# Discrimination Discovery

# The discrimination discovery task at a glance

Given **a large database of historical decision records,**
find **discriminatory situations and practices.**

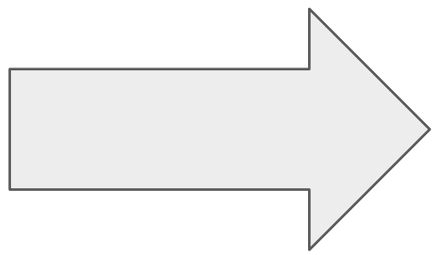
# Discrimination discovery scenario

INPUT

Database of historical *decision records*

A *criterion* of (unlawful) discrimination

A set of *potentially discriminated* groups



OUTPUT

A subset of *decision records* and *potentially discriminated* people for which the *criterion* holds

*D. Pedreschi, S. Ruggieri and F. Turini (2008). Discrimination-Aware Data Mining. In KDD, pp. 560-568.*

# The German credit score dataset

A small dataset used in many papers about discrimination (like Zachary's karate club for networks people)

**N = 1,000** records of bank account holders

**Class label**: good/bad creditor (grant or deny a loan)

**Attributes**: *numeric/interval-scaled*: duration of loan, amount requested, number of installments, age of requester, existing credits, number of dependents; *nominal*: result of past credits, purpose of credit, personal status, other parties, residence since, property magnitude, housing, job, other payment plans, own telephone, foreign worker; *ordinal*: checking status, saving status, employment

*German credit score dataset:*
*https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)*

# Defining potentially discriminated (PD) groups

A subset of attribute values are **perceived as potentially discriminatory** based on  background knowledge.

Potentially discriminated groups are people with those attribute values.

Example:

- Women (misogyny)
- Ethnic minority (***racism***) or minority language
- Specific age range (***ageism***)
- Specific sexual orientation (***homophobia***)

# Discrimination and combinations of attribute values

Discrimination can be a result of several joint characteristics (attribute values)  which are not discriminatory by themselves

Thus, the object of discrimination should be described by a conjunction of attribute values:

**Known as Itemsets**

# Association and classification rules

Association rules are if/then statements that help uncover relationships between  seemingly unrelated data in a relational database.

In a classification rule, Y is a class item and X contains no class items.

$$\mathbf{X \rightarrow Y}$$

# Definition: Association Rule

Let **D** be database of transactions e.g.

| Transaction ID | Items |
|---|---|
| 2000 | A, B, C |
| 1000 | A, C |
| 4000 | A, D |
| 5000 | B, E, F |

- Let ***I*** be the set of items that appear in the database, e.g., ***I = {A,B,C,D,E,F}***

- A **rule** is defined by **X → Y**, where **X⊂*I***, **Y⊂*I***, and **X∩Y=∅**

  – e.g.: **{B,C} → {A}** is a rule

# Definition: Association Rule

- Association Rule
  - An implication expression of the form **X** → **Y**, where **X** and **Y** are non-overlapping itemsets
  - Example:
    ***{Milk, Diaper}*** → ***{Beer}***

| *TID* | *Items* |
|---|---|
| 1 | Bread, Milk |
| 2 | Bread, Diaper, Beer, Eggs |
| 3 | Milk, Diaper, Beer, Coke |
| 4 | Bread, Milk, Diaper, Beer |
| 5 | Bread, Milk, Diaper, Coke |

**Example:**

$$\{\text{Milk, Diaper}\} \rightarrow \text{Beer}$$

- Rule Evaluation Metrics
  - Support (s)
    - Fraction of transactions that contain both **X** and **Y**
  - Confidence (c)
    - Measures how often items in **Y** appear in transactions that contain **X**

$$s = \frac{\sigma(\text{Milk, Diaper, Beer})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk, Diaper, Beer})}{\sigma(\text{Milk, Diaper})} = \frac{2}{3} = 0.67$$

# Computing support and confidence

| TID | date | items_bought |
|---|---|---|
| 100 | 10/10/99 | {F,A,D,B} |
| 200 | 15/10/99 | {D,A,C,E,B} |
| 300 | 19/10/99 | {C,A,B,E} |
| 400 | 20/10/99 | {B,A,D} |

What is the ***support*** and ***confidence*** of the rule: {B,D} → {A}

- Support: percentage of tuples that contain {A,B,D} = 75%

- Confidence:

$$\frac{\text{number of tuples that contain } \{A,B,D\}}{\text{number of tuples that contain } \{B,D\}} = 100\%$$

# Association-rule mining task

Given a set of transactions **D**, the goal of association rule mining is to find **all** rules having

- support ≥ ***minsup*** threshold
- confidence ≥ ***minconf*** threshold

Beyond the scope of the current course!

# Direct discrimination

Direct discrimination implies rules or procedures that impose ‘disproportionate burdens’ on minorities

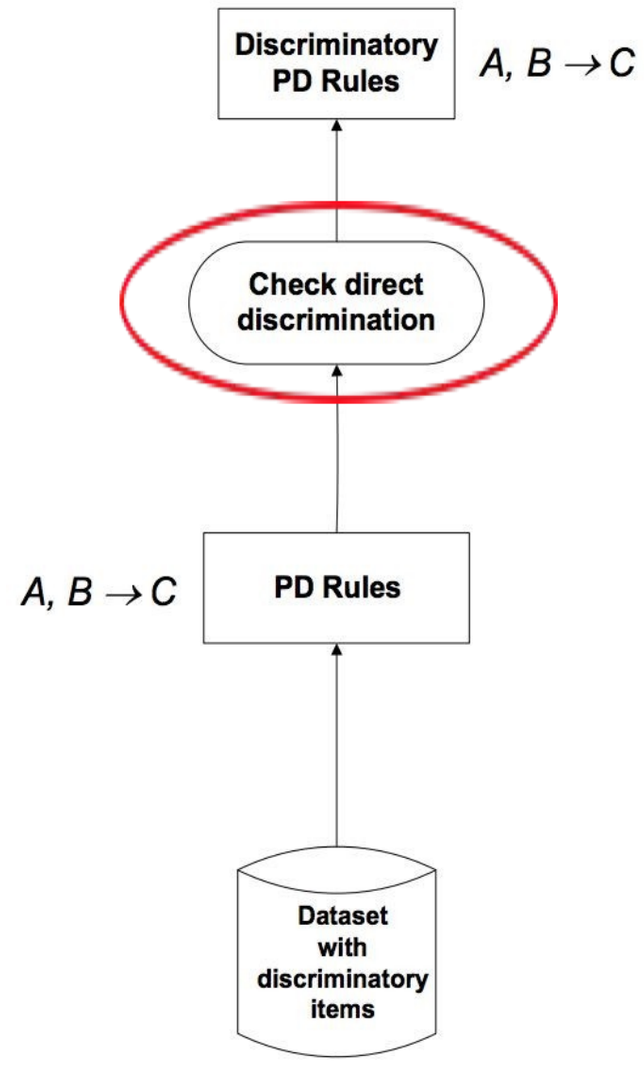
Potentially Discriminatory (PD) rules are any classification rule of the form:

A, B → C

where A is a PD group (B is called a "context")

Example:

gender="female", saving_status="no known savings” → credit=no

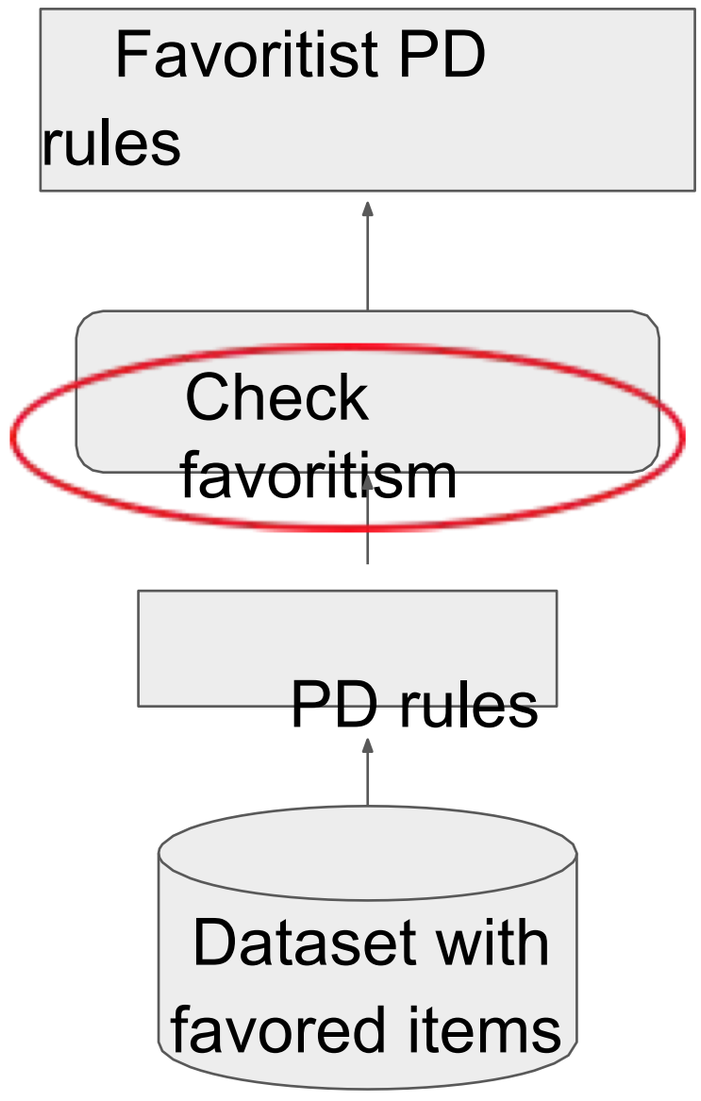


# Favoritist PD rules

Is unveiled by looking at PD rules of the form

$$A, B \rightarrow C$$

where C grants some benefit and A refers to a favored group.




**Example:**

gender="male", savings="no known savings" → credit=yes

# Evaluating PD rules through the extended lift

Remembering that conf(X → Y) = support(X → Y) / support(X)

We define the **extended lift with respect to B** of rule A, B → C as:
$elift_B$(A, B → C) = conf(A, B → C) / conf(B → C)

The rules we care about are PD rules such that:

- A is a protected group (e.g. female, black)
- B is a context (e.g. lives in San Francisco)
- C is an outcome (usually negative, e.g., deny a loan)

# The concept of $\alpha$-protection

For a given threshold α, we say that PD rule A, B → C, involving a PD group A in a context B for an outcome C, is α-protective if:

$elift_B(A, B \rightarrow C) = conf(A, B \rightarrow C) / conf(B \rightarrow C) \leq \alpha$

Otherwise, when $elift_B(A, B \rightarrow C) > \alpha$, then we say that

A, B → C is an α-discriminatory rule

# Relation of $\alpha$-protection and group representation

For a given threshold α, we say that PD rule A, B → C,

involving a PD group A in a context B for a (usually bad) outcome C, is α-protective if:

$elift_B(A, B \rightarrow C) = conf(A, B \rightarrow C) / conf(B \rightarrow C) \leq \alpha$

Note that:

$elift_B(A, B \rightarrow C) = elift_B(B, C \rightarrow A) = conf(B, C \rightarrow A) / conf(B \rightarrow A)$

This means extended lift is the ratio between the proportion of the disadvantaged group A in context B for (bad) outcome C, over the overall proportion of A in B.

# Direct discrimination example

Rule (a):

city="NYC"
→ benefit=deny
with confidence 0.25

Rule (b):

race="black", city="NYC"
→ benefit=deny
with confidence 0.75
**elift 3.0**

Additional (discriminatory) element increases the rule confidence up to 3 times.

According to $\alpha$-protection method, if the threshold $\alpha$=3 is fixed then the rule (b) is  classified as discriminatory

# Real-world example from German credit dataset

Fixing α=3:

(B) saving status = "no known savings" → credit = deny

conf.
0.18

(A) personal status = "female div/sep/mar", saving status = "no known savings" → credit = deny

conf. **elift**
0.27 **1.52**

**Rule A is α-protective.**

# Real-world example from German credit dataset

Fixing α=3:

(B) purpose = "used car" → credit = deny  conf. 0.17

(A) age = "52.6+", personal status = "female div/sep/mar", purpose = "used car" → credit = deny  conf. 1.00  **elift 6.06**

**Rule A is α-discriminatory.**