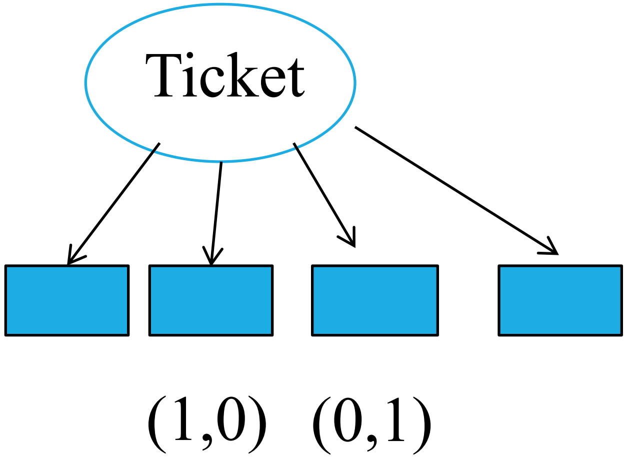
# GAIN RATIO

SYRACUSE UNIVERSITY
School of Information Studies

# GAIN RATIO

Impurity measures tend to favor attributes that have a large number of distinct values (textbook, p. 163).

- E.g., the "ticket" attribute in the Titanic data set means the ticket number. Assuming every passenger has a unique ticket number, the ticket attribute has many distinct values, and impurity measures such as IG favor such attributes.



# GAIN RATIO

What to do?

- Use domain knowledge: Does ticket number have anything to do with survival chance?
- Use gain ratio, which is IG divided by “split info.”
  - “Split info” is a penalty to a large number of splits.

In J48, the information gain measure has taken steps to avoid choosing the “ticket” types of attributes.