Active learning

Problem with supervised learning

① Require "labelled" data

↳ Hard to label

eg. trash segmentation
(humans do poor)

↳ Expensive

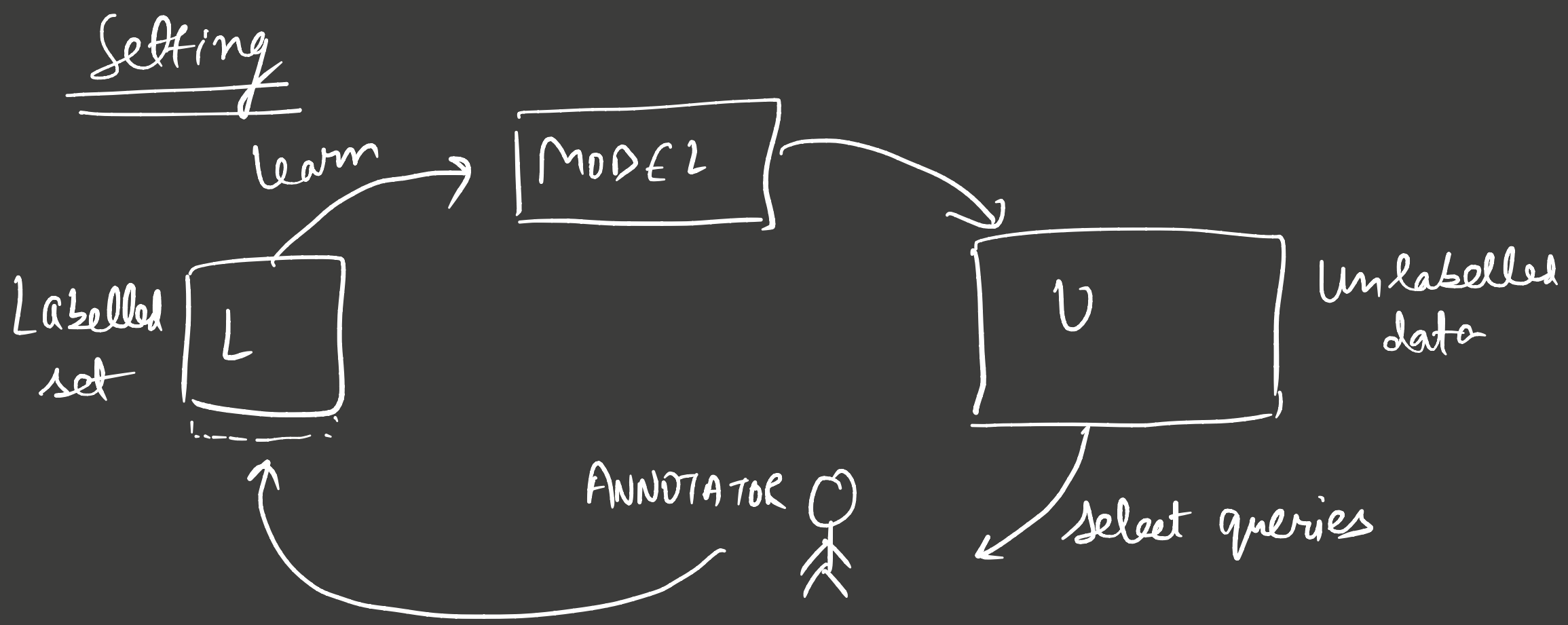
↳ Time consuming

① Unlabeled data is easy to get / cheap

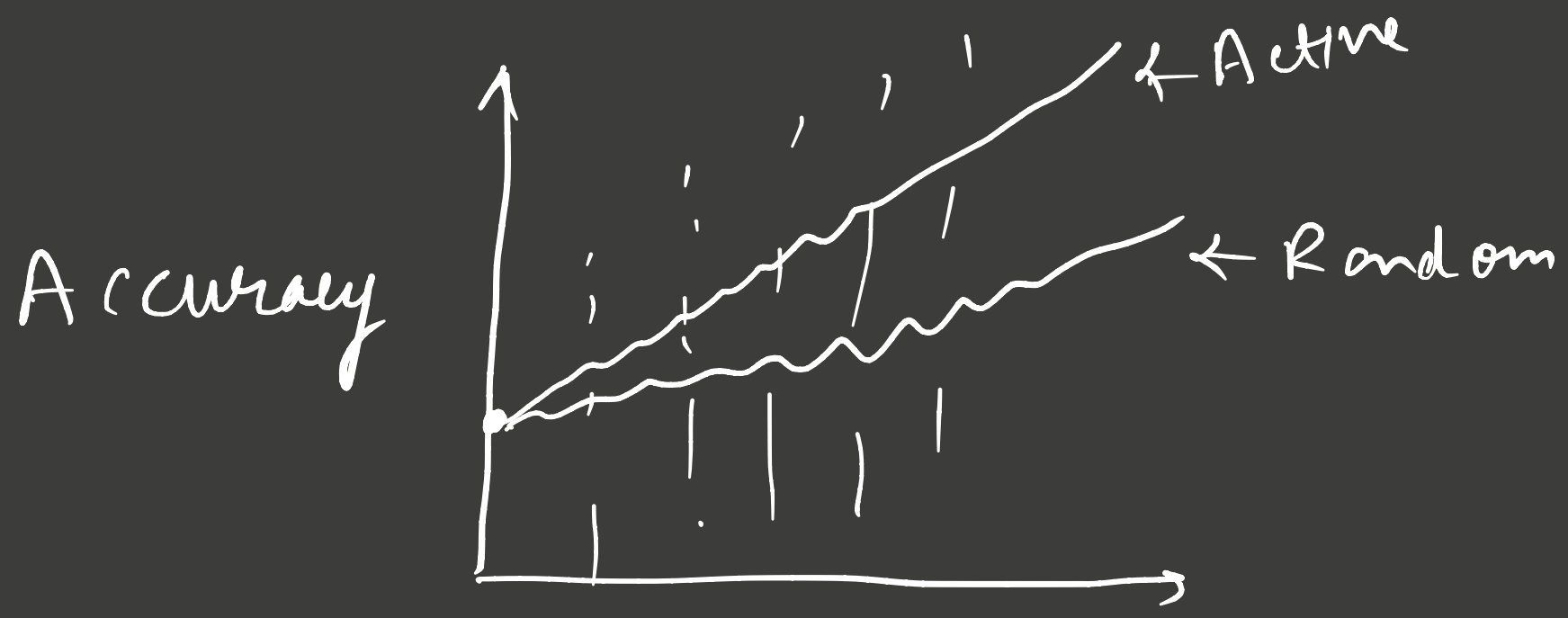
② Label "as few" as points as possible.

- Budget
- A century desired

Setting

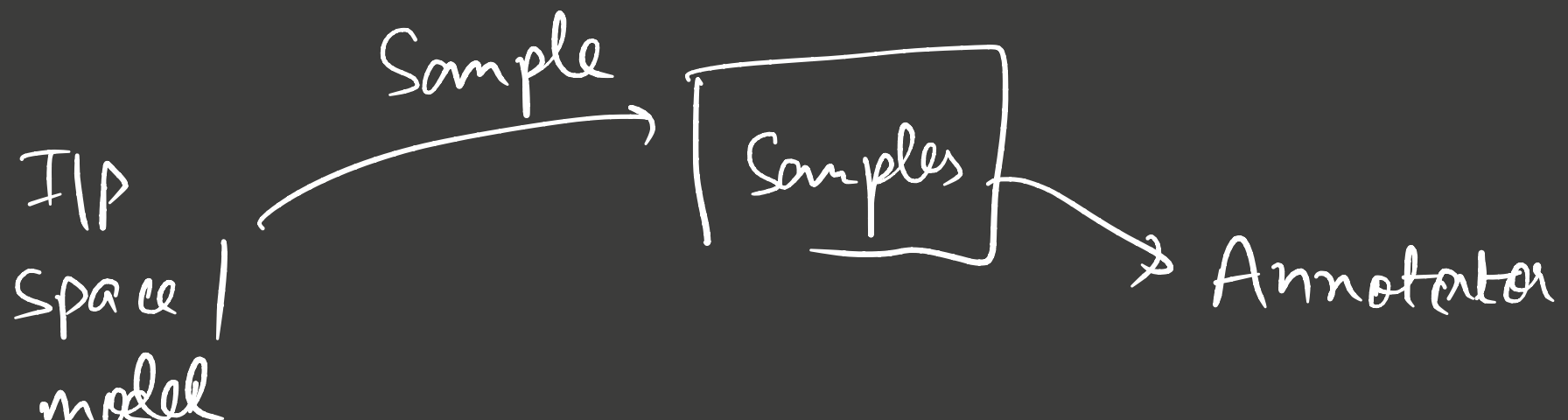


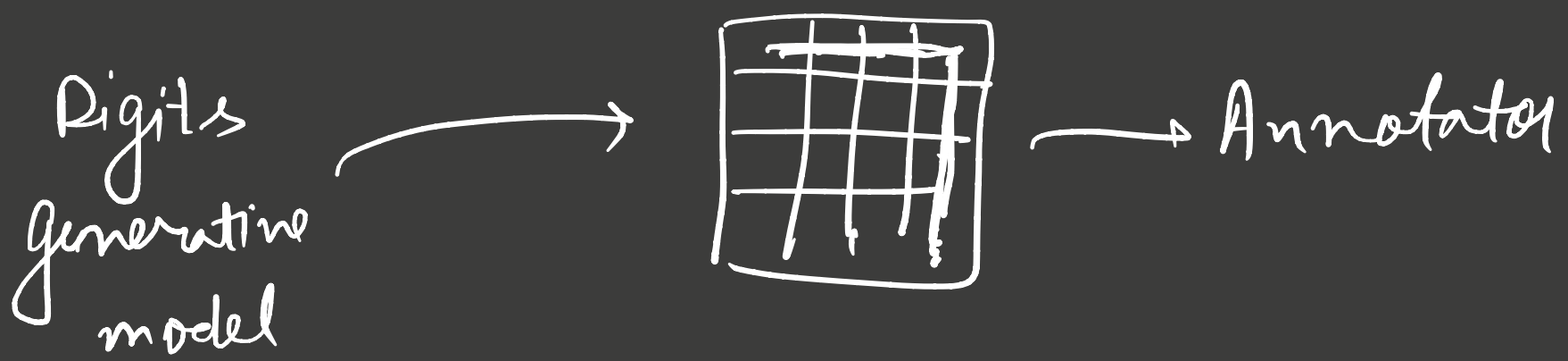
Evaluation of Active Learning



Scenarios of Active learning

(1) membership Query synthesis

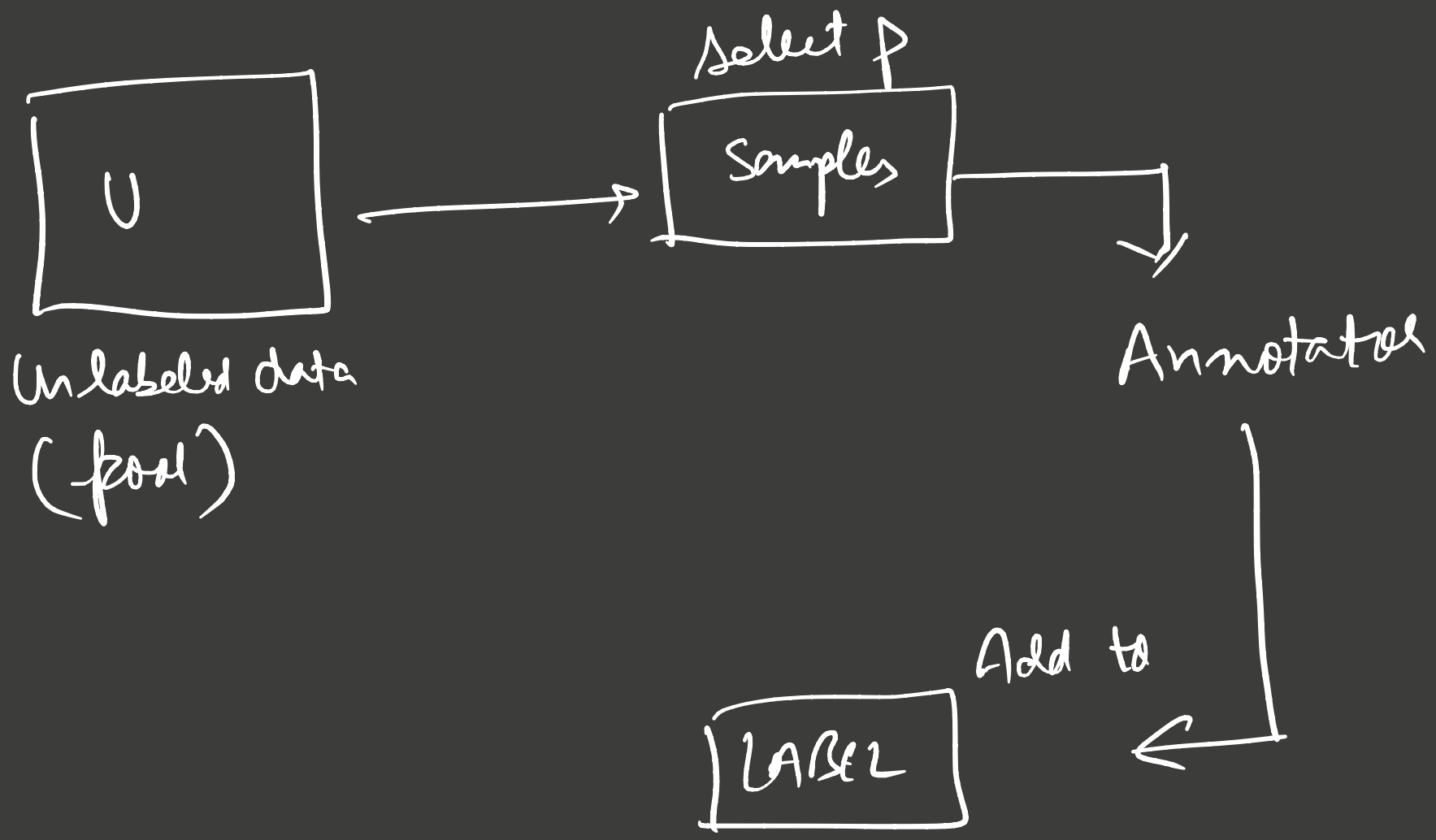


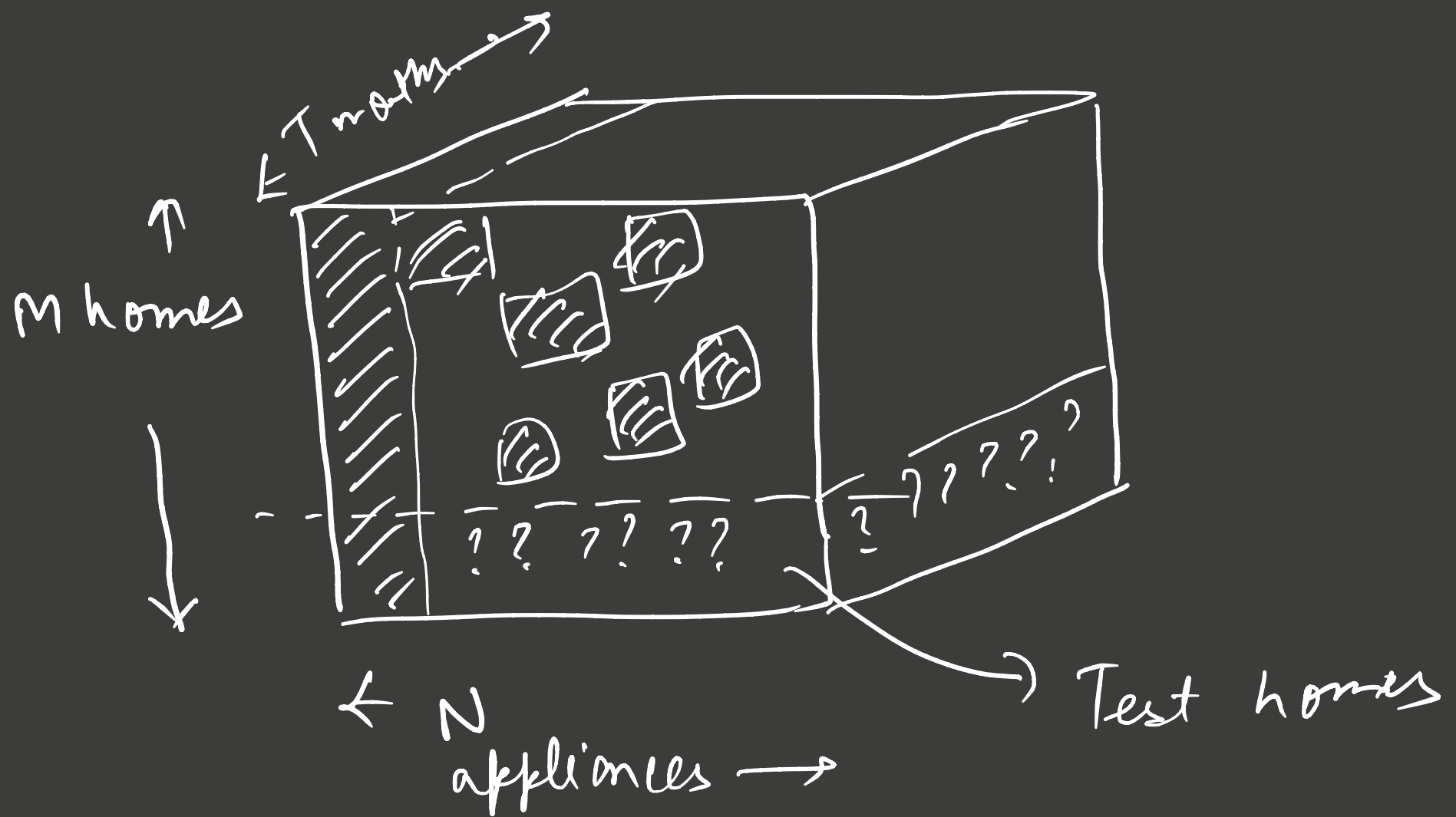


Can lead to incoherent examples.

② Stream based

3 Pool based.


Select
U
Samples
Unlabeled data
(pool)
Annotator
Add to
LABEL



0) Given budget of X sensors

Choose < home, appliance > ⟹

Maximize accuracy.

Query Strategies (for Pool based Sampling)

(1) Uncertainty sampling

→ query points you're most uncertain about

→ least confident estimate

$$x^*_{LC} = \arg\max_x \left(1 - P_\theta(\hat{y}|x)\right)$$

$\hat{y}$ = Class with max. probability

→ Margin sampling

$$x^*_M = \underset{x}{\operatorname{argmin}} \left(P_\theta(\hat{y}_1 | x) - P_\theta(\hat{y}_2 | x) \right)$$

$\hat{y}_1, \hat{y}_2$ are the two most probable classes

0) How will you do uncertainty sampling for KNN (KK classificat~)

Votes (+) ~ Votes (−)

Query by committee (QBC)

1) Create a committee of models $C = \{\theta^1, \theta^2 \ldots \theta^c\}$ trained on labeled set (L) but represent competing hypothesis.

2) Each of the model votes

3) Most informative instance = most disagreement in votes (entropy high)

eg. OBC with Random Forest

θ^1 = Model 1 = R.F. with seed 1

θ^2 = " 2 = " " " 2

⋮

eg. OBC with KNN

θ^1 = KNN with K=1

θ^2 = ----

QBC for regression

⇒ Compute variance amongst committee

⇒ Choose instances with highest variance.