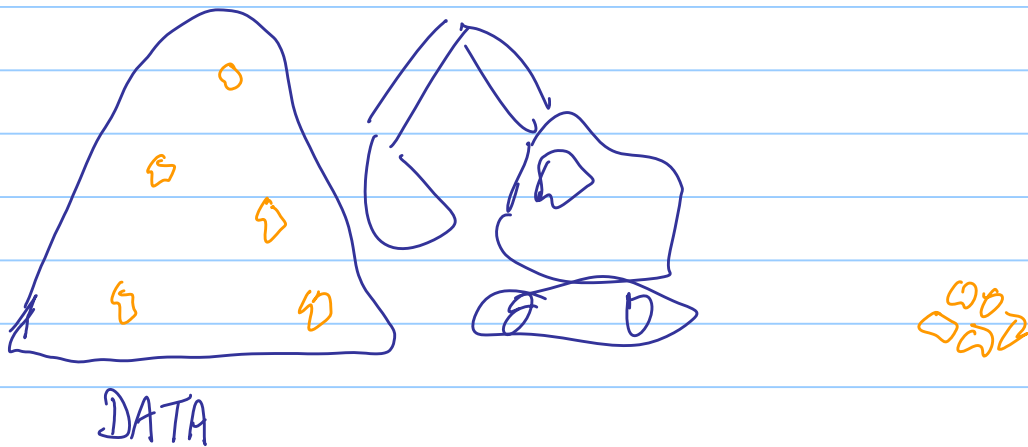


CISC 873 Data Mining

Note Title

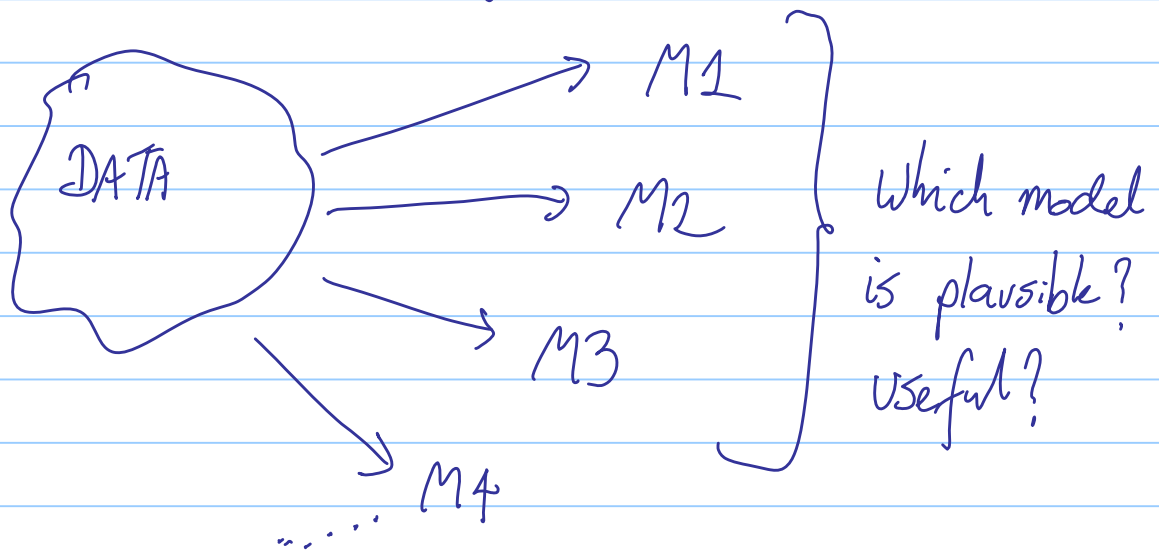
9/16/2011



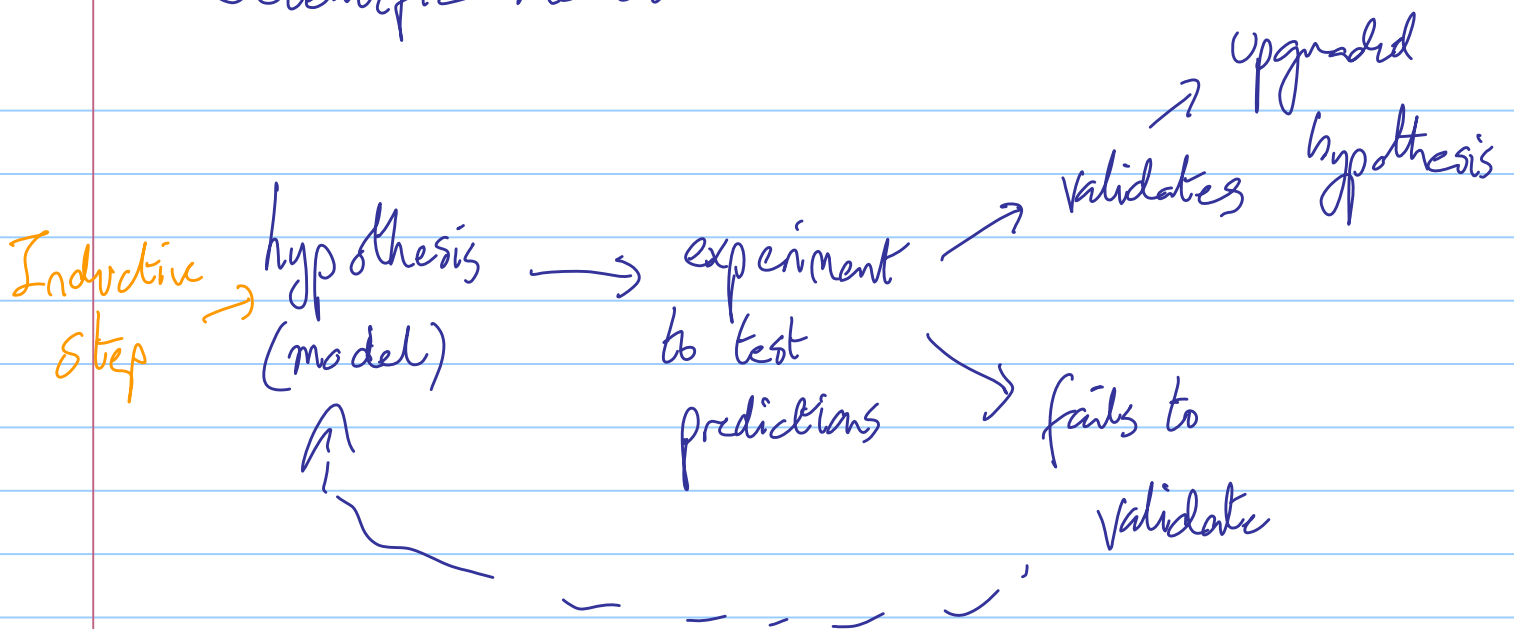
Data Mining \equiv Knowledge Discovery

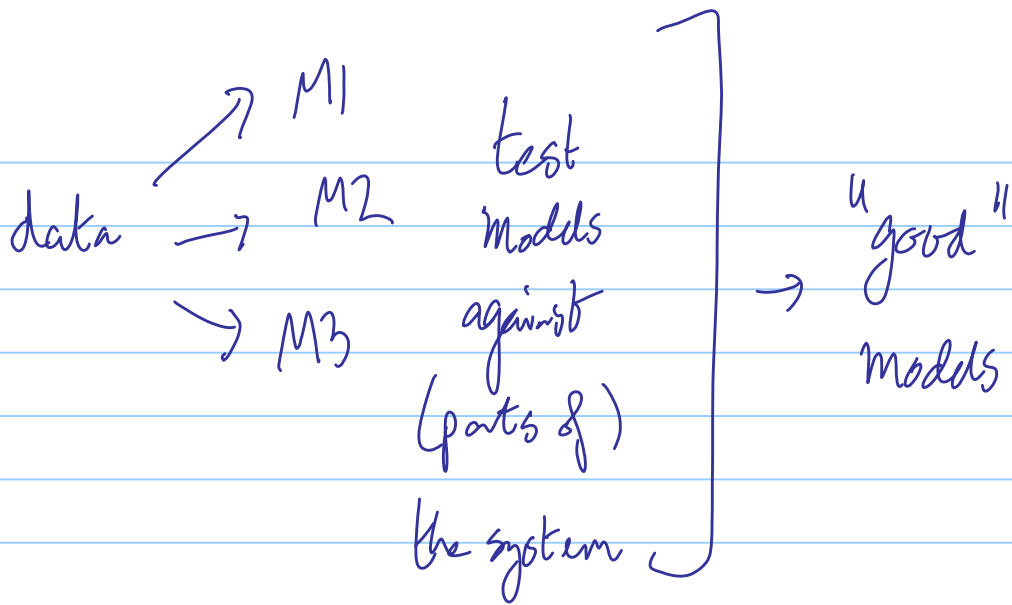
Inductive modelling of data

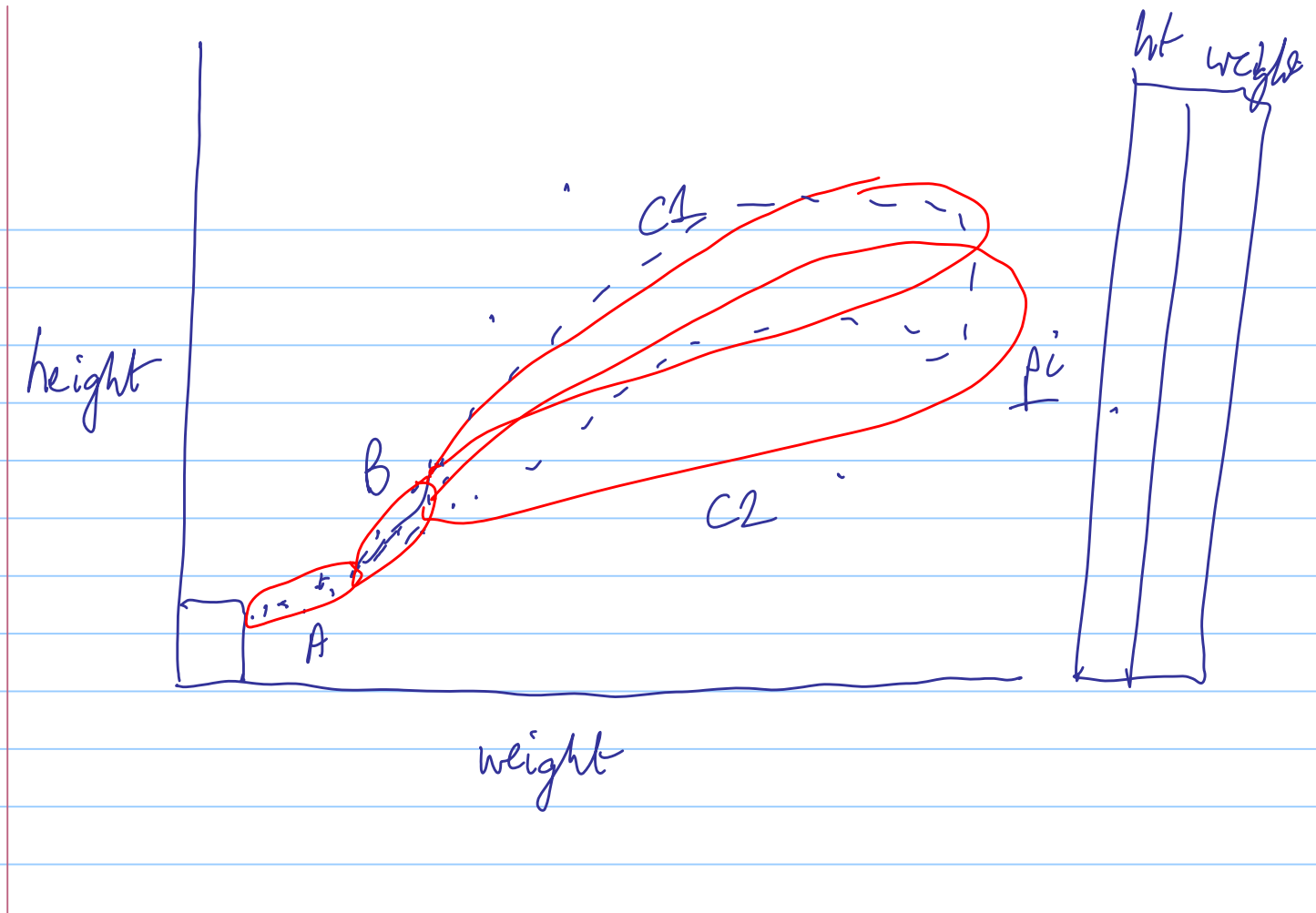
↳ Data tells you its structures

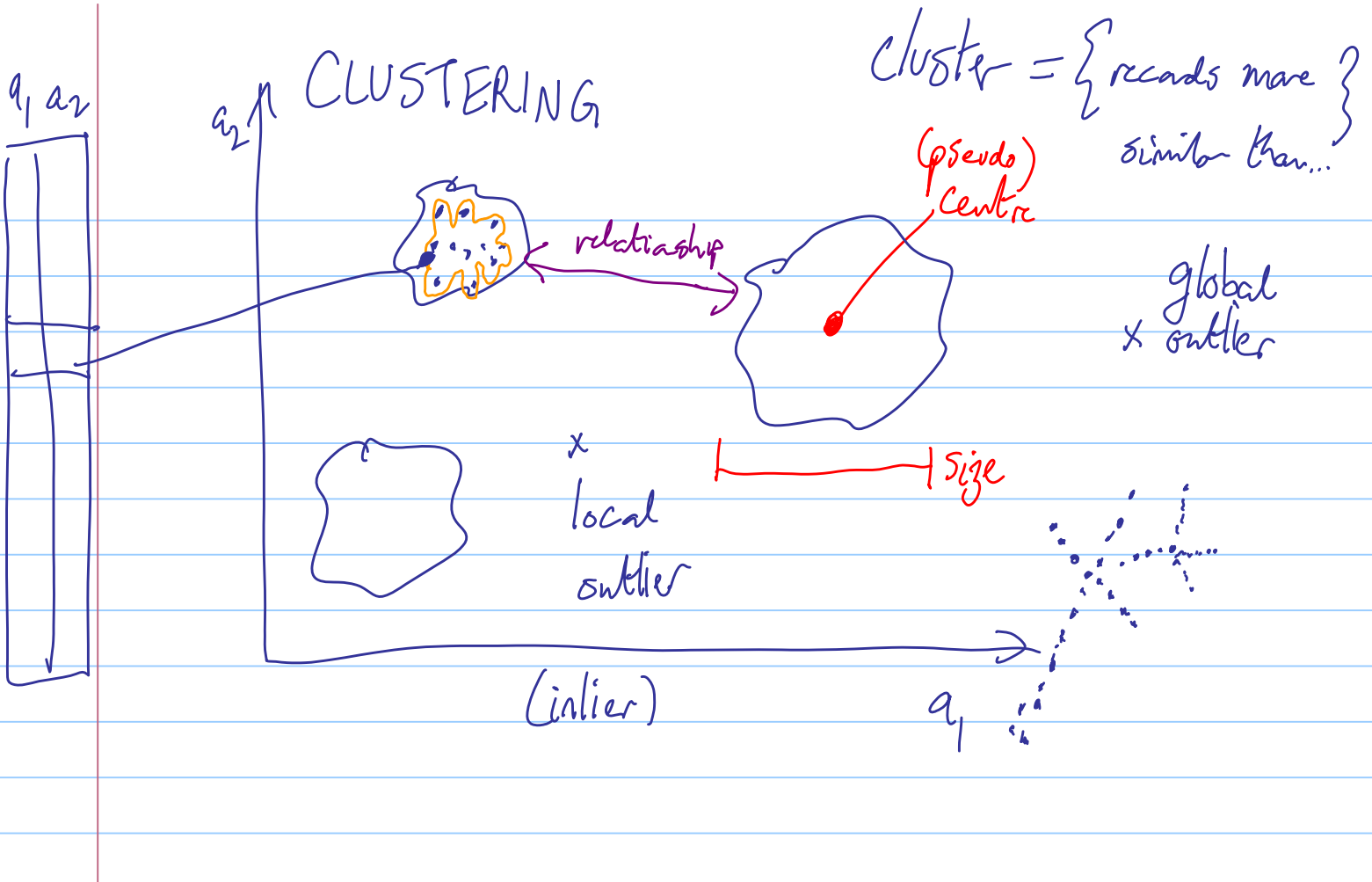


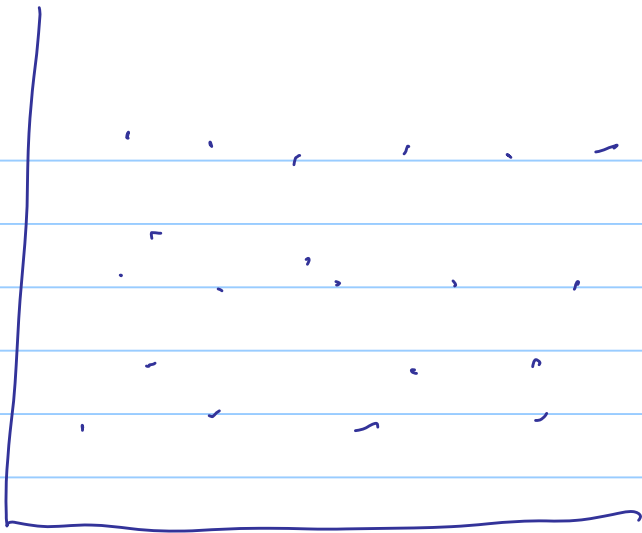
Scientific method

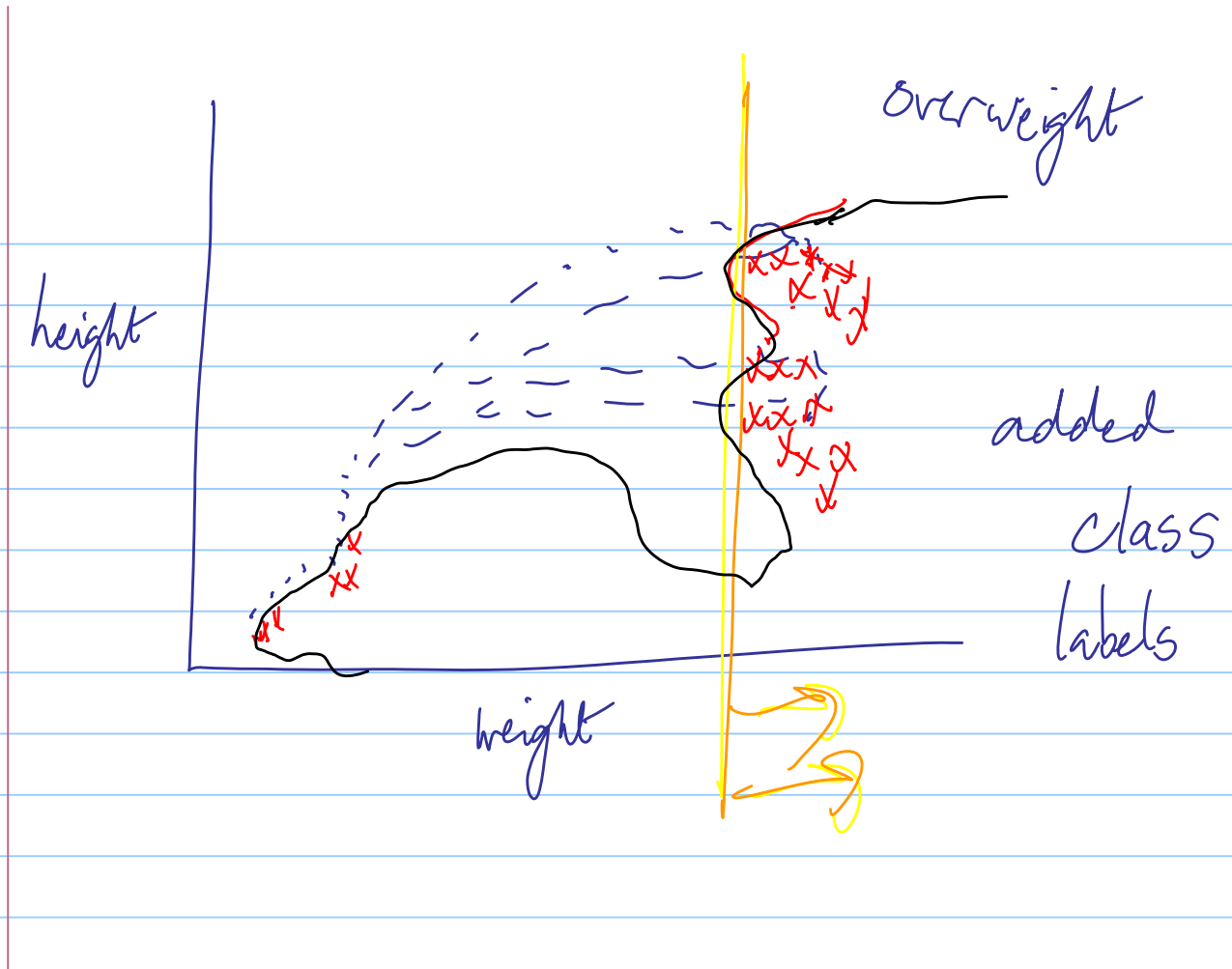






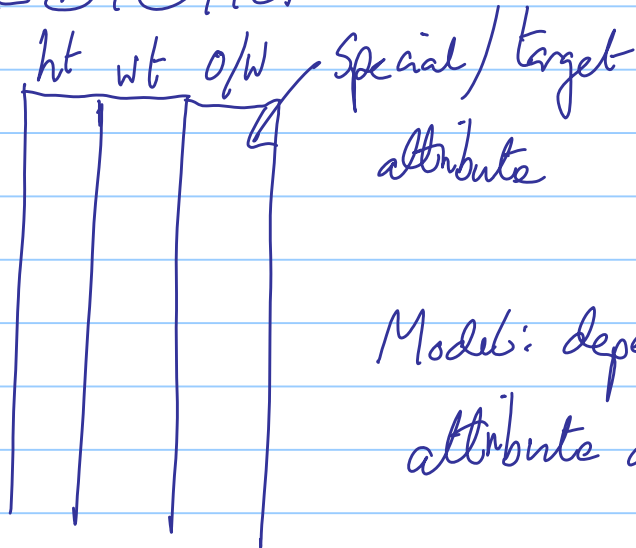






model = boundary between the classes

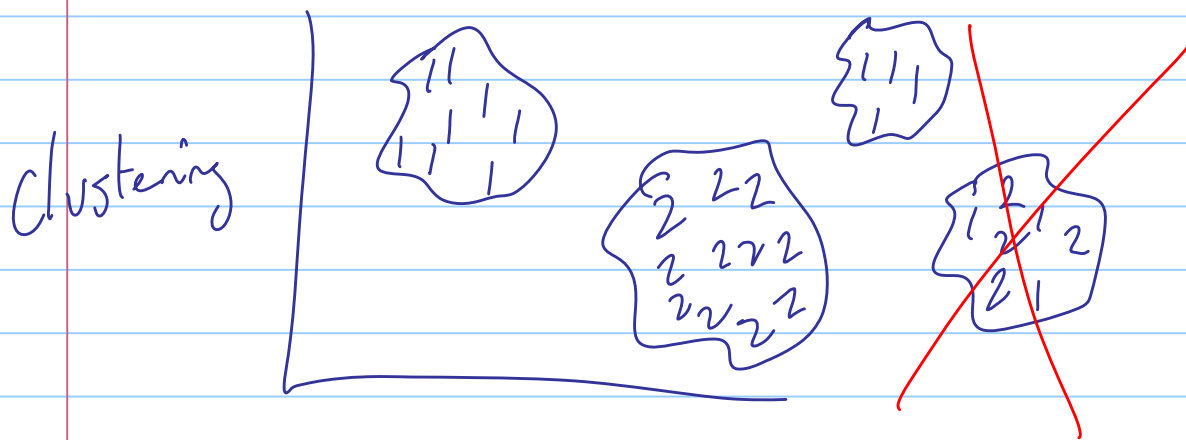
PREDICTION



Model: dependence of target attribute on other attributes

Model is used to predict class/target value
for new, previously unseen records.

Labelled data (2 classes)



Classification vs regression

target attri
few fixed values

target attribute
continuous variable

Issues:

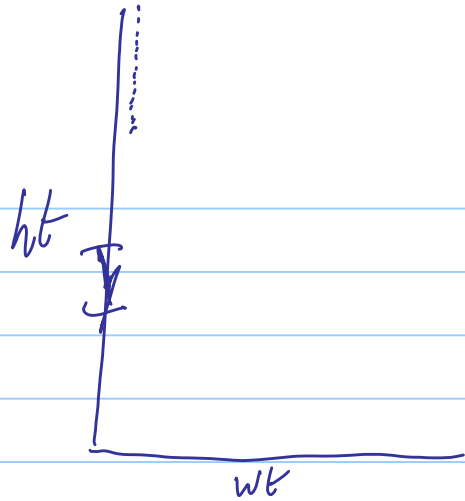
① representation of the data

height angstroms
weight tons

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

apples oranges

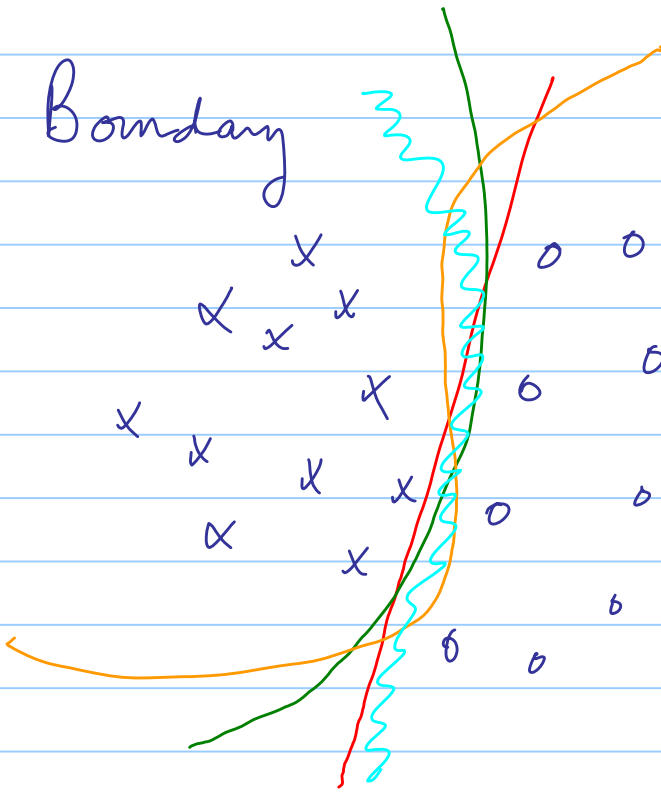
Significance & magnitude



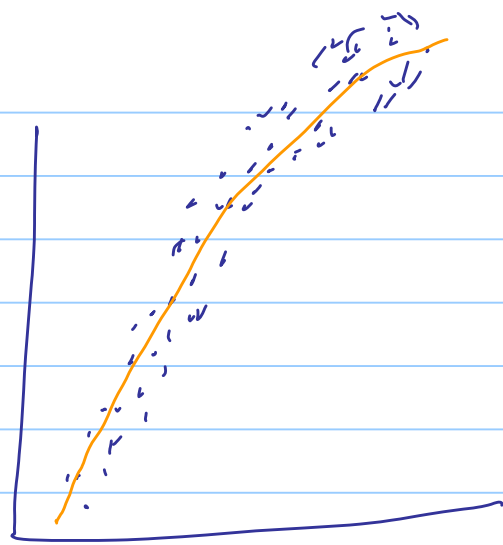
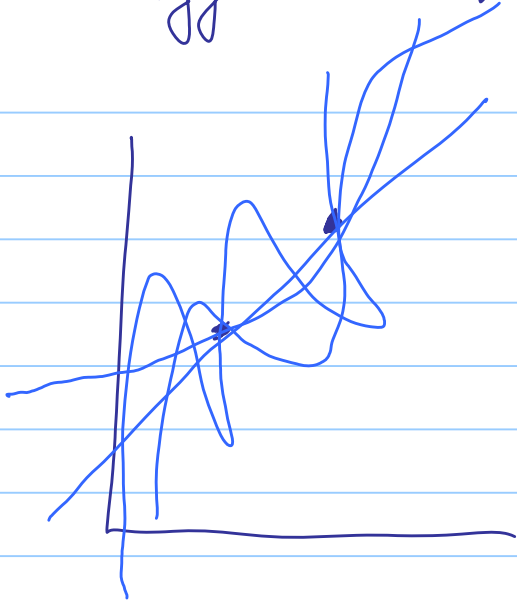
Similarity

Clustering ← choice of similarity

Boundary



Bigger data \rightarrow more possible models \rightarrow harder choices?



Supermarket

profit

p_1 p_2 p_3 ...

10,000

cost 1

1 0 3 0 ... 17 0 ... 0

$\sum f_{ij} \times \text{profit}_j$

(06,000)
month

