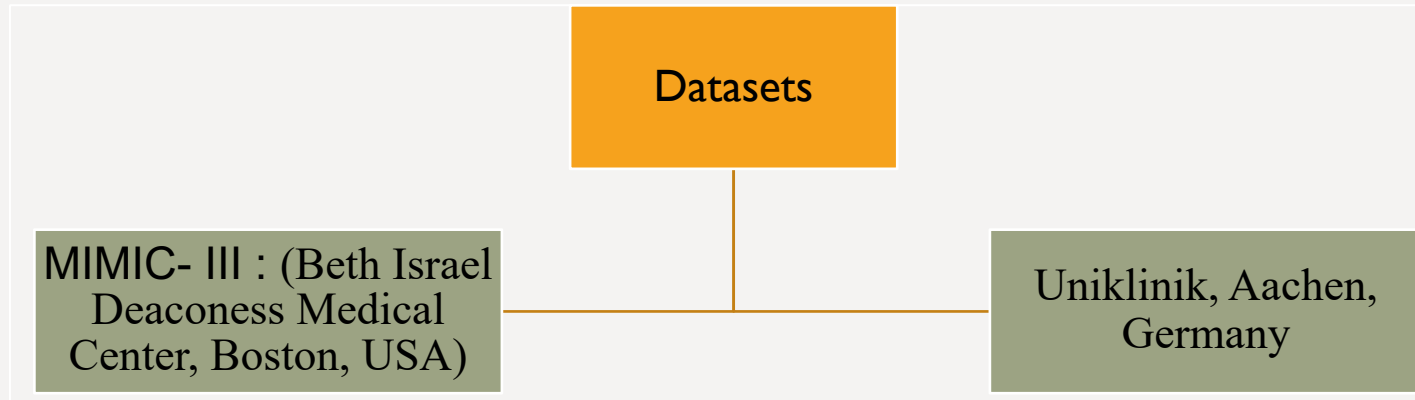


FEATURE SELECTION FOR PATIENT POPULATION DISCRIMINATION

Kumari Shikha

Supervisor : Jayesh Bhat



Data :

- Nearly 17900 patients from each dataset
- 100 features

Questions :

- If we pick up a patient out of these 35800, can we figure out which dataset he/she belongs to?
- Do all 100 features matter?
- What do "**non discriminating**" features tell us?

MY THESIS WORK : INVESTIGATE THE HETEROGENEITY
BETWEEN TWO PATIENT POPULATIONS



The problem is posed as a
**binary classification
problem**



**Feature-set subsets with
high classification
accuracy** are discriminatory
and can be used to predict
the data source of a
randomly drawn patient
from the two populations.



**Subsets that generate
accuracy close to chance**
because classifiers for them
cannot distinguish the data
source of a randomly-drawn
patient.

