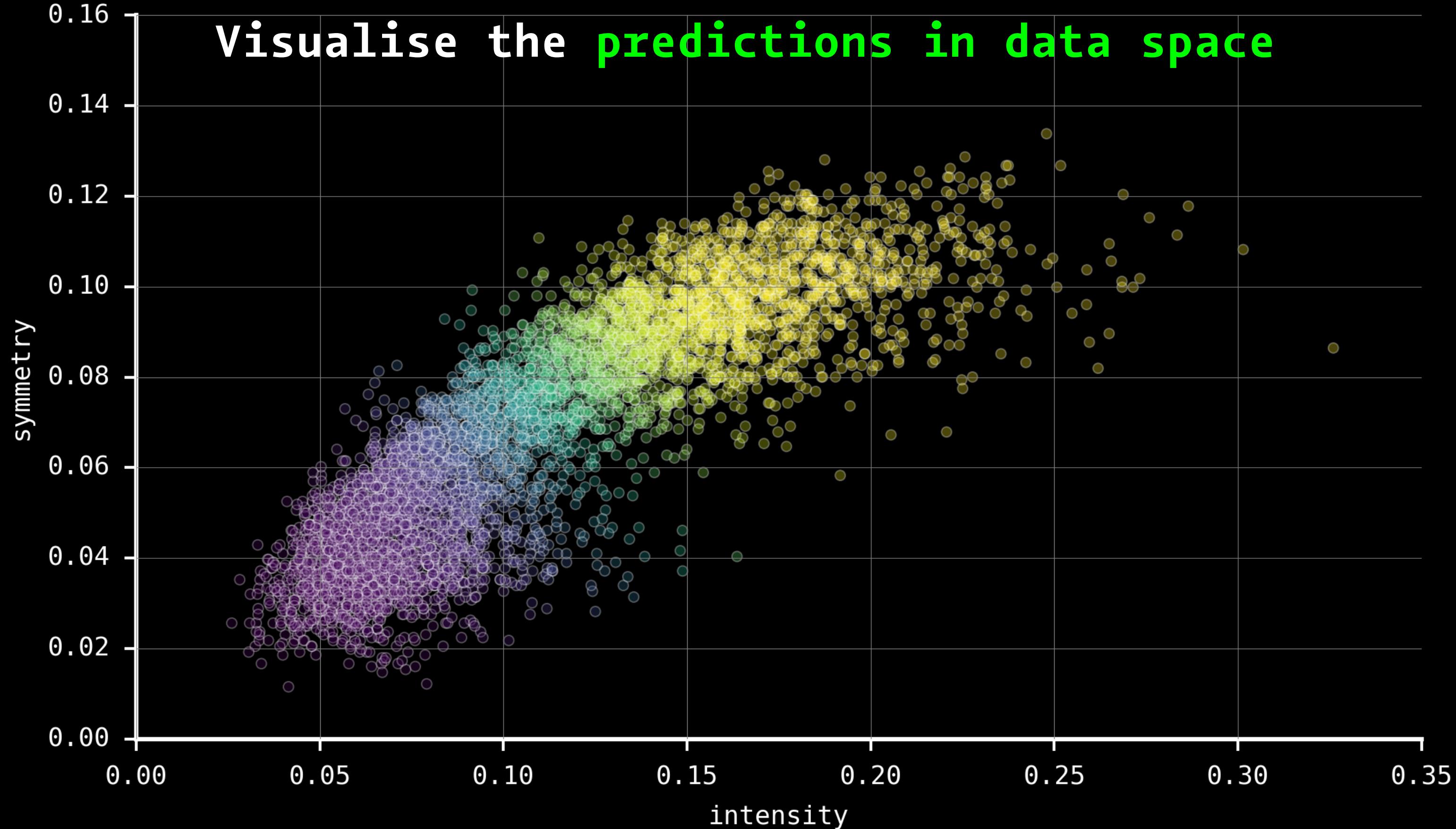
Identify the features - Symmetry & Intensity



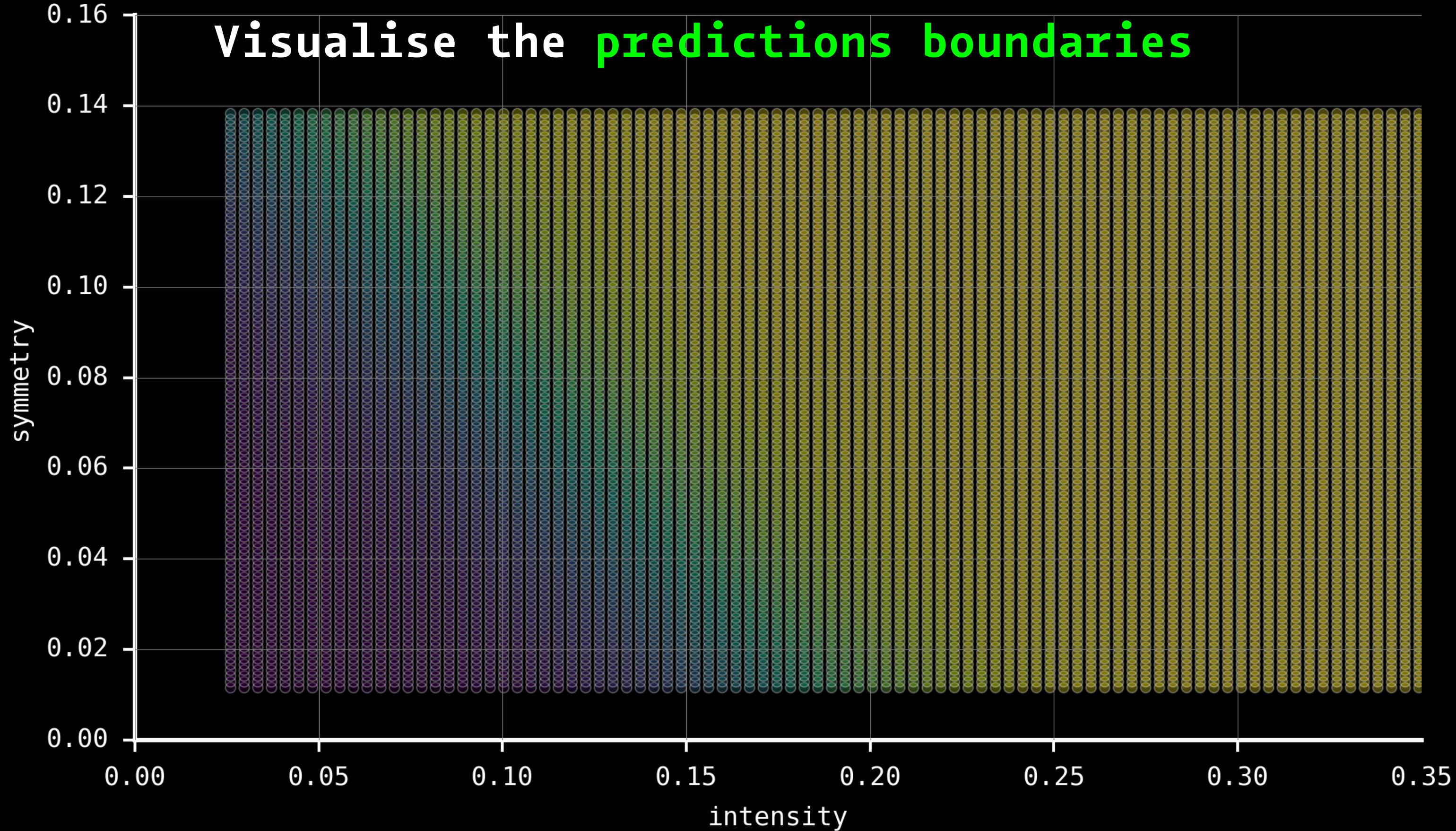
Visualise the reduced feature space



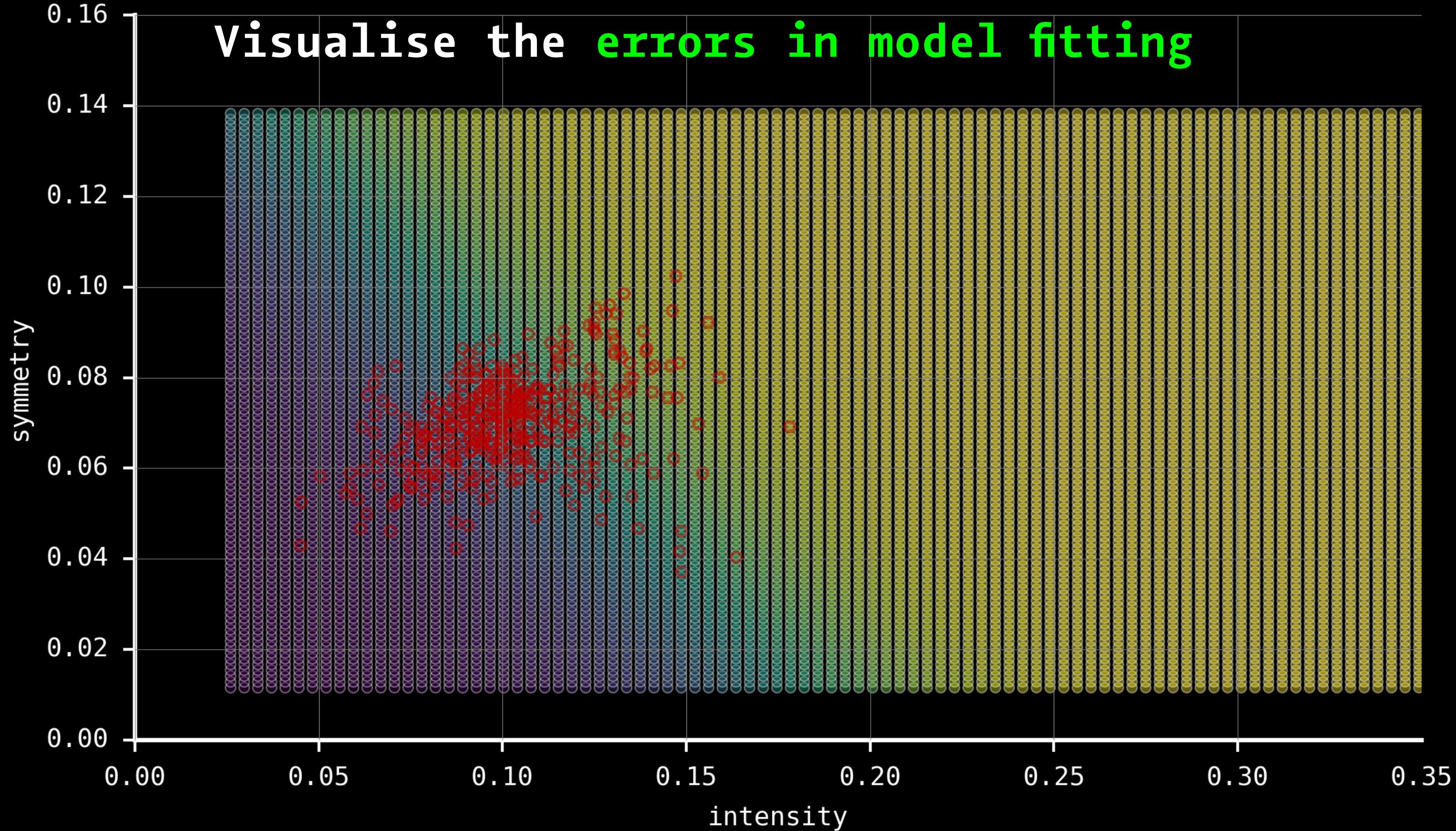
Visualise the predictions in data space



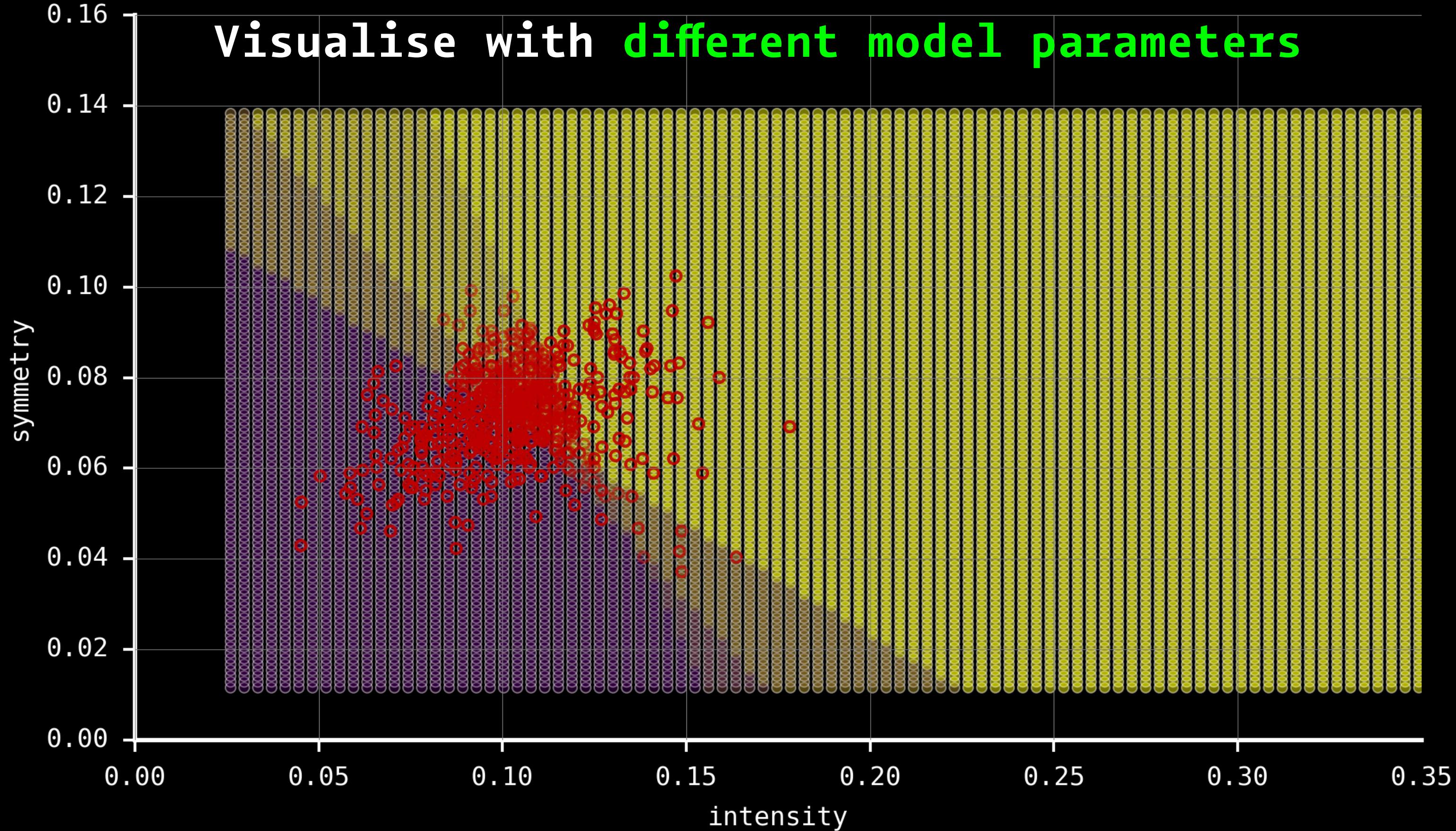
Visualise the predictions boundaries



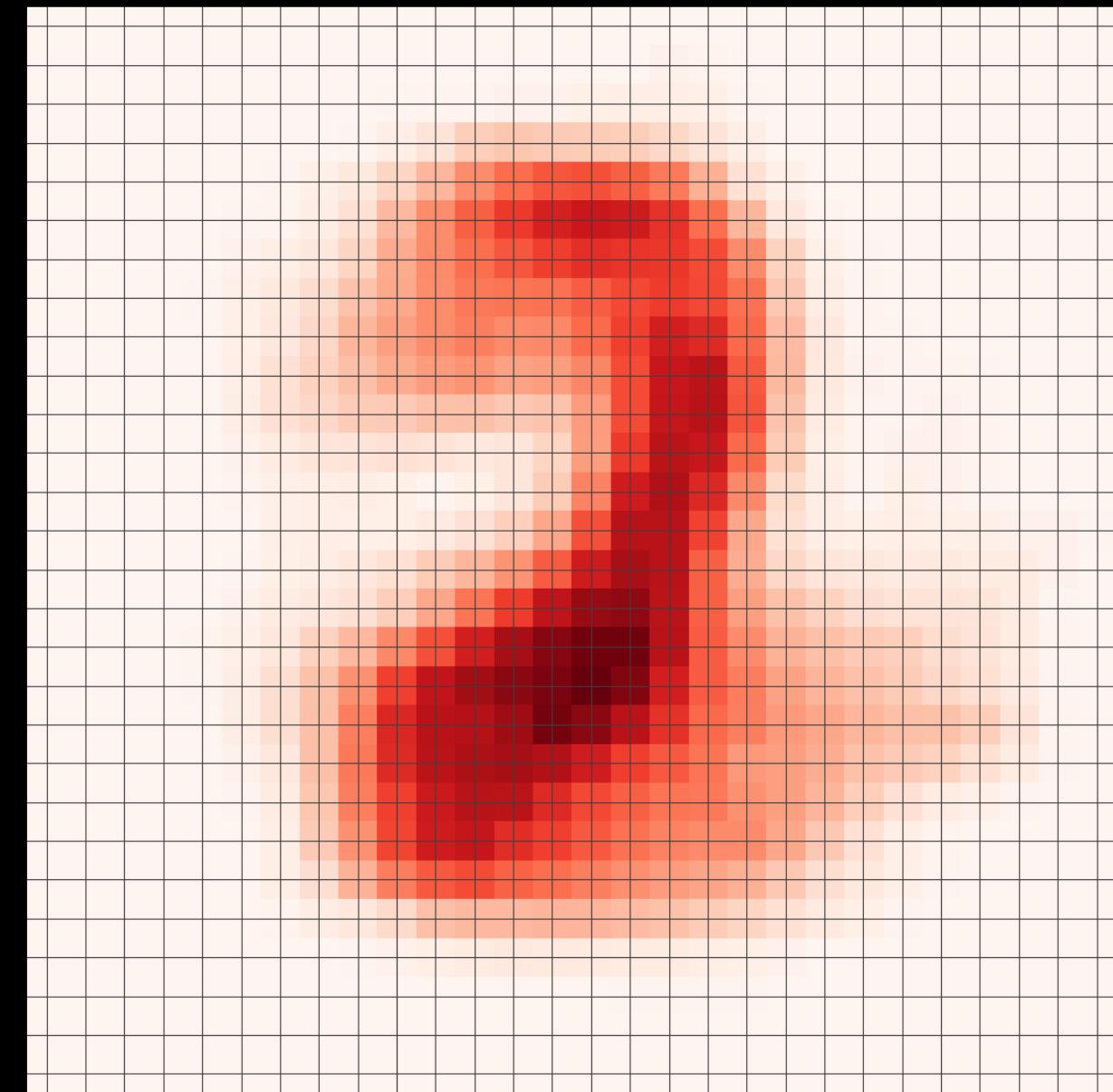
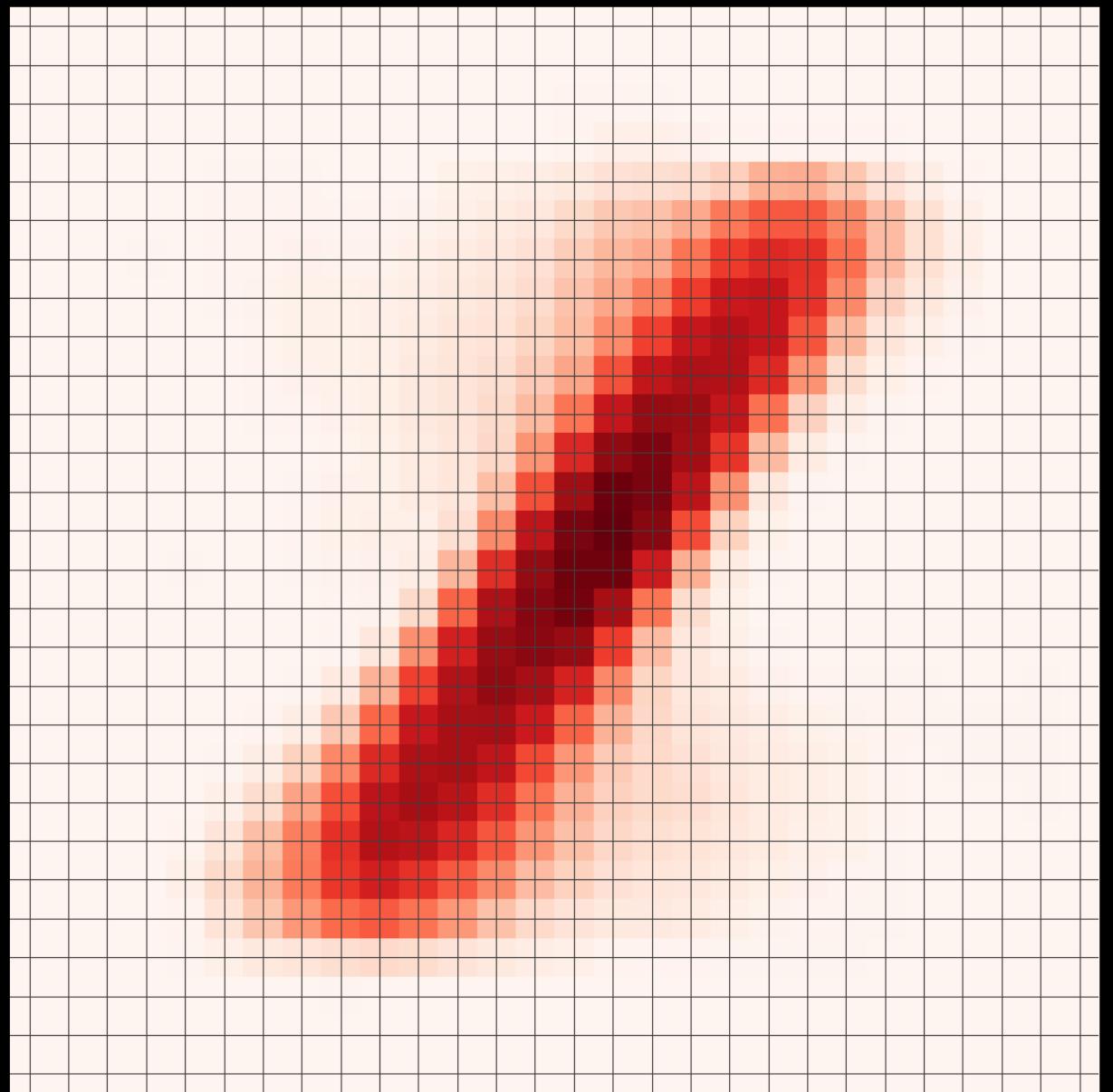
Visualise the errors in model fitting



Visualise with different model parameters



Easy to visualise errors in data space



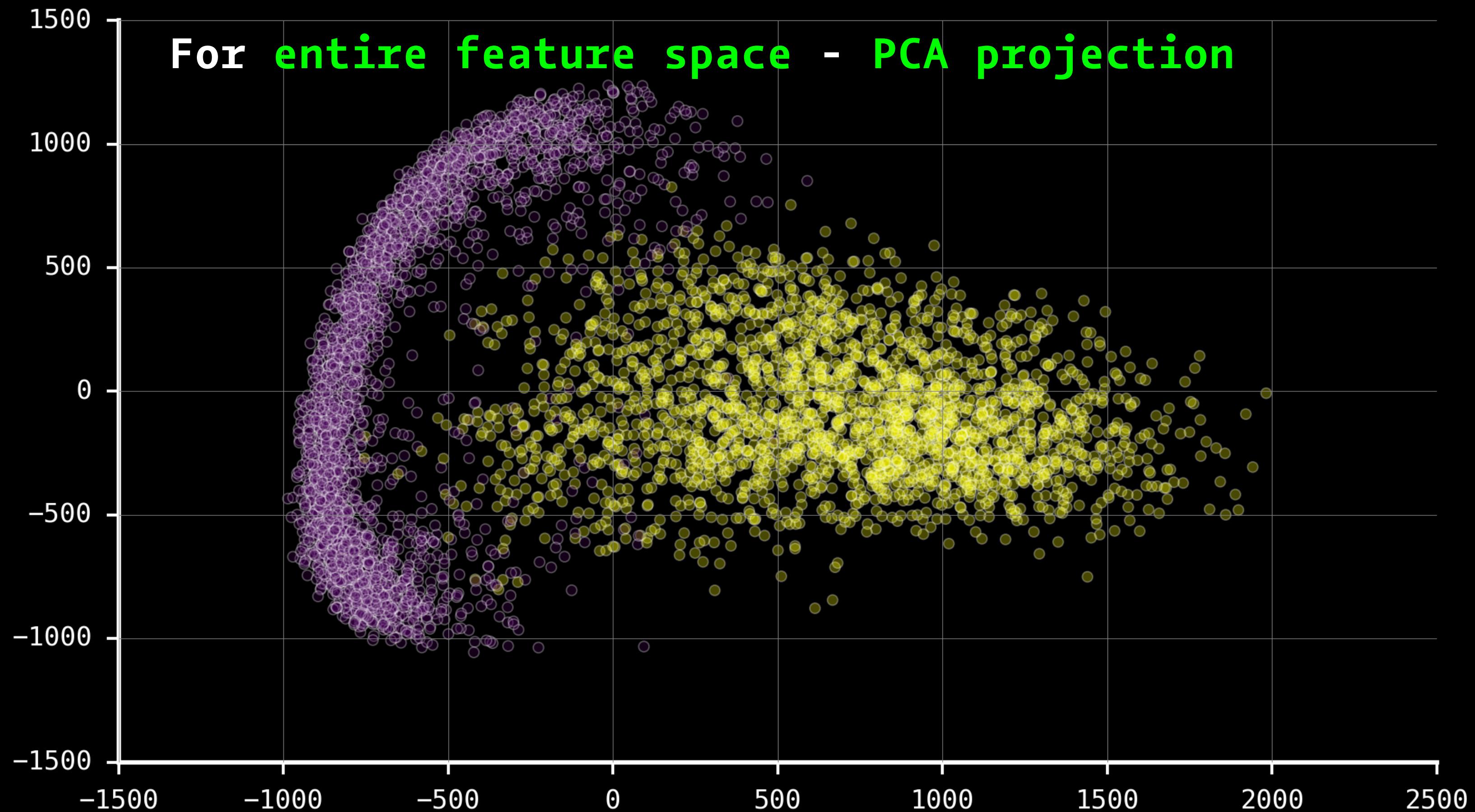
How to scale for large p ?

Curse of dimensionality

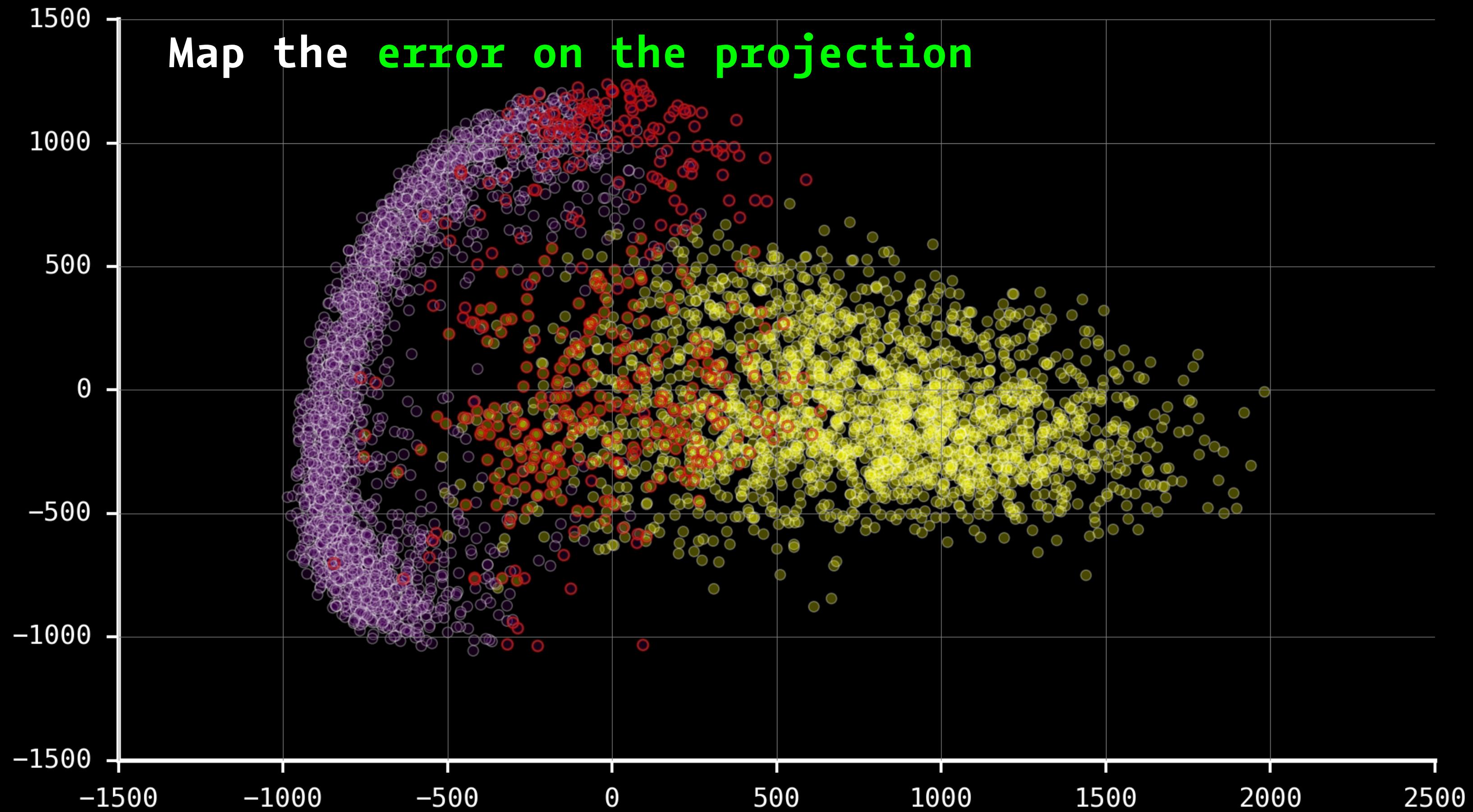
Mesh approach **computationally expensive**

Need to use **projections**

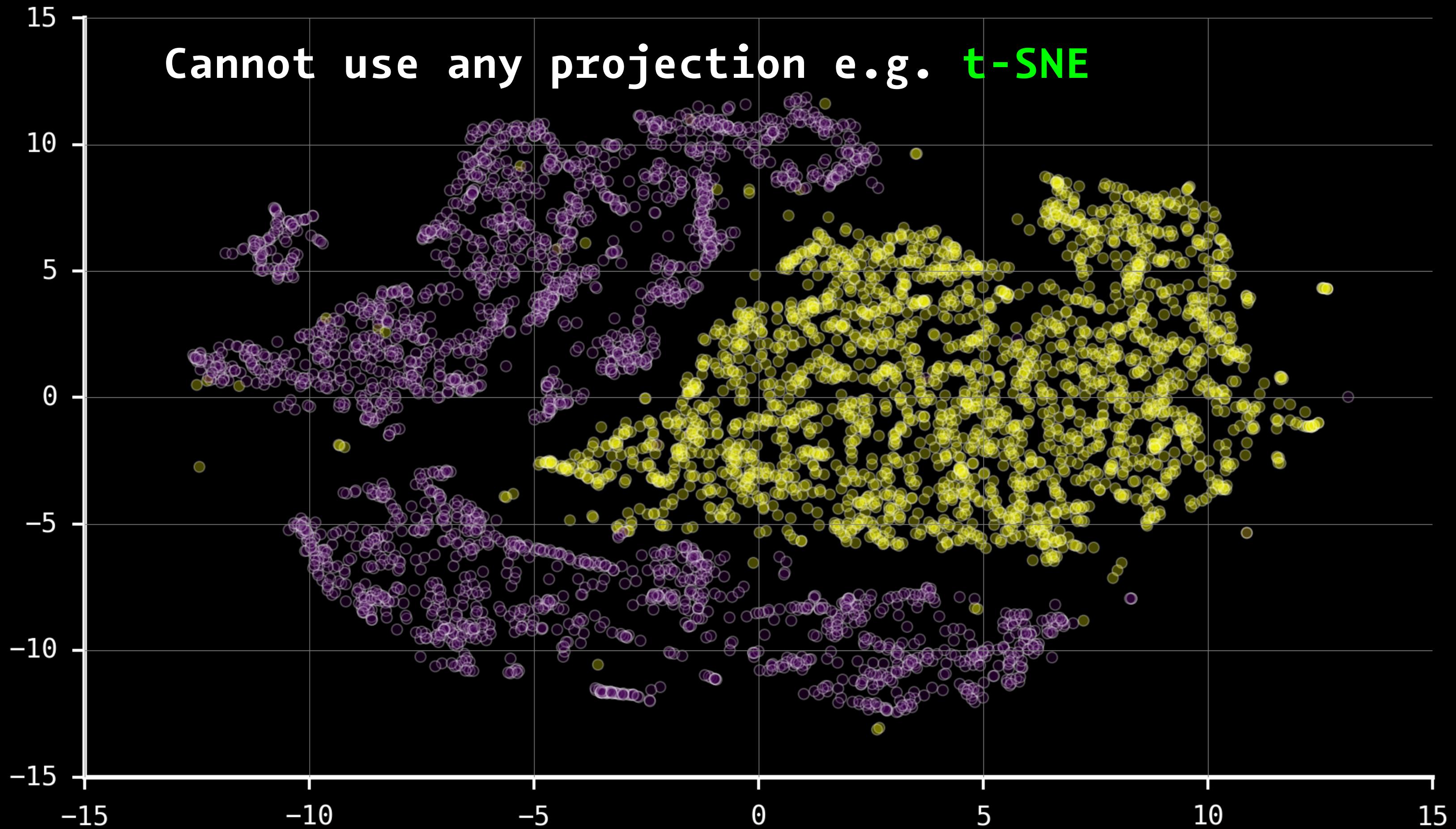
For entire feature space - PCA projection



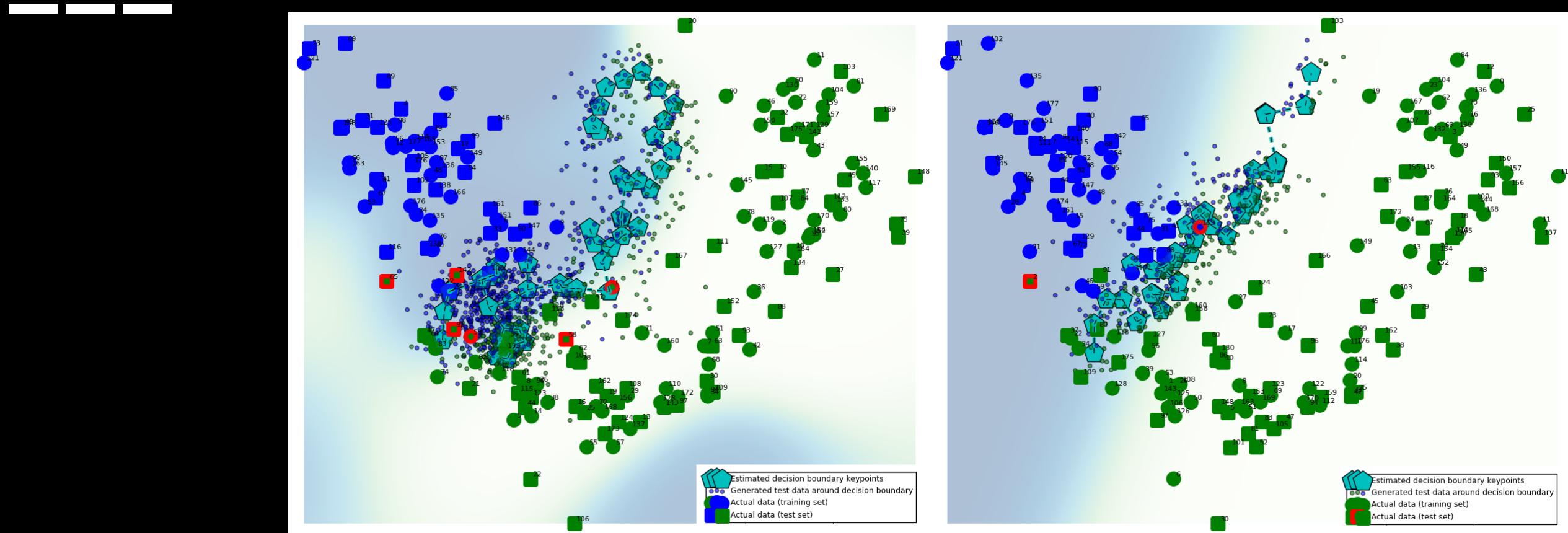
Map the error on the projection



Cannot use any projection e.g. t-SNE



High-p Boundary Classifiers



Github: [highdimensional-decision-boundary-plot](#)

Regression: Large n

NYC Taxi Trip Data

n ~ 10M (in just one month)

p = 20, geo location (drop & pick up), fare breakup, passenger no. etc.

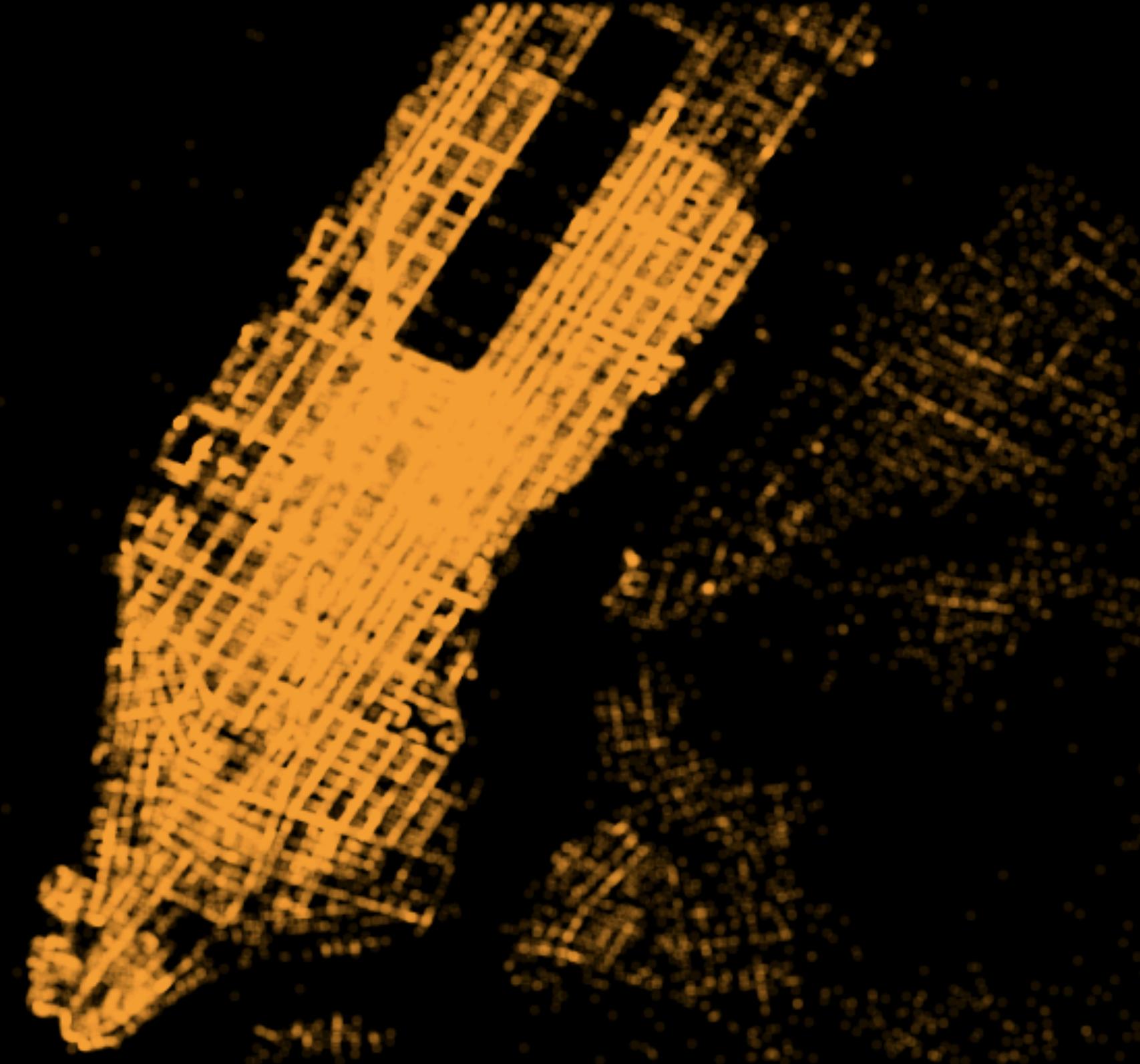
Data-Viz Issue

Plotting is hard e.g.

alpha

Sampling (~1%) may be effective

Require careful tuning parameters e.g. overweighting unusual values



Binning Helps

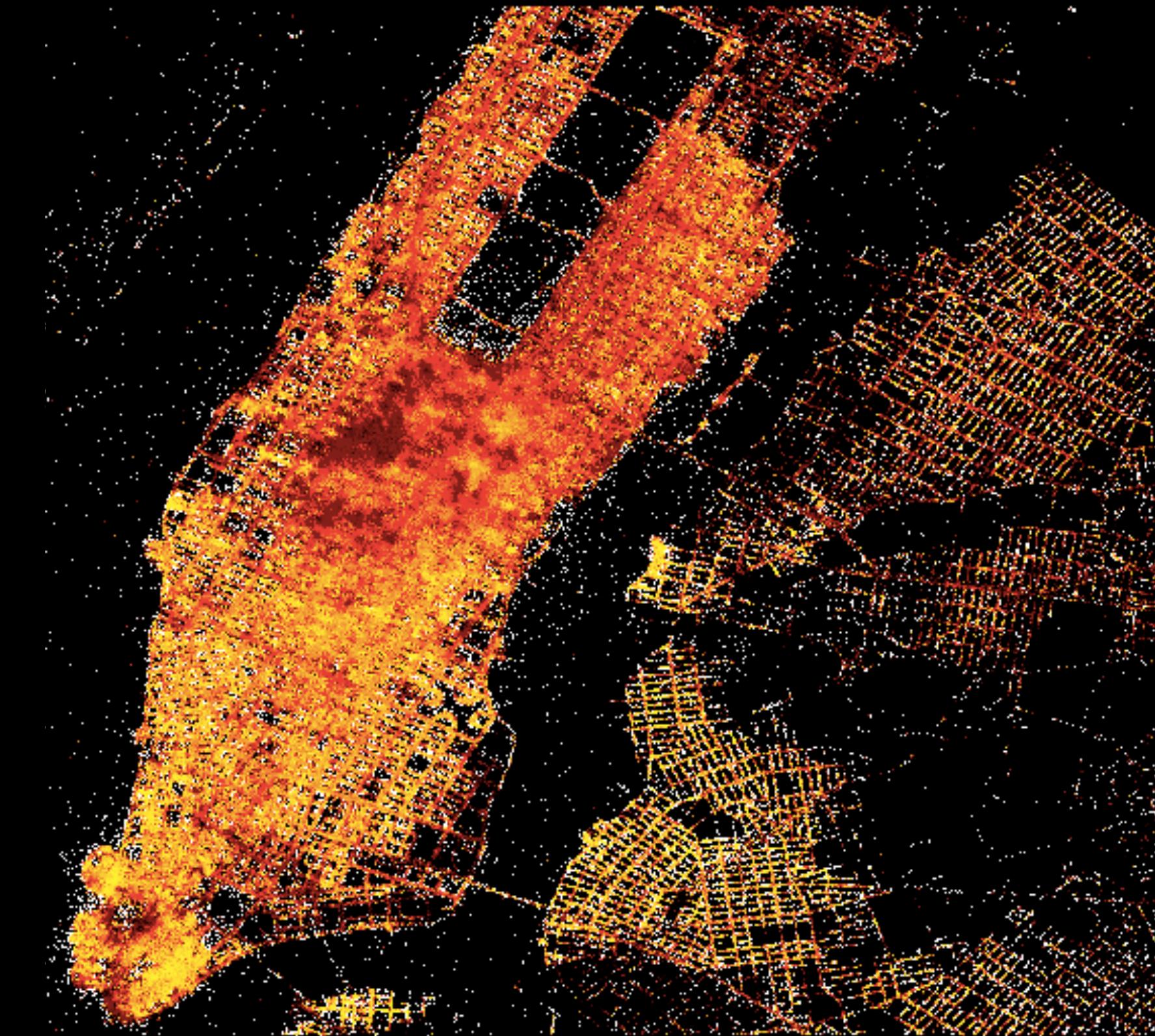
“**Bin - Summarize - Smooth**: A framework for visualising big data” - Hadley Wickam
Package in R: '**BigVis**' (2013)

Recent Interactive implementation in Python
Package in Python: '**Datashader**' (2016)

Vis Data Space

Plot the **probability** of
getting a tip

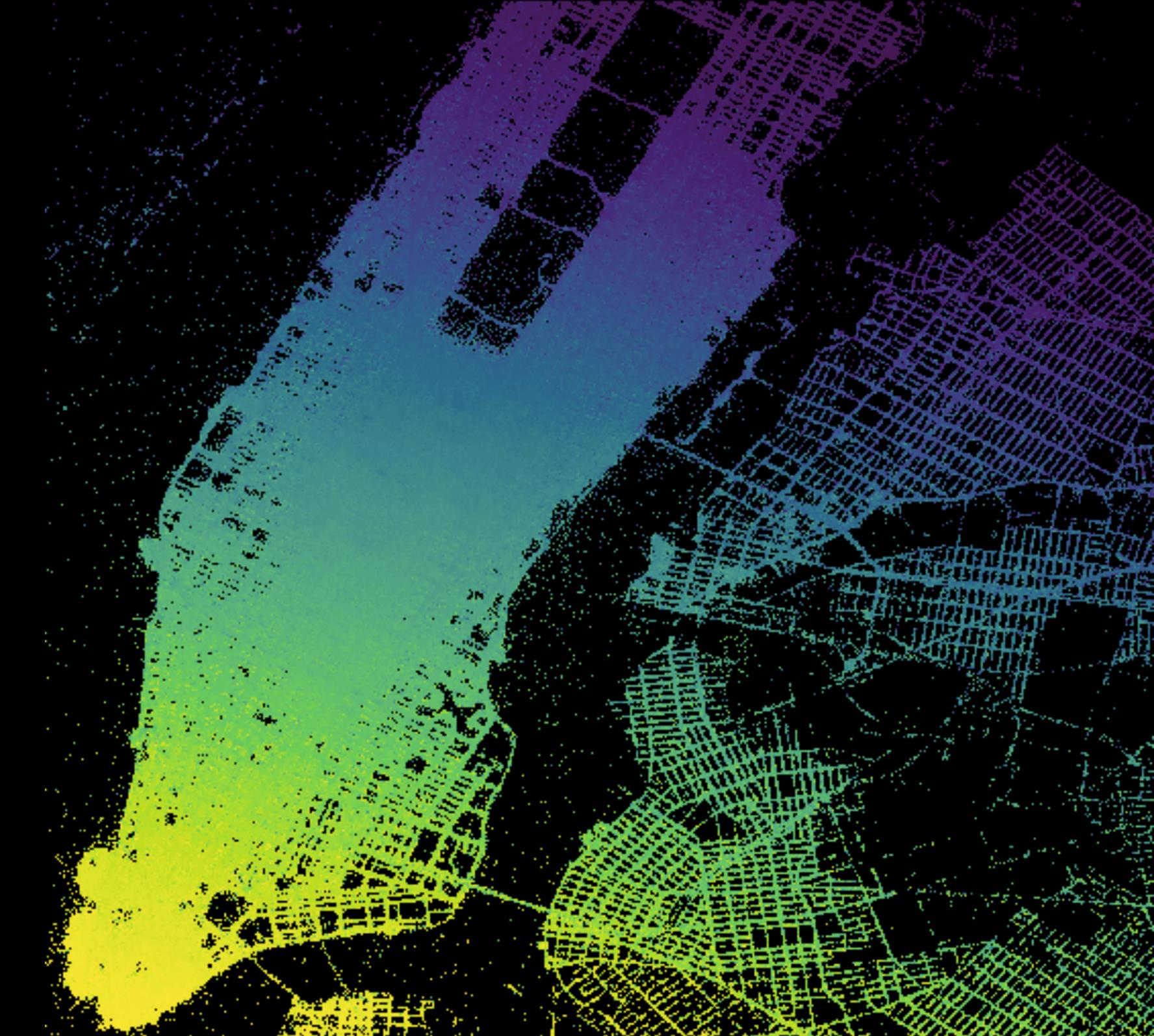
Start to see the
patterns in the
visualisation



Vis Predictions

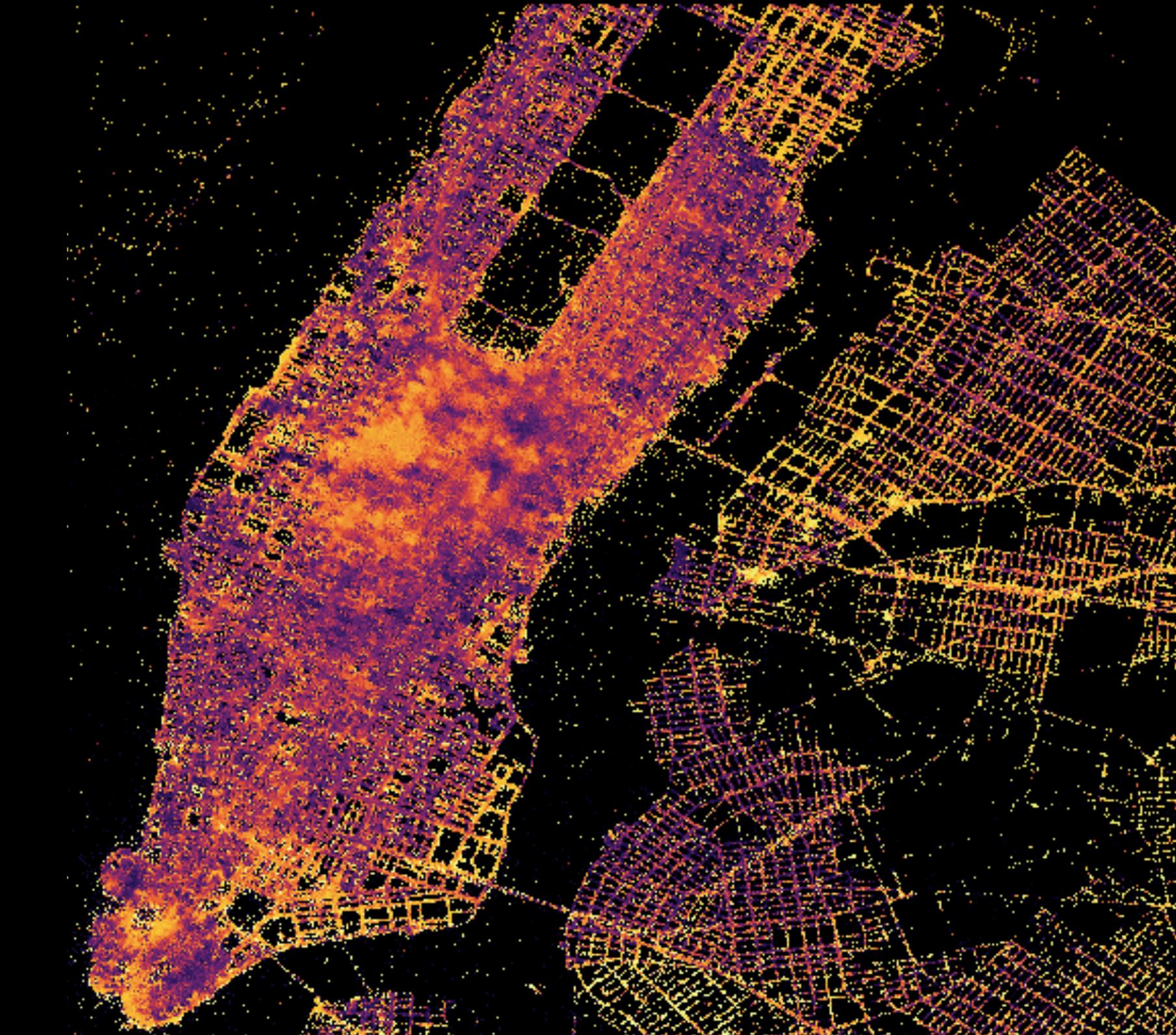
Predict the **probability** of getting tip

Simple Linear Model -
drop coords, passenger
count, time and day of
week



Vis Errors

Visualise the **errors** in
tip probability
distribution



N-Models Challenge

Model Explosion

Entire Model Space

+ Add *Tuning* Models

+ Add *Bootstrap* Models

+ Add *Ensemble* Models

+ Add *Cross-Validation* Models

N-Models Challenge

Keep track of **prediction** &

errors

Keep track of **model output**

parameters

Tidy Model

Augment **predictions & errors** to
dataset

Create **output parameters** data
frame

Visualise like **Tidy Data**

Managing N-Models

"Managing Many Models in R"

by Hadley Wickham

"Broom Package in R"

by David Robinson

p/n/N Model-Vis challenge

p -- High dimensional data

n -- Large and big data

N -- Multiple models

n/p/N Model-Vis approach

p -- use Projections

n -- use Binning or Sampling

N -- use Tidy Model

Model-Viz

Similar challenges to Data-Viz
More an Art, than a Science
Essential in ML Model Pipeline
Both to Explain or to Predict
Scope for easier tooling

Model-Vis

Slides and Code

<http://modelvis.amitkaps.com>

Mini-Site and Explanation

Visualising ML Models

Amit Kapoor
@amitkaps

amitkaps.com