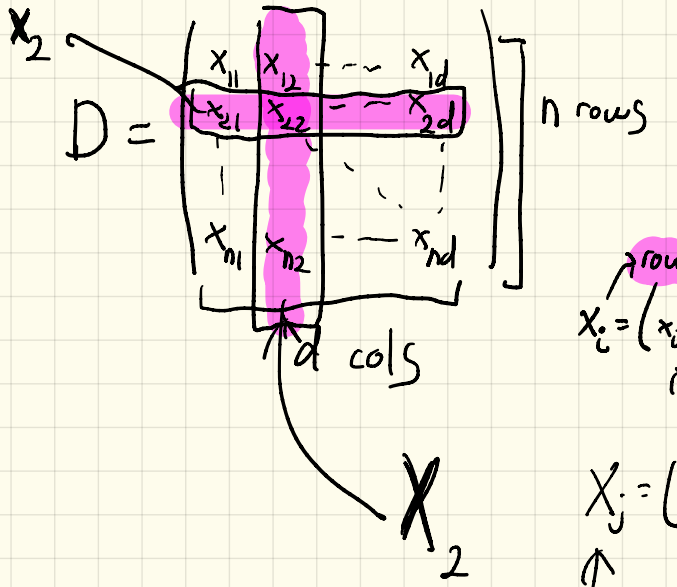
Data, Dist, & Similarity

8/19

Data matrix

$n \times d$

n: # rows; d: # of cols → dim of data



cols are attributes $(X_1, X_2, \dots, X_d)$

rows are observations $(x_1, x_2, \dots, x_n)$

x_i: rows of data matrix

$x_i = (x_{i1}\ x_{i2}\ x_{i3} \dots x_{id})$

↑ 1st attrib of obs i

$X_j = (x_{1j}\ x_{2j} \dots x_{nj})$

↑ j^{th} attrib of obs 2

cols (of data matrix)

univariate analysis:
analysis w/ 1 attrib

bivariate analysis
analysis w/ 2 attrib

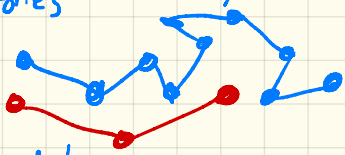
multivariate analysis
analysis w/ more than 2

rows: entities, instances
records, points, tuples

cols:
attribs, properties,
features, variables, fields

	X_1	X_2	X_3	X_4	X_5
x_1	5.9	3.0	4.2	1.5	Iris-versicolor
x_2	6.9	3.1	4.9	1.5	Iris-versicolor
x_3	6.6	2.9	4.6	1.3	Iris-versicolor
x_4	4.6	3.2	1.4	0.2	Iris-setosa
x_5	6.0	2.2	4.0	1.0	Iris-versicolor
x_6	4.7	3.2	1.3	0.2	Iris-setosa
x_7	6.5	3.0	5.8	2.2	Iris-virginica
x_8	5.8	2.7	5.1	1.9	Iris-virginica

Questions!
- what is x_3?
- what is X_4?
- what is the dim of this data? 5
- what is the size of this data 8
- Give 1 example of a dataset that doesn't easily fit into a data matrix
  - DNA
  - words / key phrase / n-grams
  - trajectories



  - image data

turn into something fits into data matrix
feature extraction

Note Most techniques assume each entity is ind - prob not true most of the time
Sometimes rep as data graph
nodes - entites
edges - relationships between entities