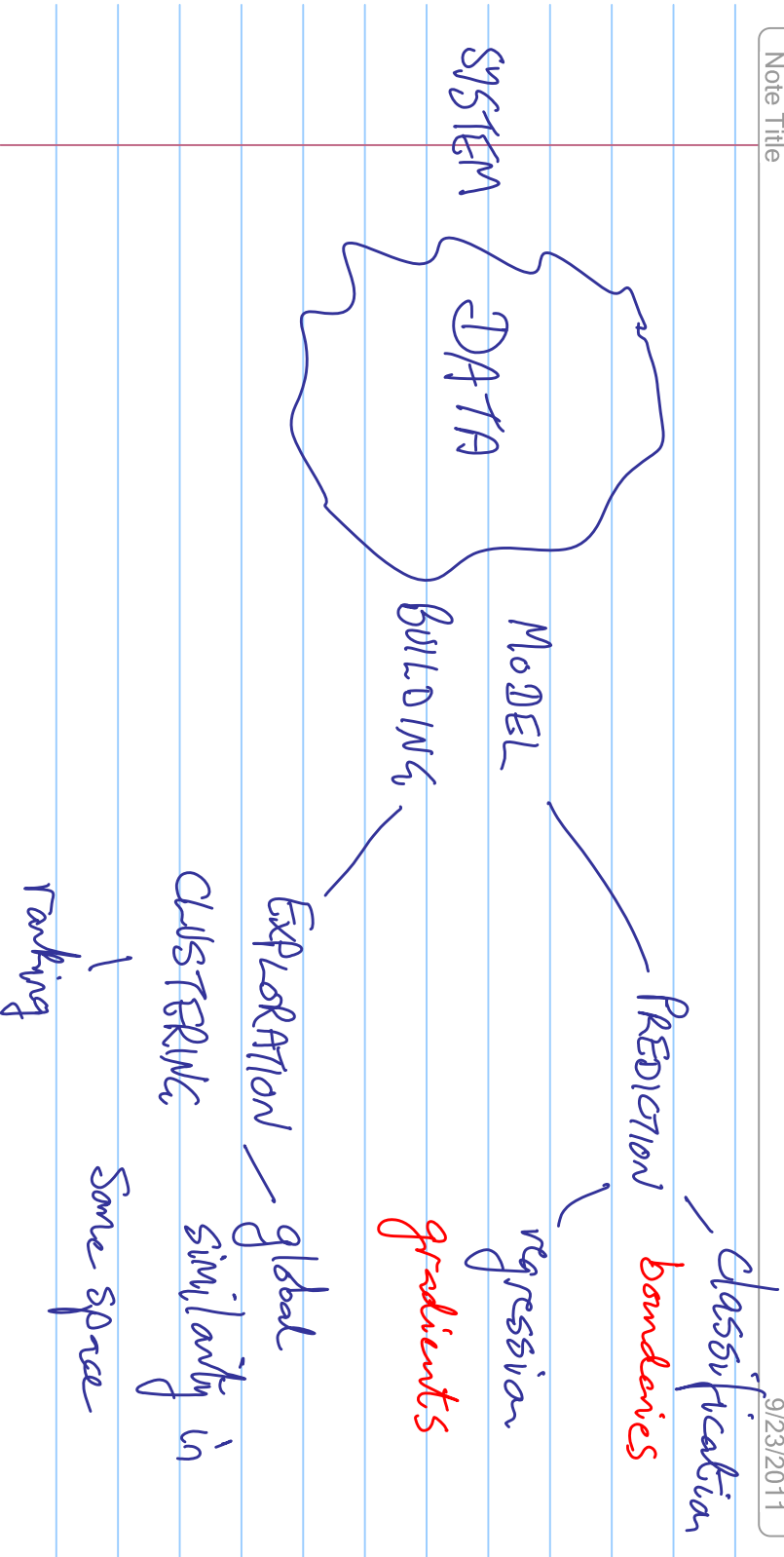
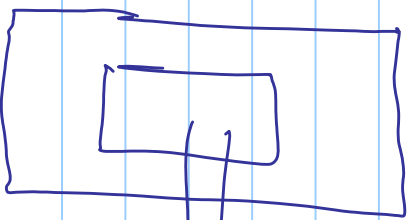


873 Class 2



DATA - matrix / table



Sample from a larger universe
(convenience)

- wrong data - CHECK

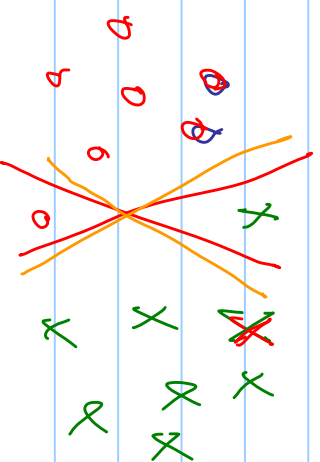
- missing data - why? missing at random?

missing not at random

- Sampling

BIAS

VARIANCE



"Good" model — a good match of algorithm & data

accident?

⇒ build more than 1 model

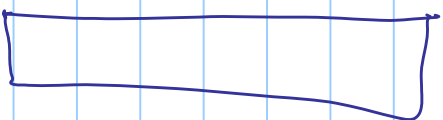
— Use diff. parameters

— want consistent performance

in a small range of parameter space

Absolute magnitudes

→ transform to z-scores



$$\frac{a_{ij} - \text{mean}}{\text{std dev}}$$

mean

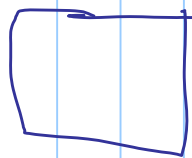
std dev

"1" Stage "1" of dataset

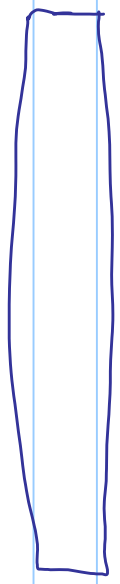
thin



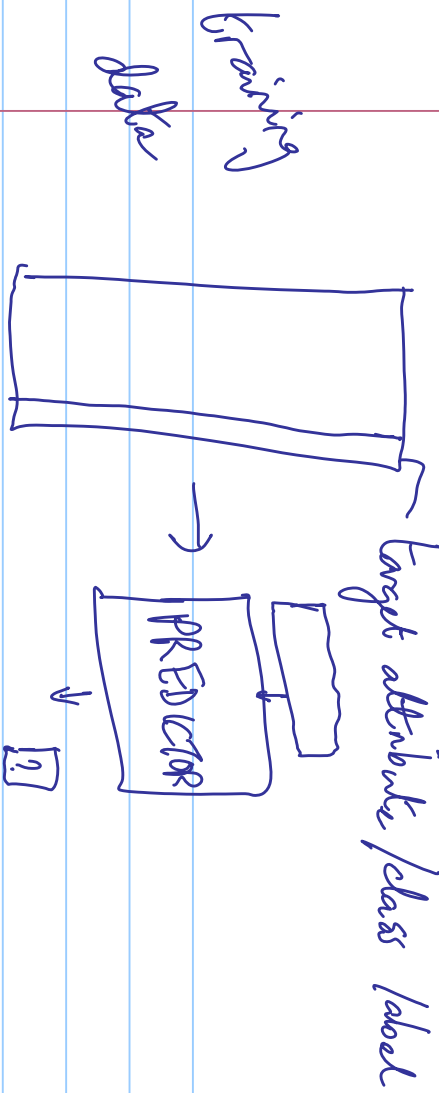
spare



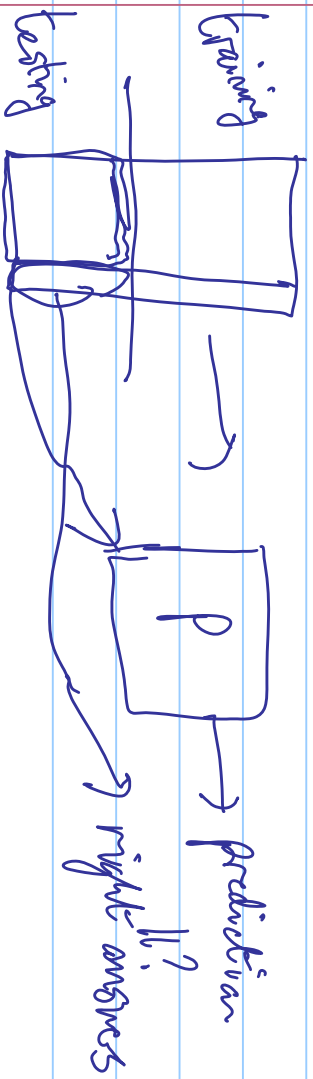
wide



PREDICTION



To see how well a predictor works



Prediction accuracy

Cross validation - repeat with diff splits

of training & test data

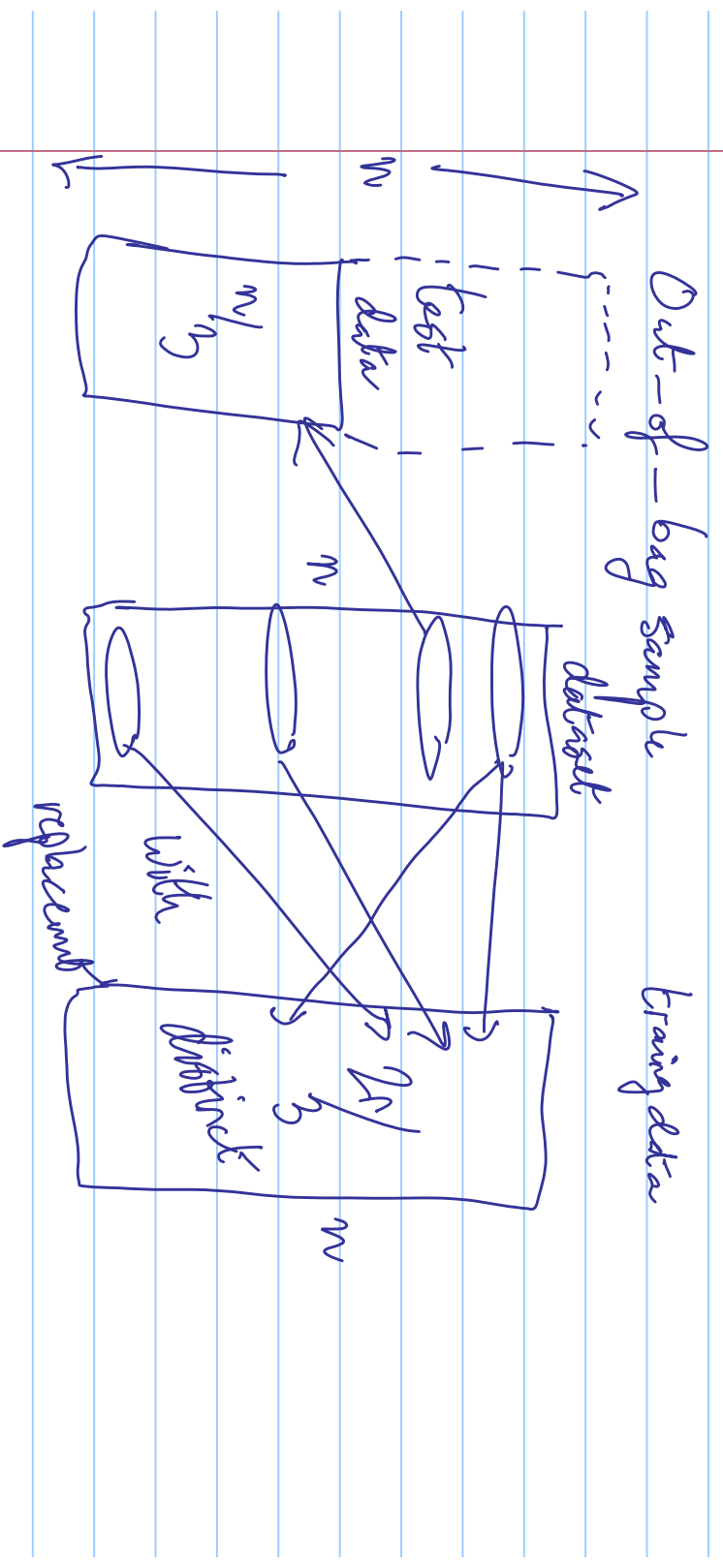
Confusion matrix predicted to be

Is

	A	B
A	100	27
B	25	98

150	2
50	98

RIGHT ANSWERS



Dataset 1

State of the Union - first 4 / president

"word"

0	0	0	0	Integrative
0	0	1	0	Complexity
0	0	0	1	TC

IC - measure complexity of speech

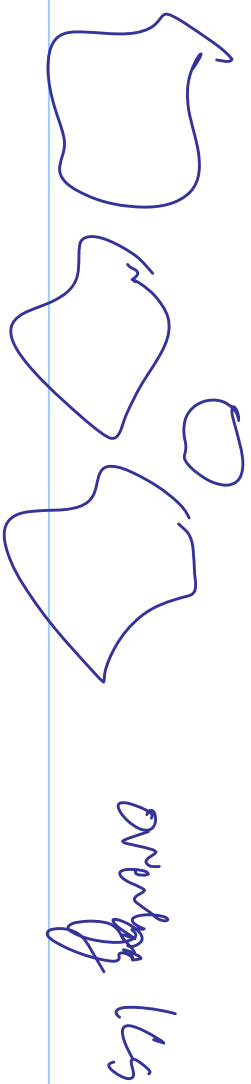
- more than 1 possibility

- extend ideas beyond superficial

Suedfeld UBc

- human scored

clustering



Mission

1. Select your technique
2. be ready to give 10-15 min presentation
next week about your technique
3. 2 weeks away — first presentation on dataset

Problems of language

— normalization of data

3 kinds of words

— common words "the", "is"

— topic of document

— other

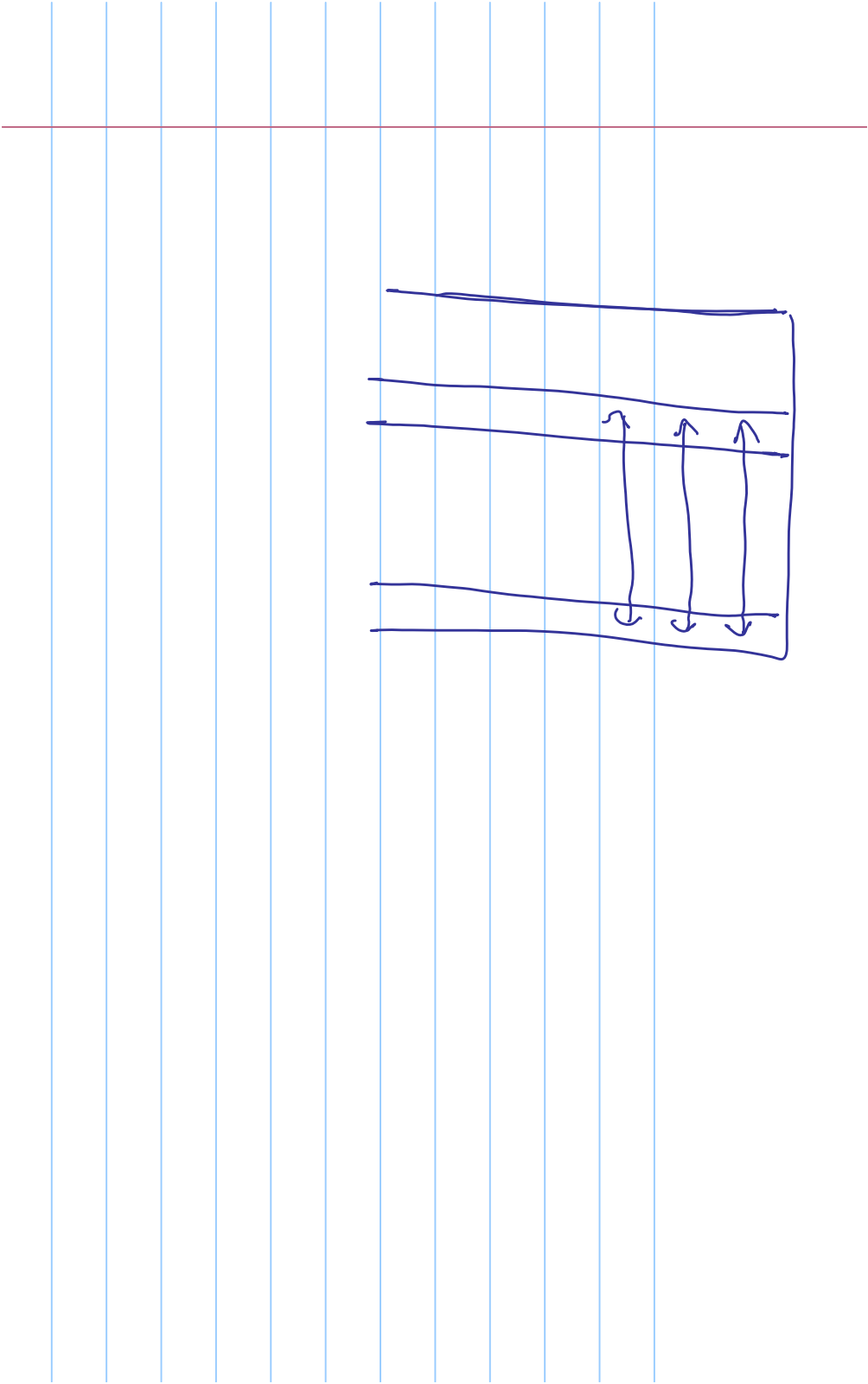
normalize by length of document

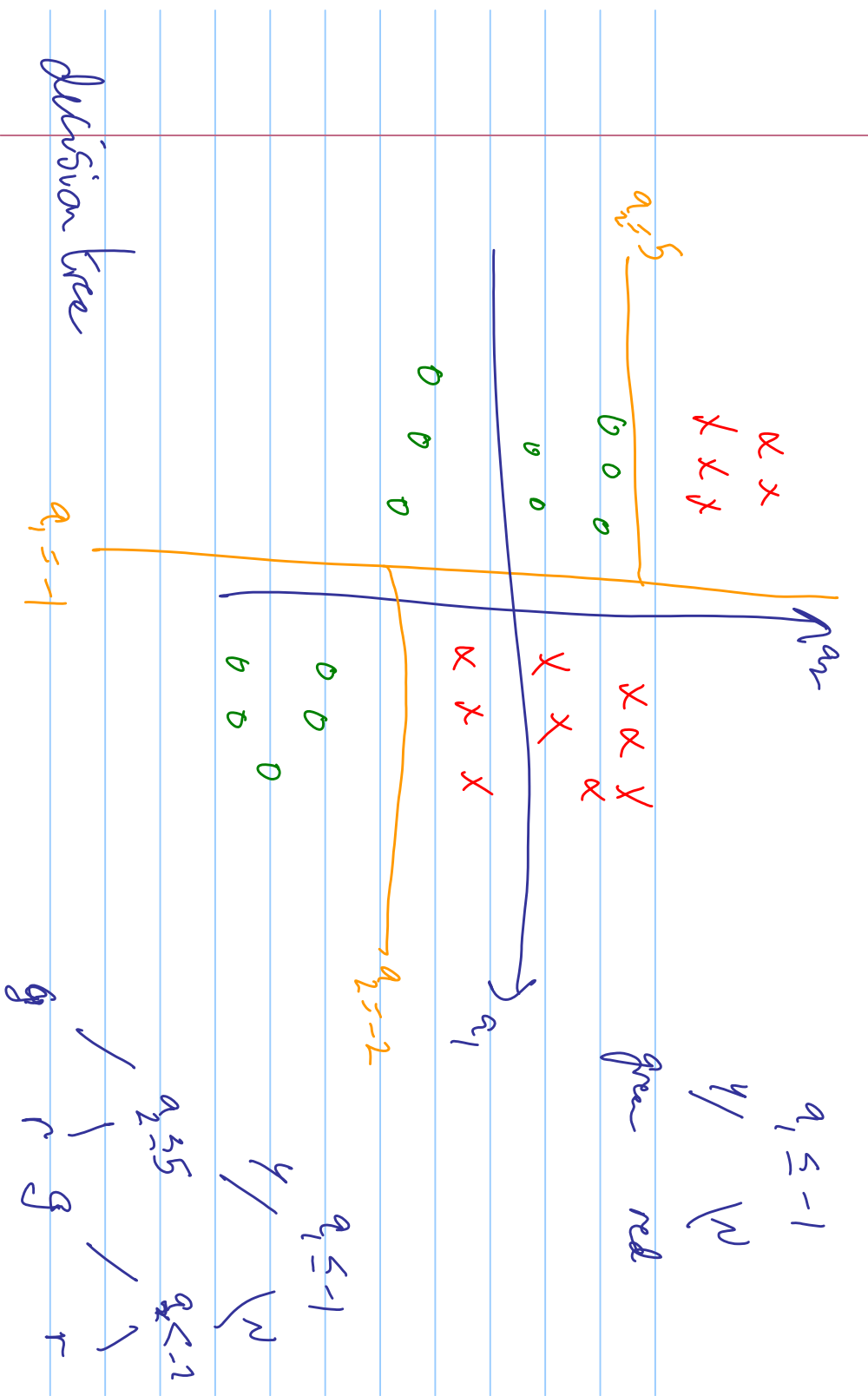
normalize by columns \rightarrow ? 3-grams

! sparse

non-gram 3-grams - mean of non-gram entries
standard

Predictions

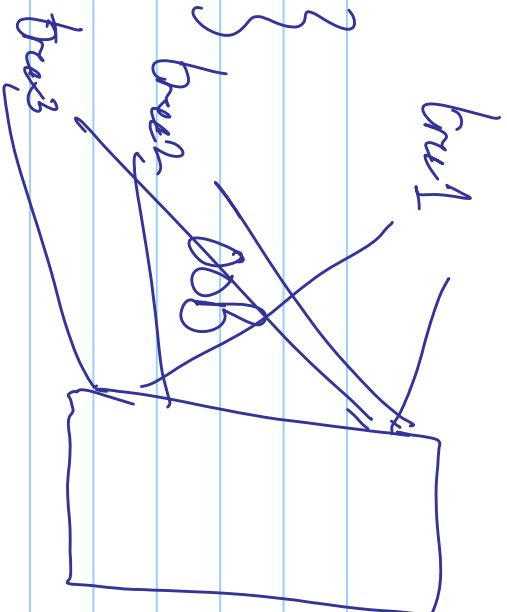




Decision tree

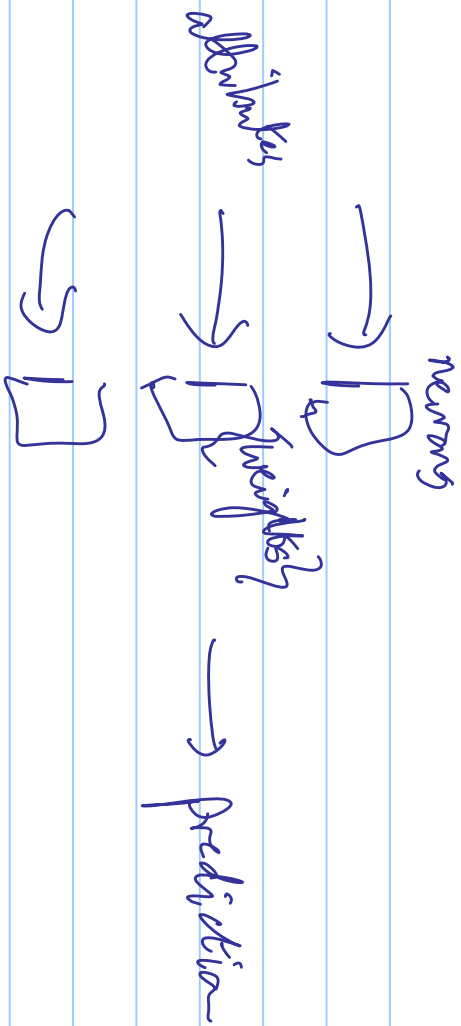
B. Random forests

{ decision trees }

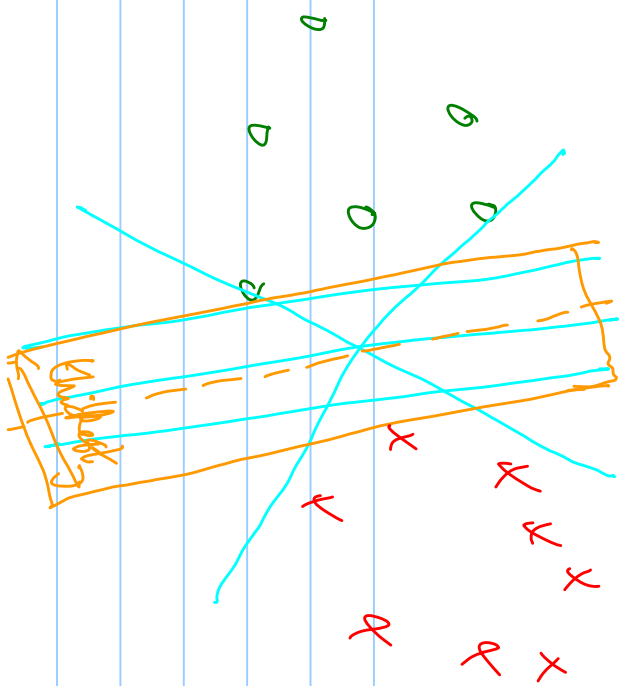


internal node - choose test from a fresh set of attributes - mtry

A. Supervised neural networks



C. Support vector machines



D. Rules

If $a_1 > 3$ & $a_5 = 17$ & $a_{12} < 3$ then class A

⋮

restriction on structure!

E, investigation

Clustering

- high-dimensional spaces are non-intuitive

each attribute: big- v , close to 0, big tr

record in 1000-dimensional space

how likely is it to be close to origin

a_1 close to 0

a_2 close to 0

\vdots

$\left(\frac{1}{3}\right)^{1000}$

a_{1000} close to 0

When is a close to b

a_1 matches b_1 $loss$

a_2 matches b_2 $(1/3)$

\vdots

a_k matches b_k

How good is a clustering?



intra-cluster hi similarity

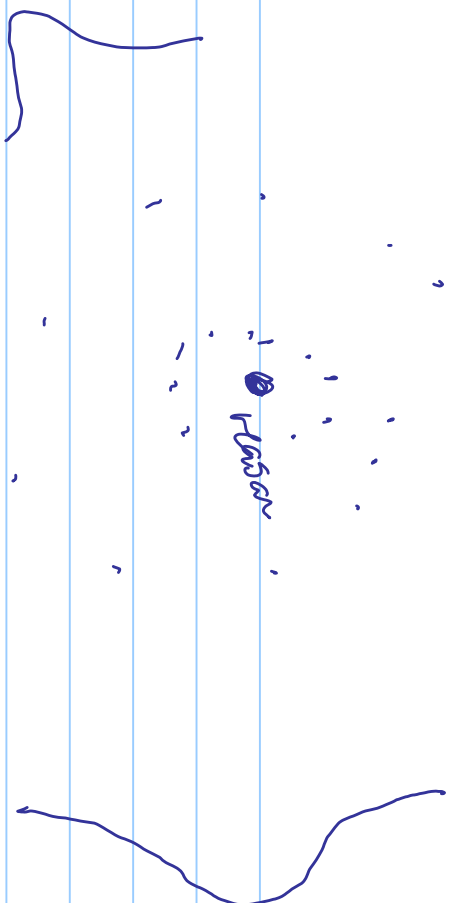
inter-cluster low similarity

3. Hierarchical clustering

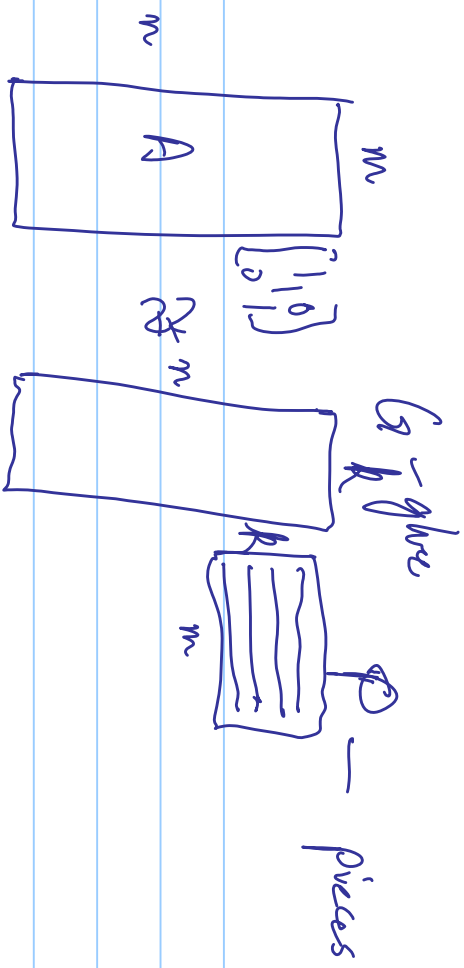
4. distance-based

distribution-based



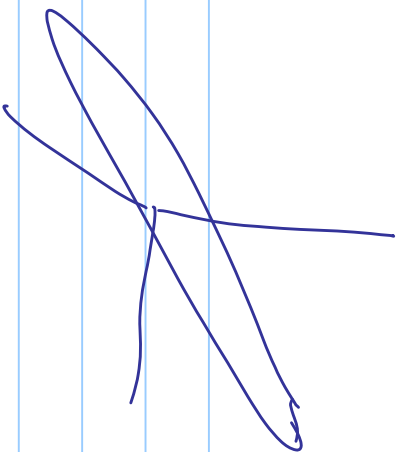


1. independent component analysis
2. singular value decomposition / semi discrete
5. Latent Dirichlet allocation



3 interpretations:

① rows of P are axes, cols of Q are coordinates

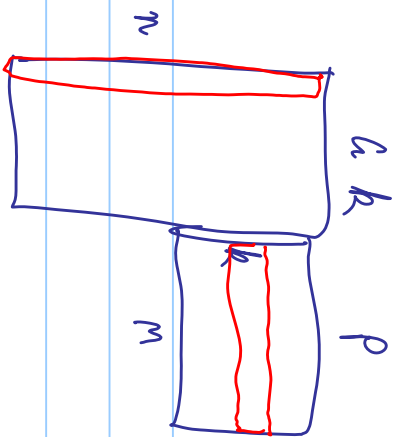


② factor interpretation

rows of P are underlying/latent factors

C mixes them to give the observed factors

③



$$A = \sum_{j=1}^p P_j$$

$$(n \times 1) \times (1 \times m) \rightarrow n \times m$$

