DATA INVESTIGATION APPROACHES

(Adapted from Data Science for Business by Provost and Fawcett, pp. 20-23)

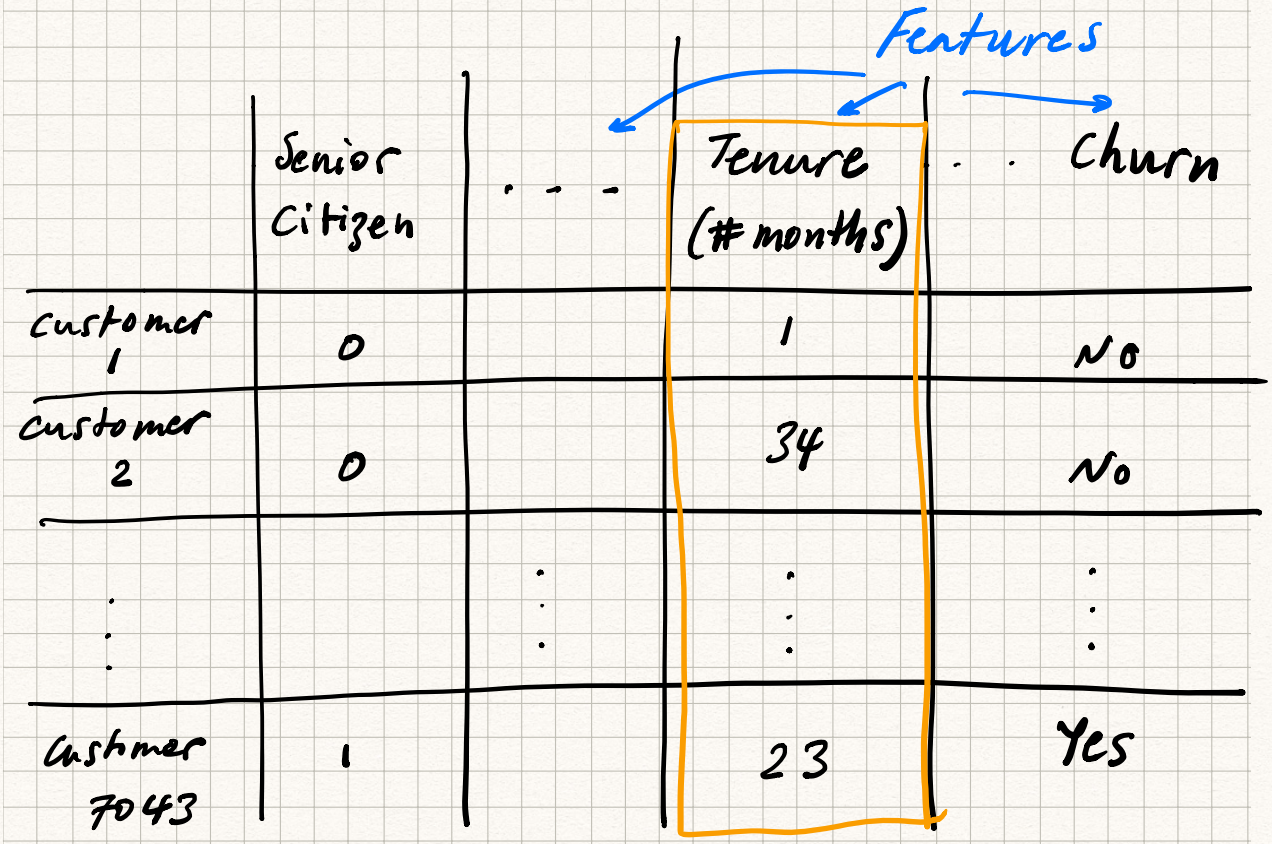
DATA = TABLE OF VALUES

Telecom churn data set

Features

	Senior Citizen	. . .	Tenure (#months)	. . .	Churn
Customer 1	0		1		No
Customer 2	0		34		No
⋮		⋮	⋮		⋮
Customer 7043	1		23		Yes

PATTERN 1: REGRESSION



Predict the tenure of a customer given the values of all the other features for that customer.

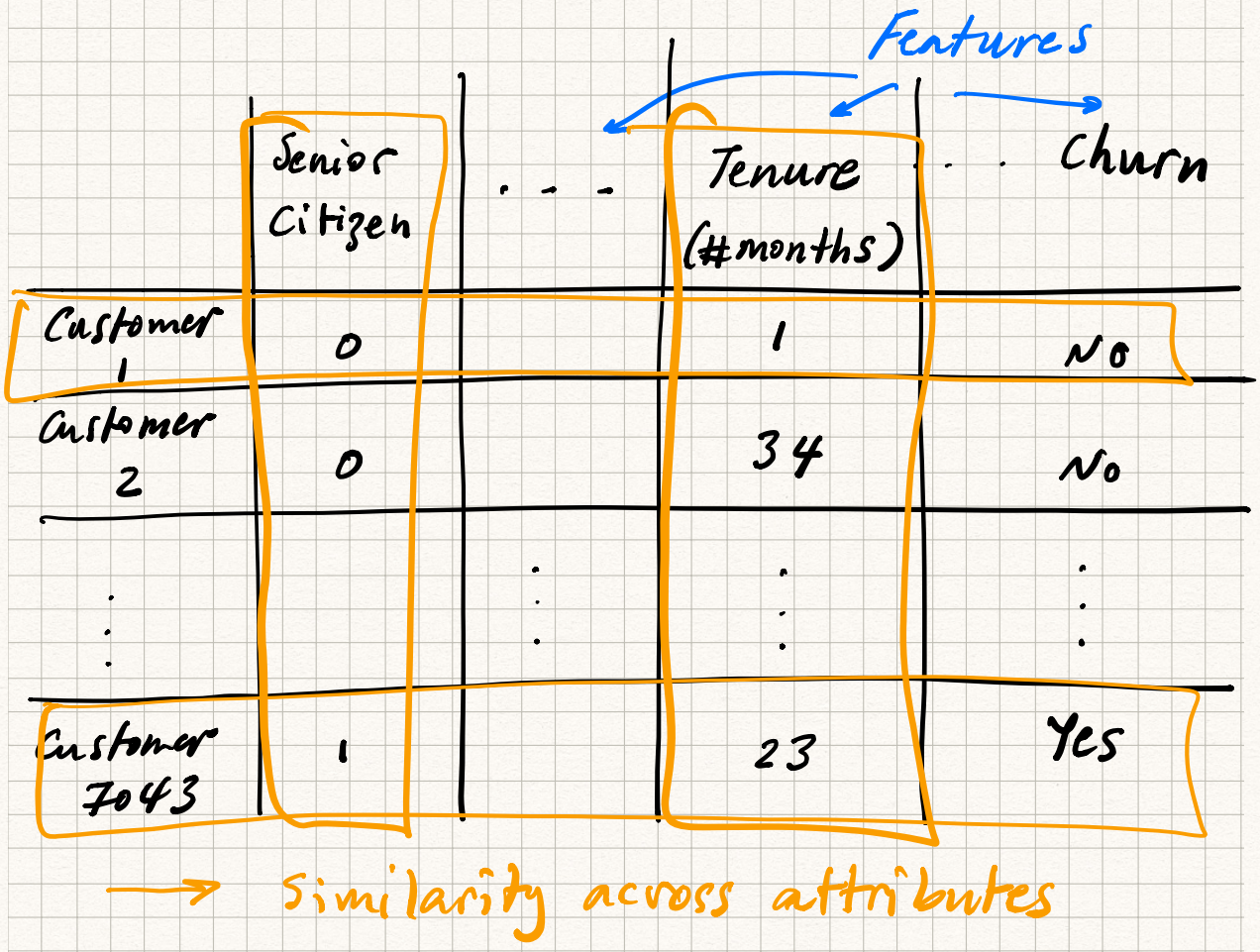
PATTERN 2: CLASSIFICATION

Features

	Senior Citizen	. . .	Tenure (#months)	. . .	Churn
Customer 1	0		1		No
Customer 2	0		34		No
⋮		⋮	⋮		⋮
Customer 704	1		23		Yes

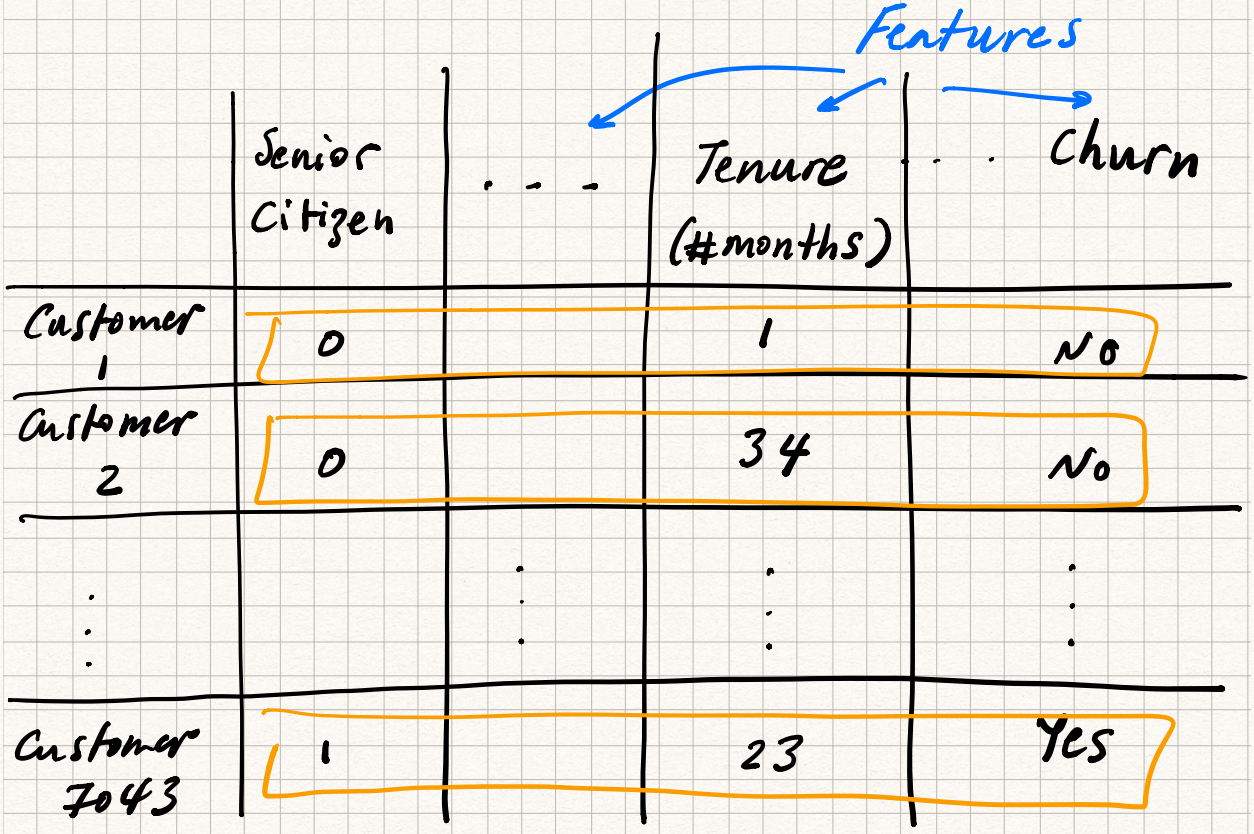
Predict if a customer will churn or not given the values of all the other features for that customer.

PATTERN 3 : SIMILARITY



Given a set of features and a value or range of values for each feature, which Customers are similar?

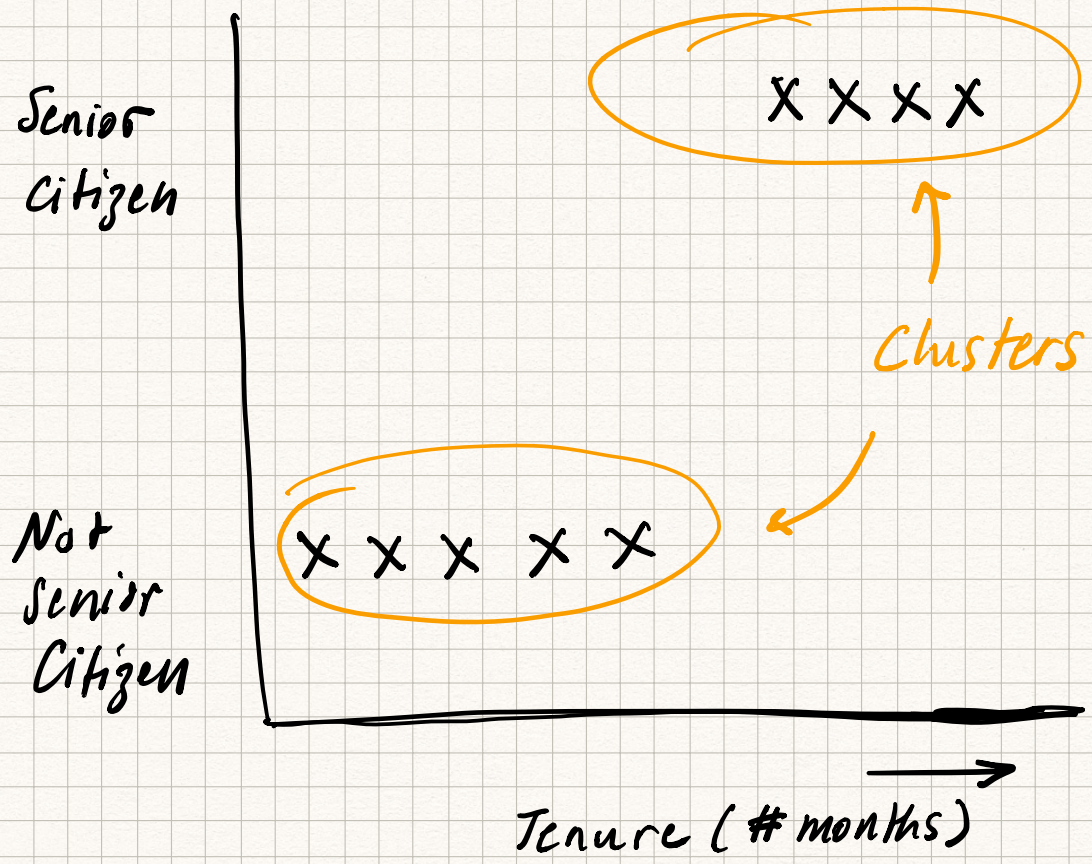
PATTERN 4: CLUSTERING



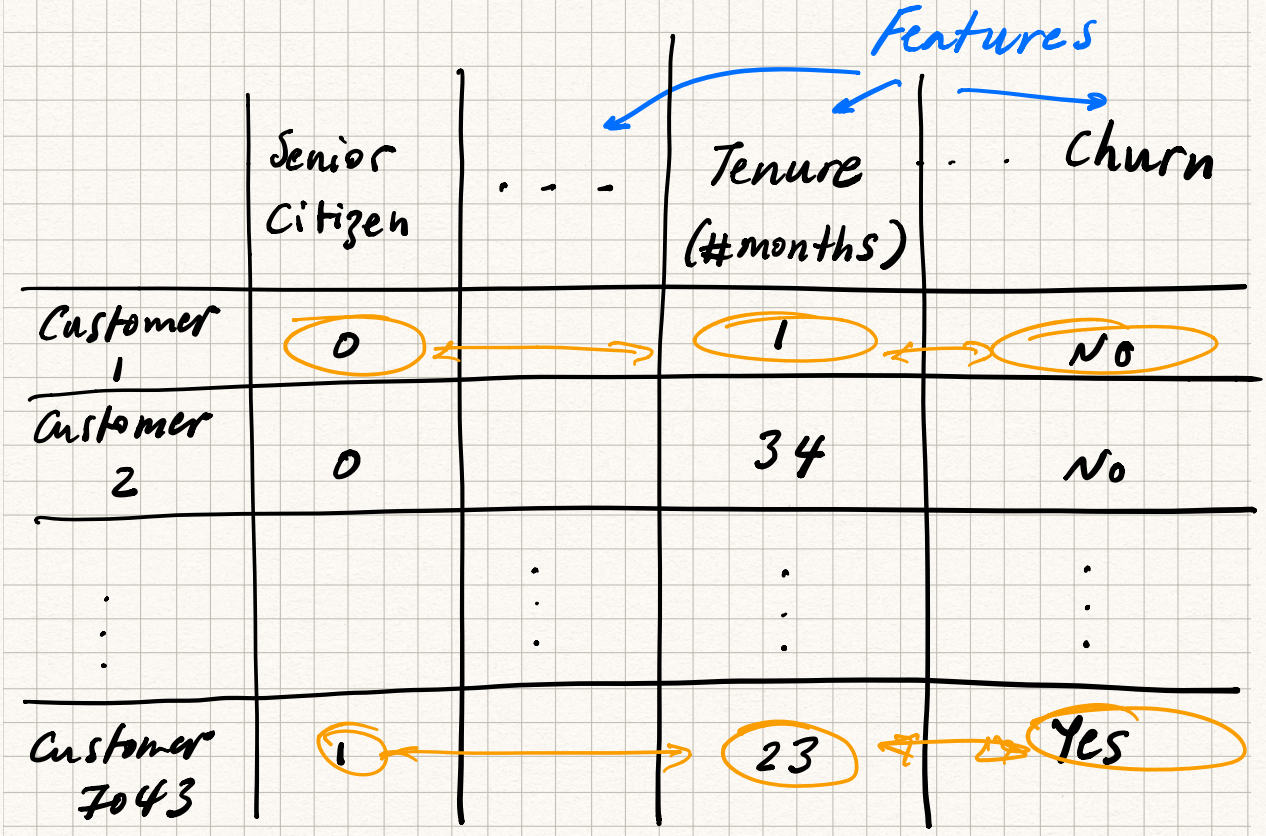
Do customers cluster together?

(on some or all attributes)

PATTERN 4: CLUSTERING EXAMPLE



PATTERN 5: CO-OCCURRENCE (ACROSS ATTRIBUTES)



- Which feature values always/often occur together?
- Which feature values never/rarely occur together?

PATTERN 5: CO-OCCURRENCE (IN THE SAME ATTRIBUTE)

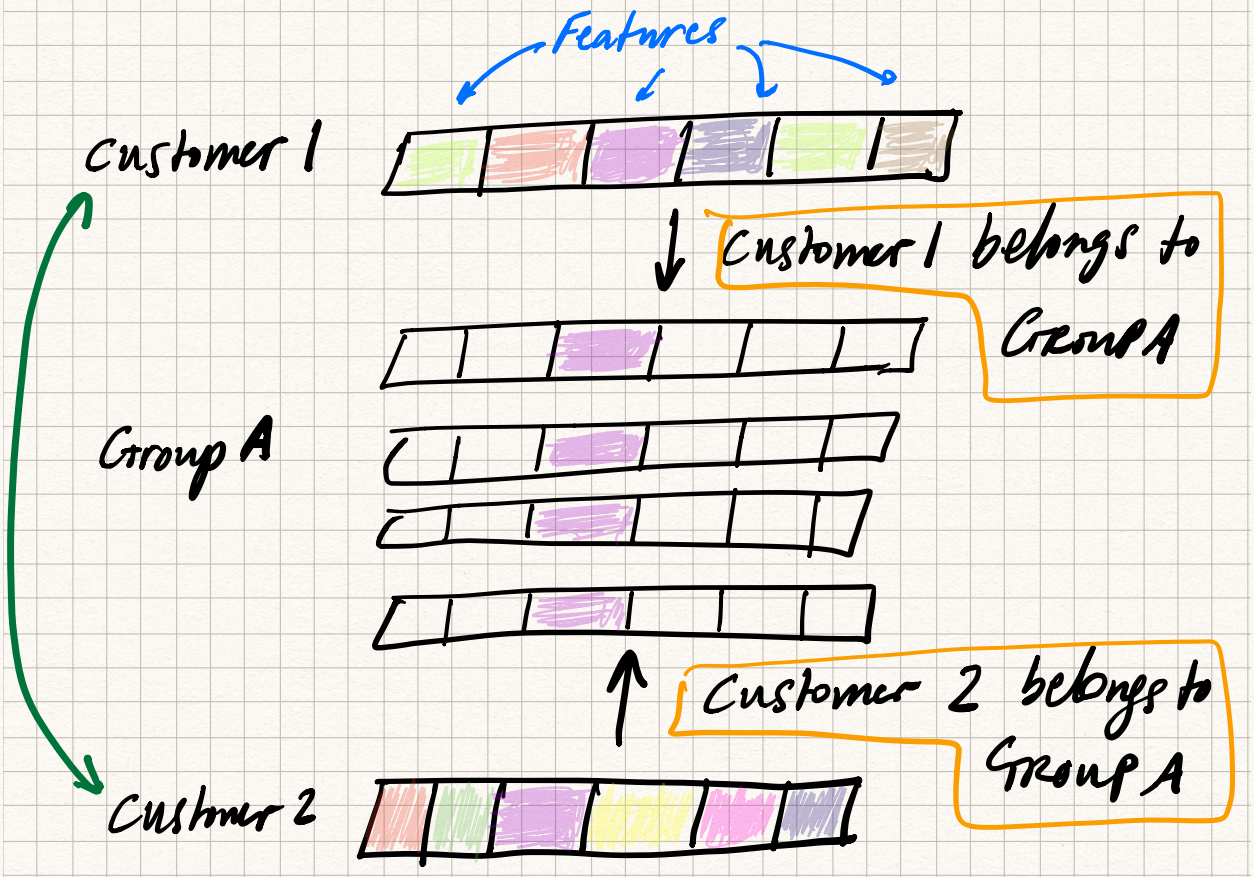
Features

	Senior Citizen	Friends With	Tenure (#months)	...	Churn
Customer 1	0	16, 22, 45	1		No
Customer 2	0	16, 22, 260	34		No
⋮		⋮	⋮		⋮
Customer 7043	1	12, 4264	23		Yes

Customers 1 and 2 have 2 friends in common.

- Which values occur always/often together in the same feature?
- Which values never/rarely occur together in the same feature?

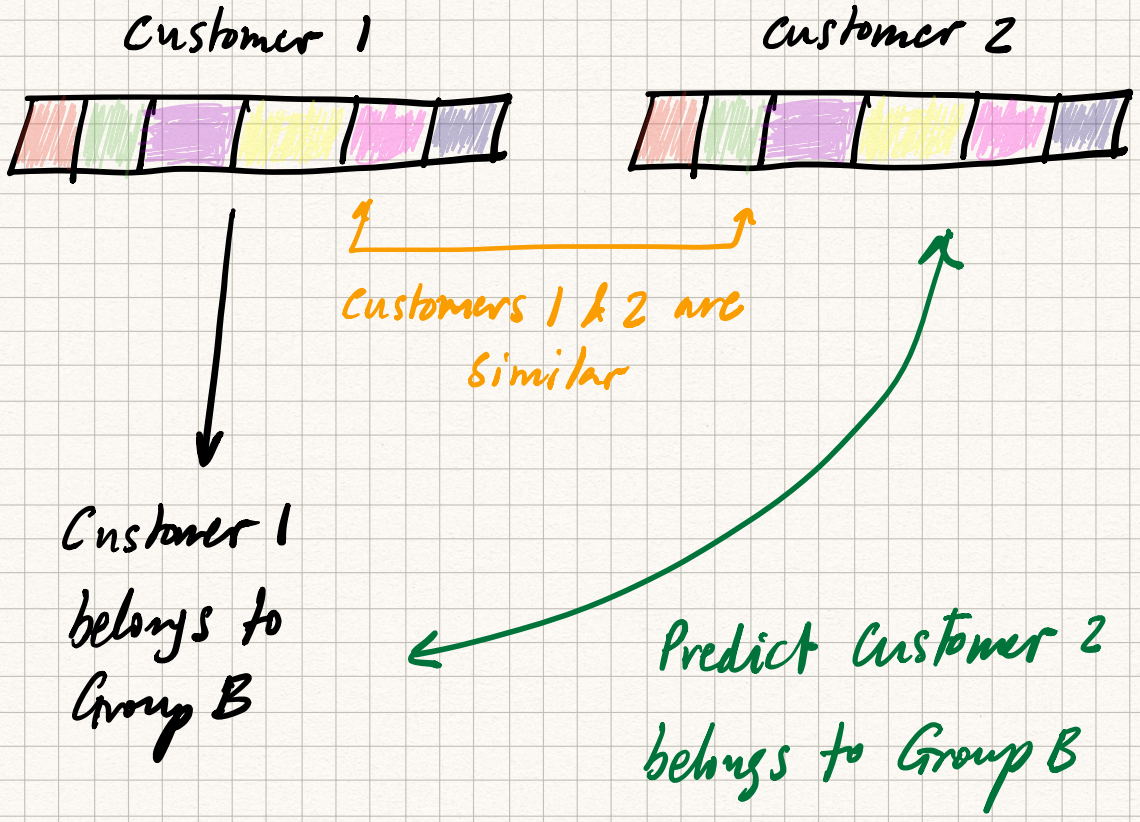
PATTERN 6: LINK PREDICTION THROUGH GROUP SIMILARITY



Group A: The set of customers who share the same 3 friends.

Predict that customers 1 and 2 will "respond" similarly or have the same interests - they are Linked.

PATTERN 6: LINK PREDICTION THROUGH INDIVIDUAL SIMILARITY



Group B: The set of customers who responded to the last marketing campaign.

PATTERN 7: FEATURE ENGINEERING

- Which features can be ignored/deleted?
  - We'll look at some techniques later based on total variance explained
- Which features need creation?
  - Combination of existing features (e.g., monthly charge / tenure)
  - Invention of new features (e.g., # of senior citizen friends)
  - Extension of existing features (e.g., $tenure^2$, log(monthly charge))