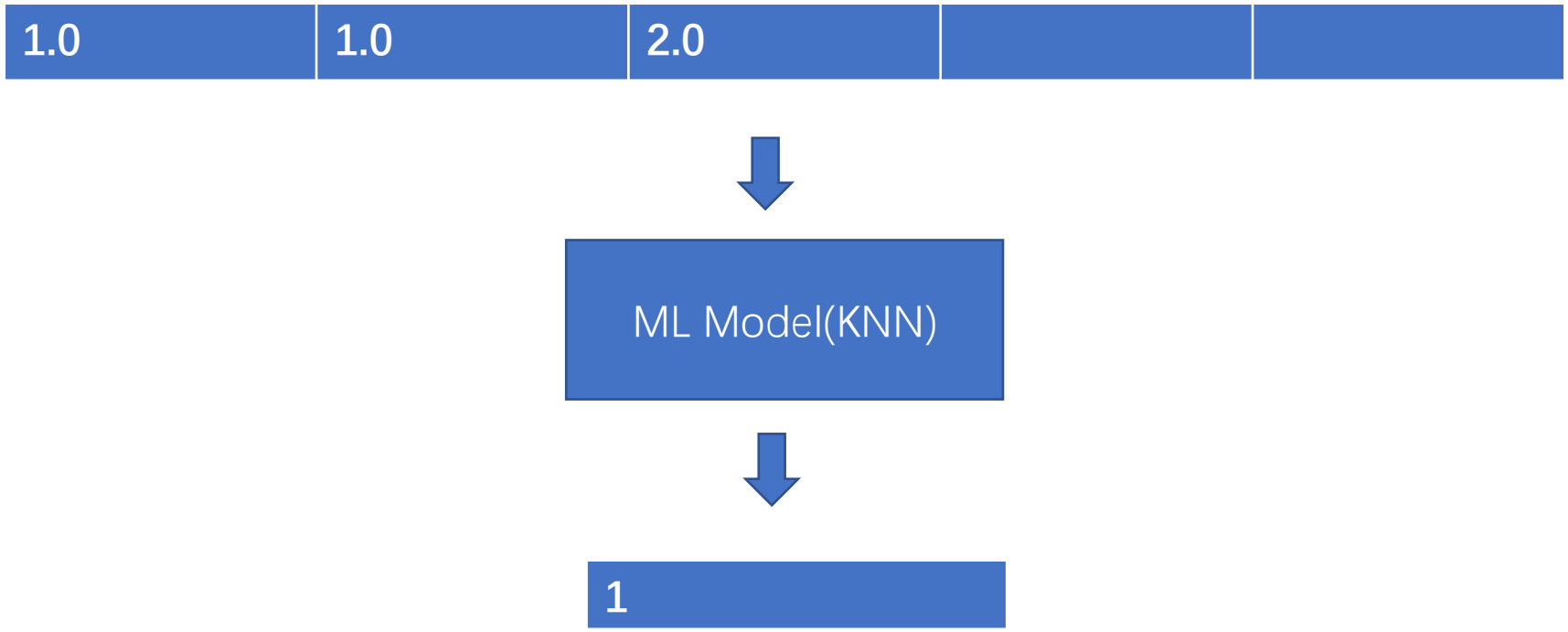
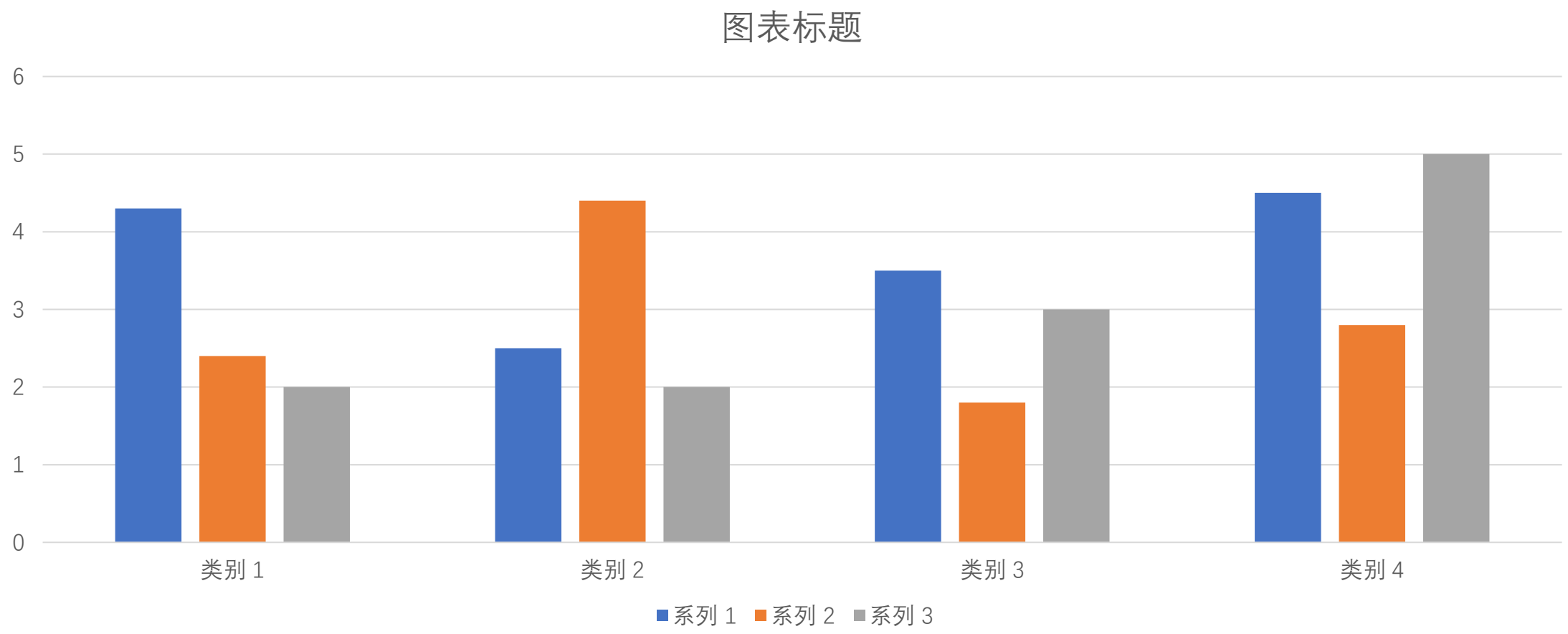
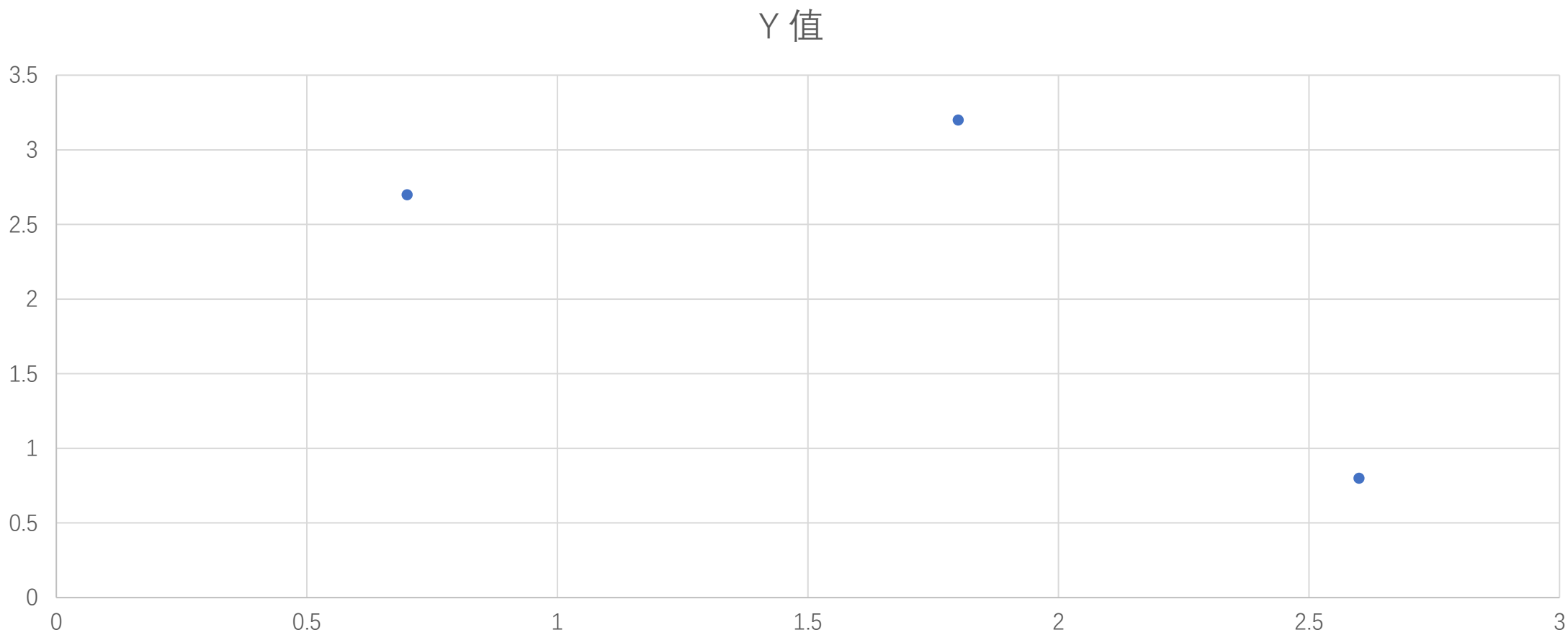
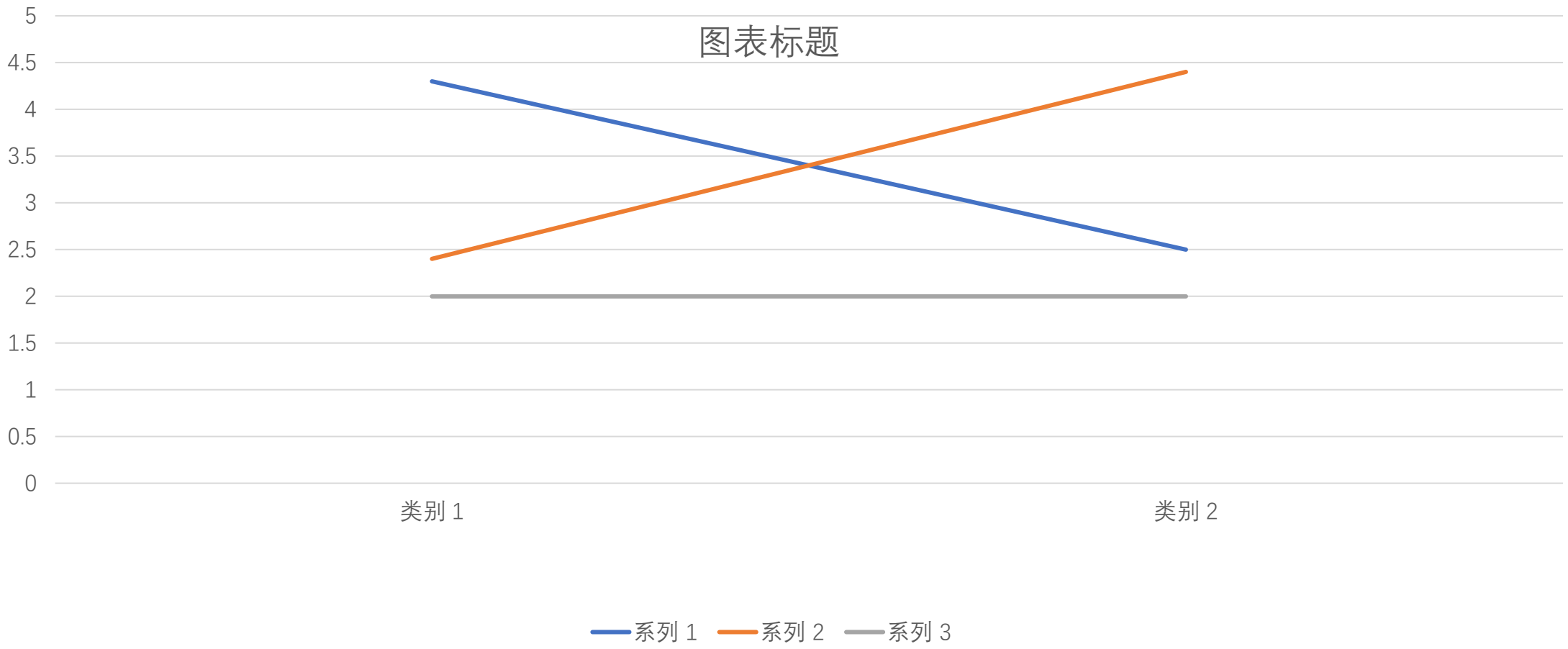
构造更多质量好的特征！ --Goal

数据向量化 – 特征工程

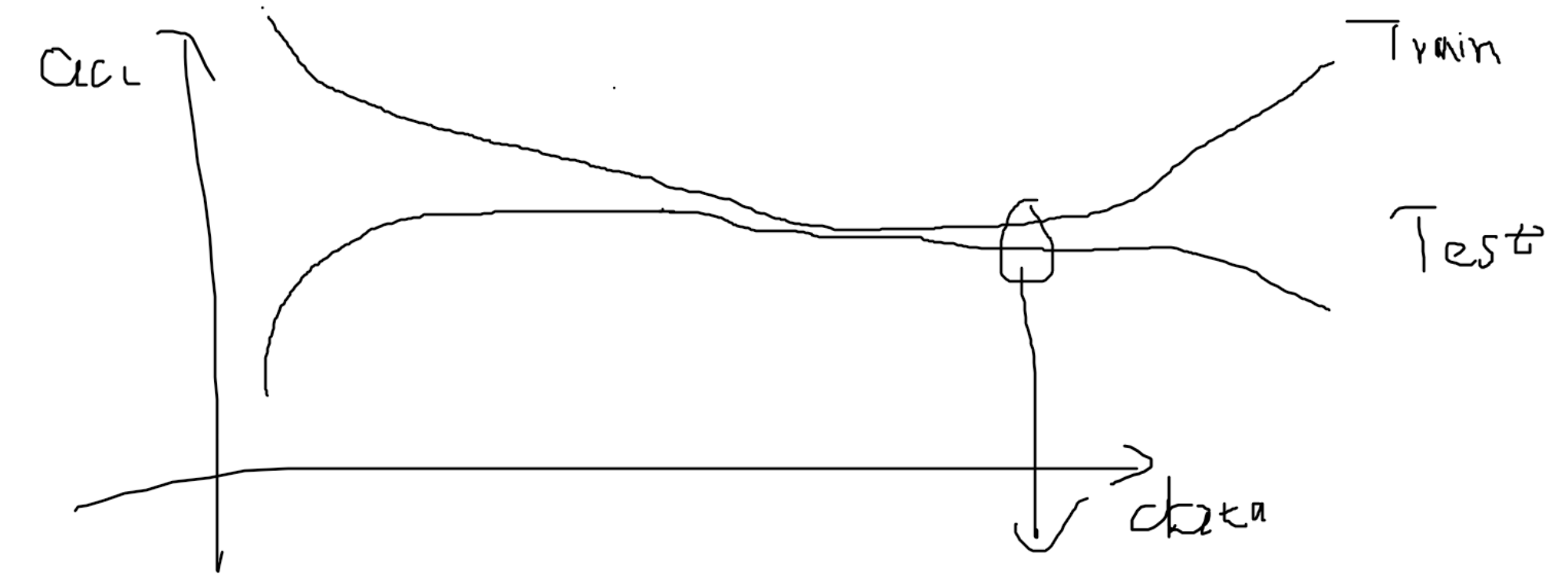



图表标题
6
5
4
3
2
1
0
类别 1
类别 2
类别 3
类别 4
系列 1
系列 2
系列 3

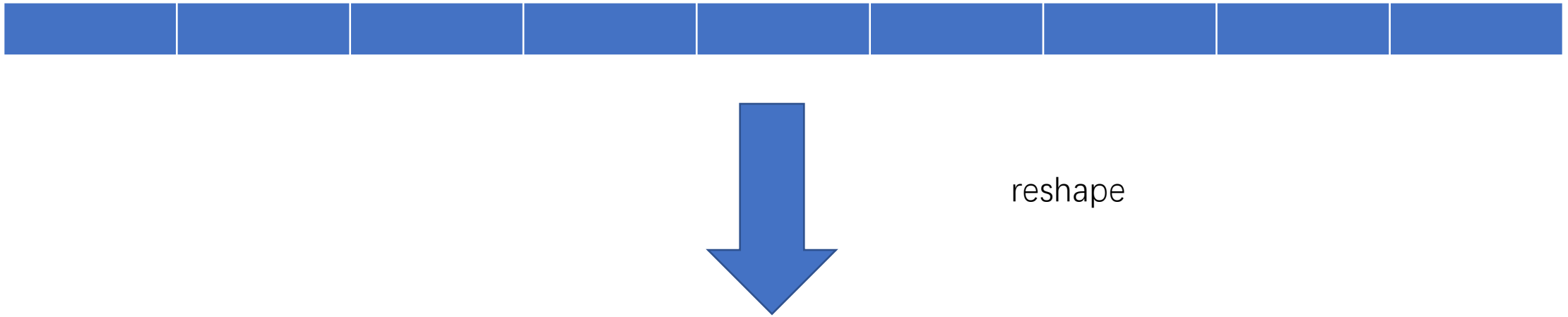

Y 値
3.5
3
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3

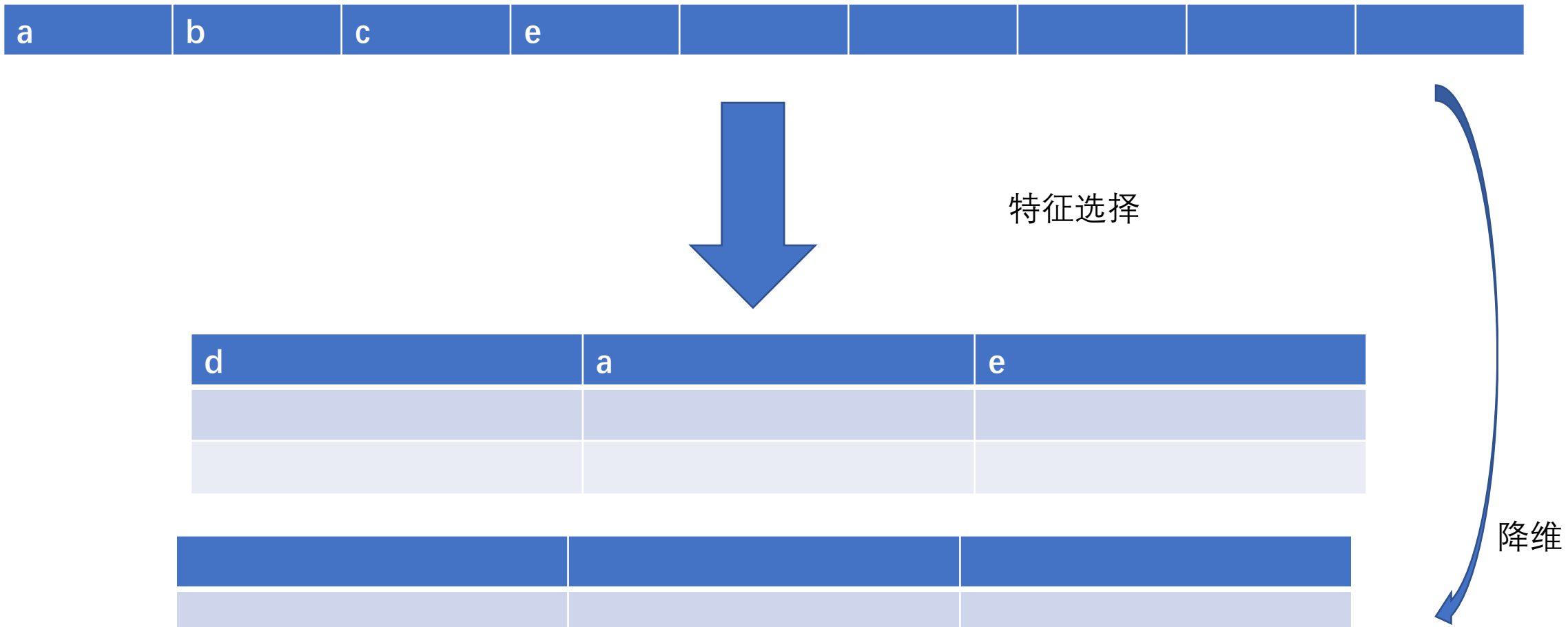

图表标题
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
类别 1
类别 2
系列 1
系列 2
系列 3

9:37

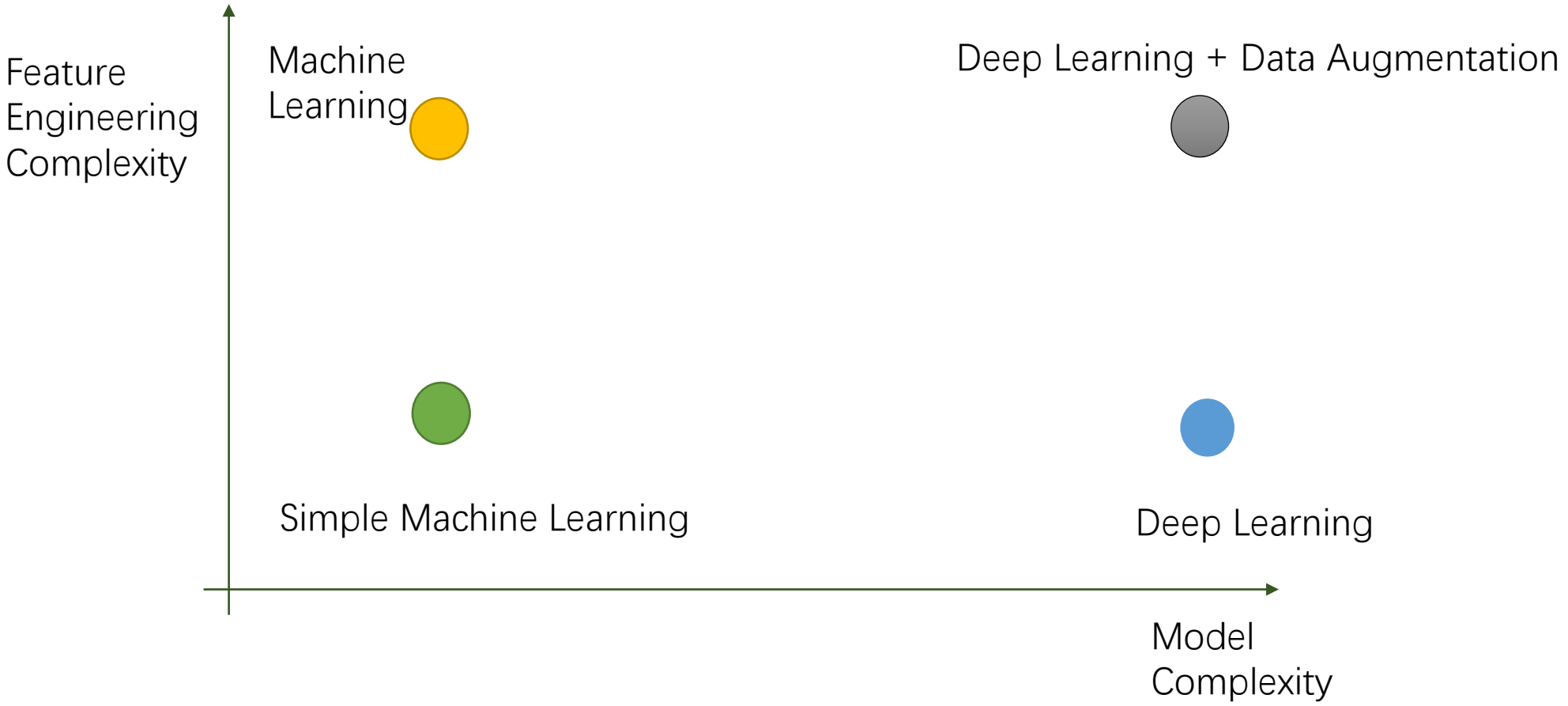


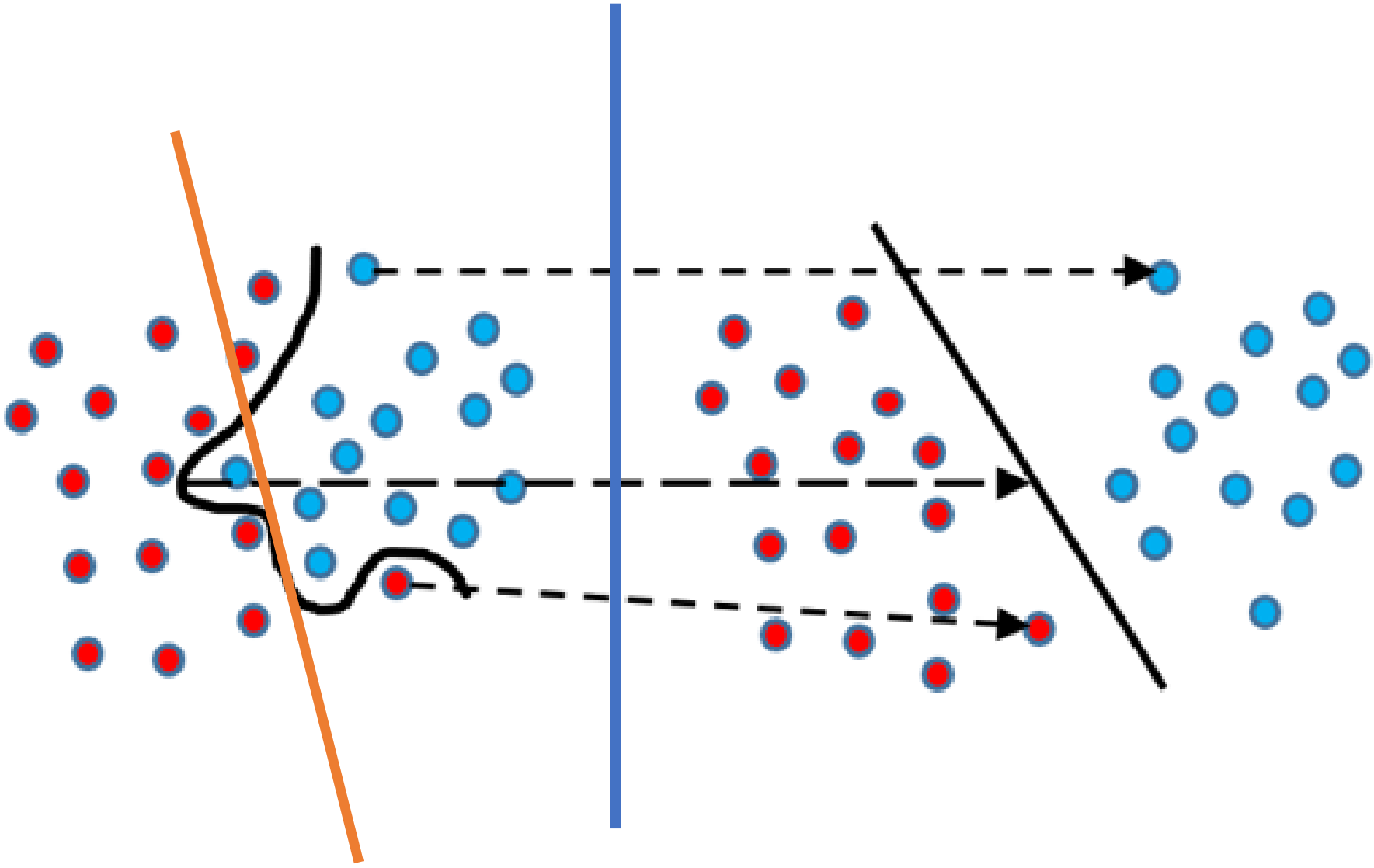
- Vec a + Vec b = Vec c
- For i in range(10):
  - c[i]=a[i] + b[i]
  - ->c = a + b


reshape


a
b
c
e
特征选择
d
a
e
降维

- Y = wX + b
- 模型被保存的文件 = 权重W,b + meta (函数形式)


Feature Engineering Complexity
Machine Learning
Deep Learning + Data Augmentation
Simple Machine Learning
Deep Learning
Model Complexity

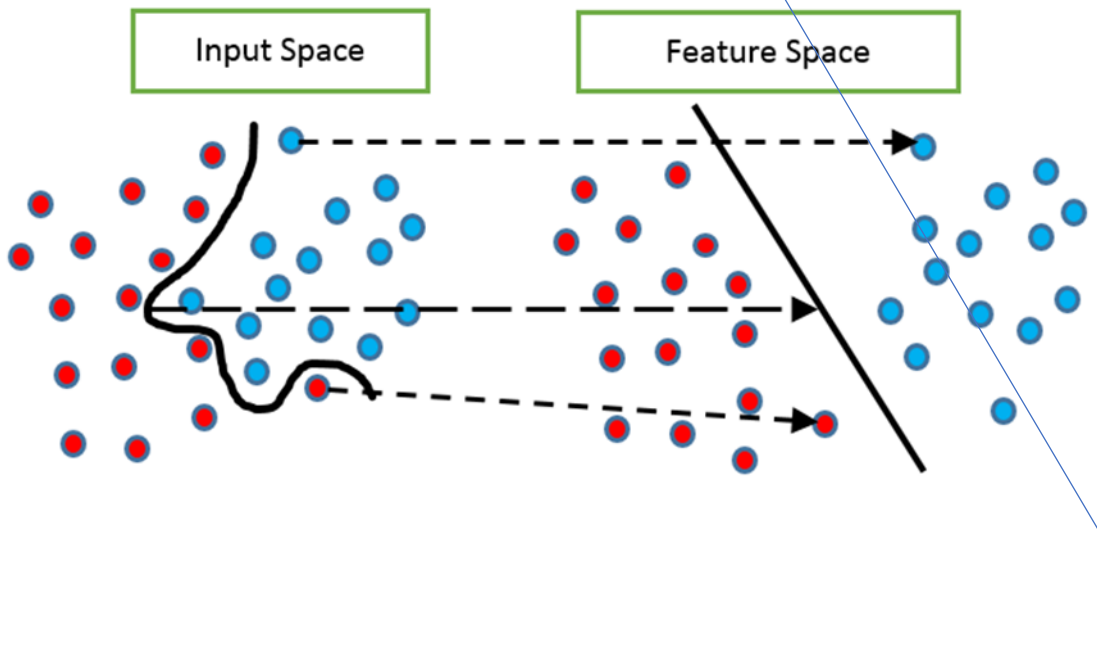


Data Augmentation

空值处理

- 缺失较多，可以丢弃
- 填充
  - 填充固定数值
  - 离散数值 – 众数或新数值
  - 连续数值 – 均值
- ML预测

- Label Encoder
  - Company_type：　big, middle, small.
  - 0,1,2
- One encoder
  - Company_type -> c_t_a, c_t,_b, c_t_c
  - [1,0,0], [0,1,0], [0,0,1]

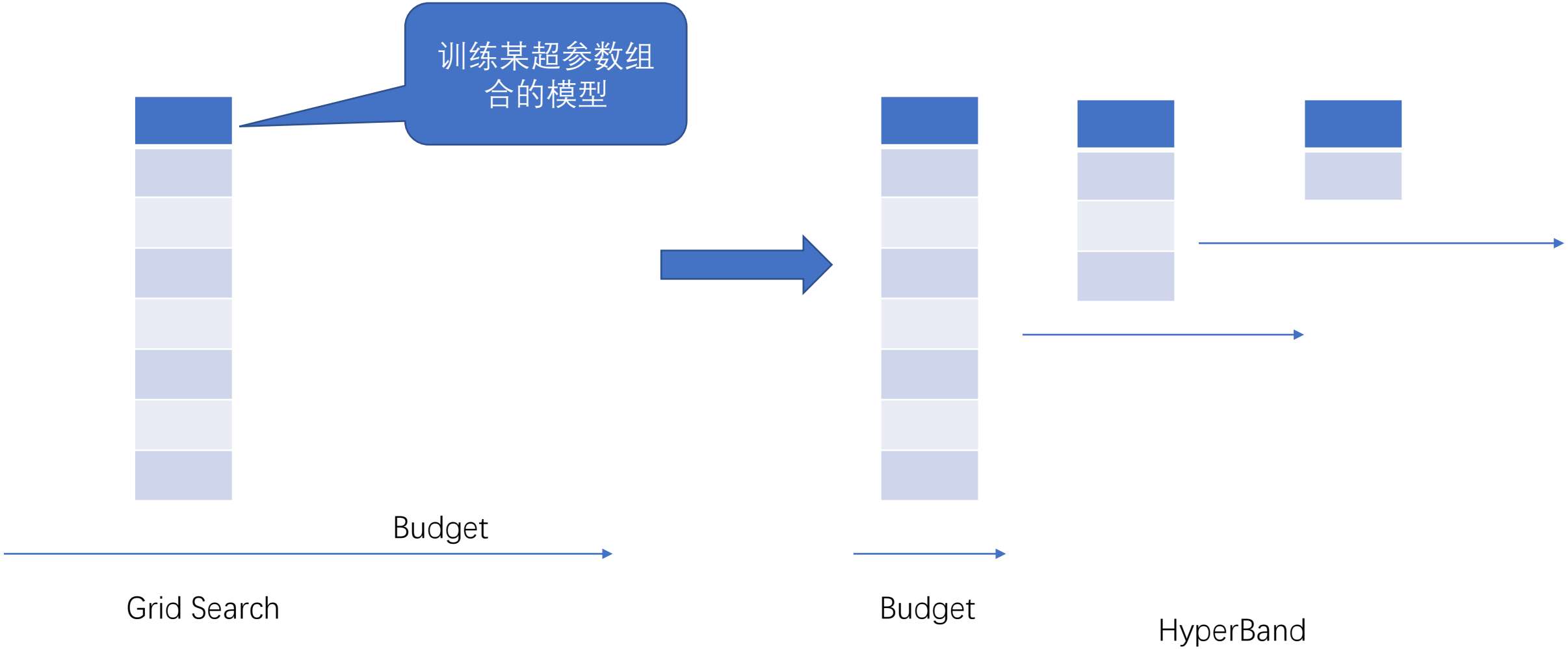

Input Space
Feature Space

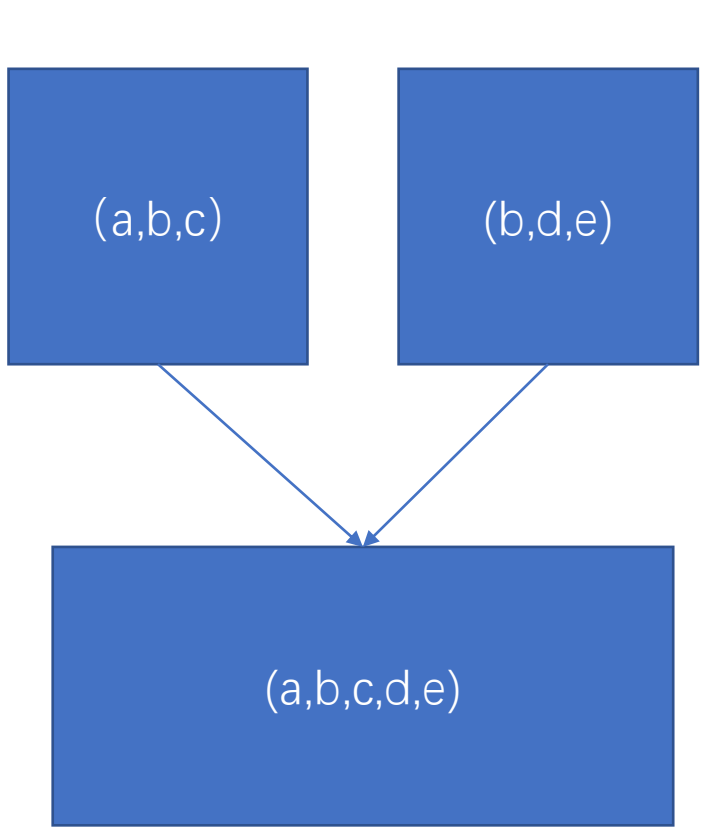

0
100
500

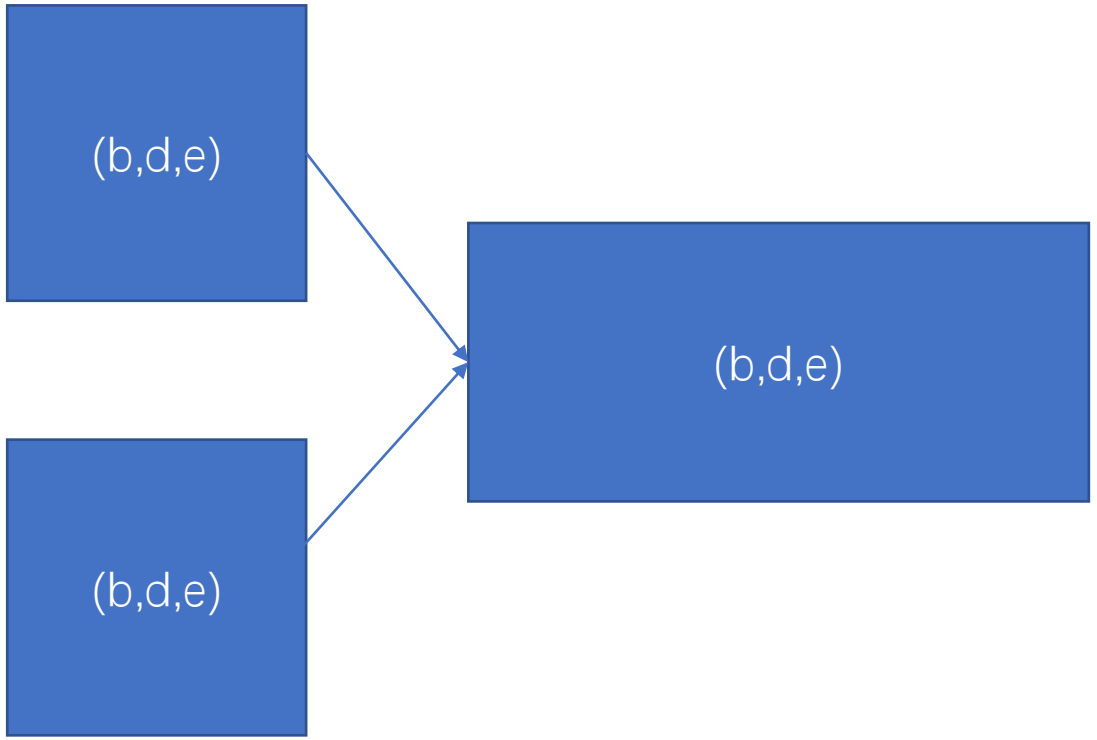
Embeded feature selection

- Y = WX + B
  - W反应特征重要性
- Decision tree

AutoML – HPO




(a,b,c)
(b,d,e)
(a,b,c,d,e)


(b,d,e)
(b,d,e)
(b,d,e)