Model-based Discrimination Analysis: A Position Paper

Qusai Ramadan[1], Shayan Ahmadian[1], Daniel Strüber[1],
Jan Jürjens[1,2] and Steffen Staab[1,3]

[1] University of Koblenz-Landau, Koblenz, Germany

[2] Fraunhofer-Institute for Software and Systems Engineering ISST, Dortmund, Germany

[3]University of Southampton, UK

Tuesday 29th May 2018, Sweden, Gothenburg

Basics

- Decision-making software may lead to undesirable discrimination:
  - exploiting sensitive data (e.g., race)
  - learning correlations between a set of data.

1. Black box approaches
2. White box approaches

} Require the software to be implemented.

How to discover potential discrimination during the software design phase?
(i.e., before having a faulty implementation)

Motivating Example 1

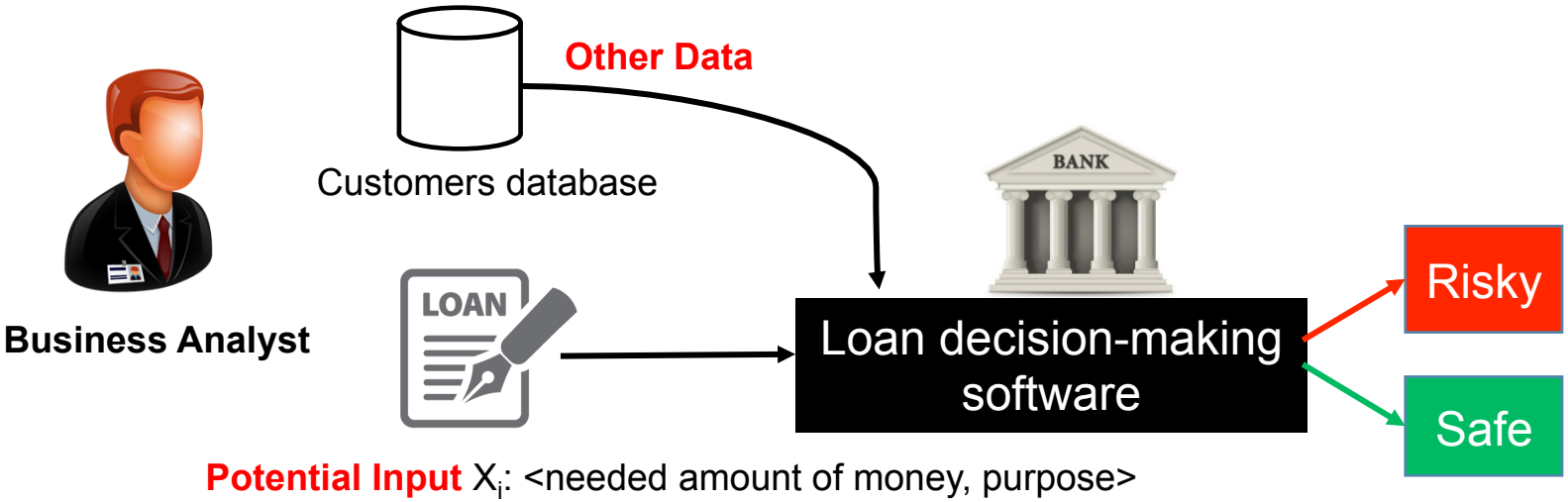
- Bank offers three services
  - Zero-Fee Money Transfer (for international merchants)
  - Vacancies Announcement (for domestic persons educated in accounting).
  - Apply for a loan.

- **Policy**: The bank disallows discriminating between the **loans applicants** based on their **citizenship**.

Business Analyst

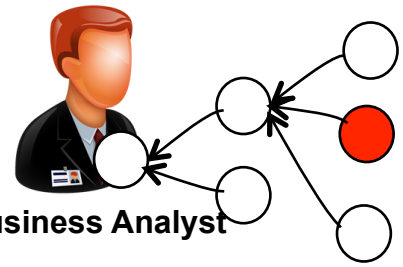
Input X_i: <needed amount of money, purpose>

Motivating Example 1



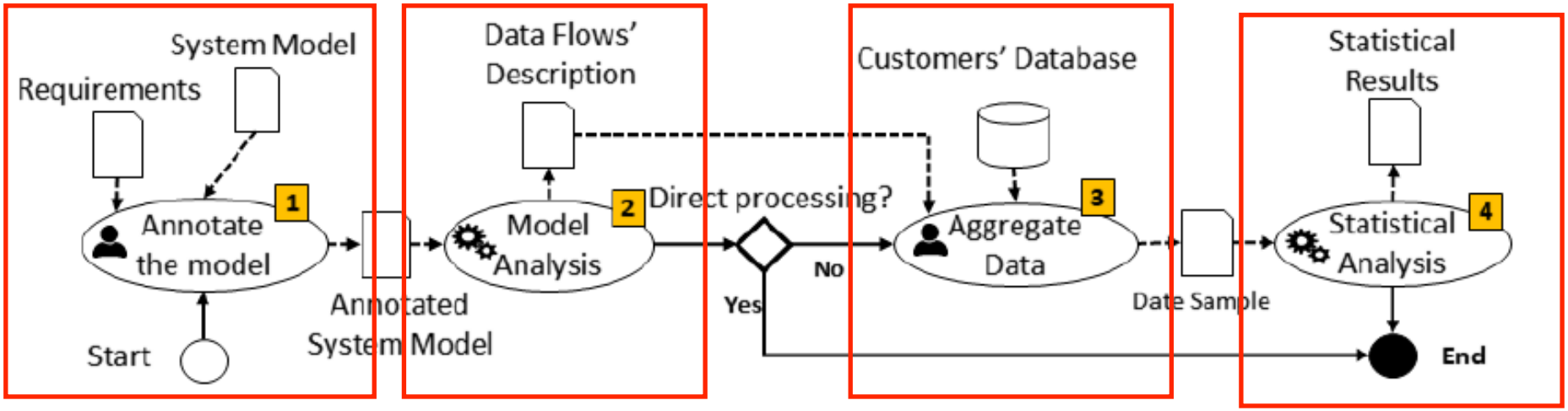
Potential Input X_i: <needed amount of money, purpose>

Roadmap: Model-based Analysis



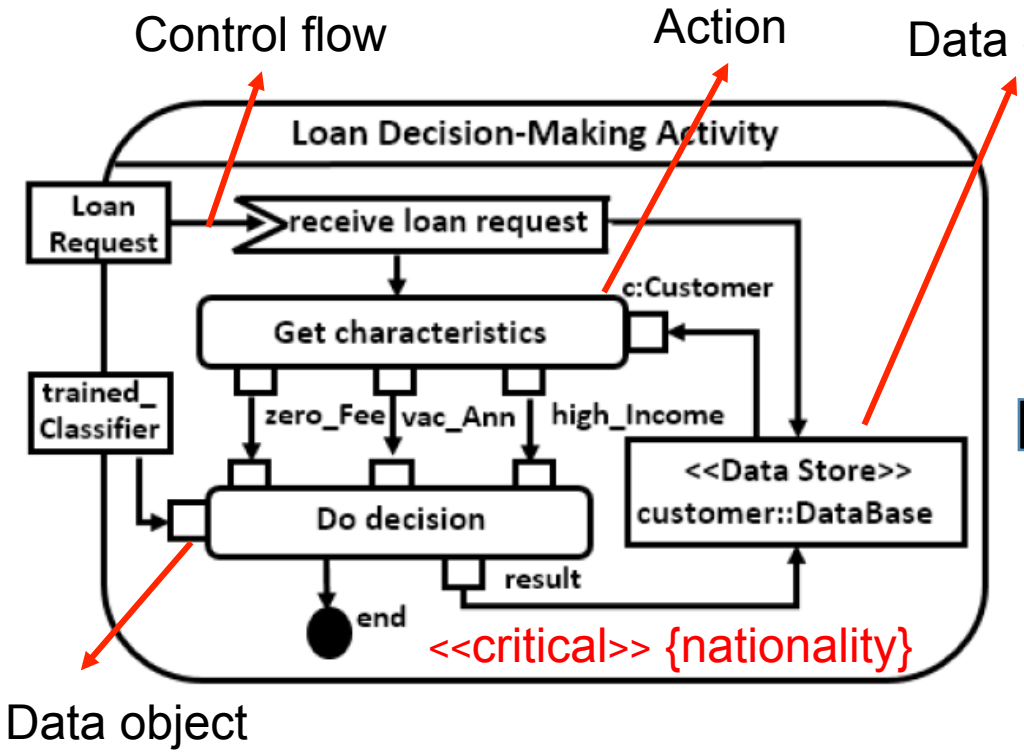
A Domain Expert and a Business Analyst

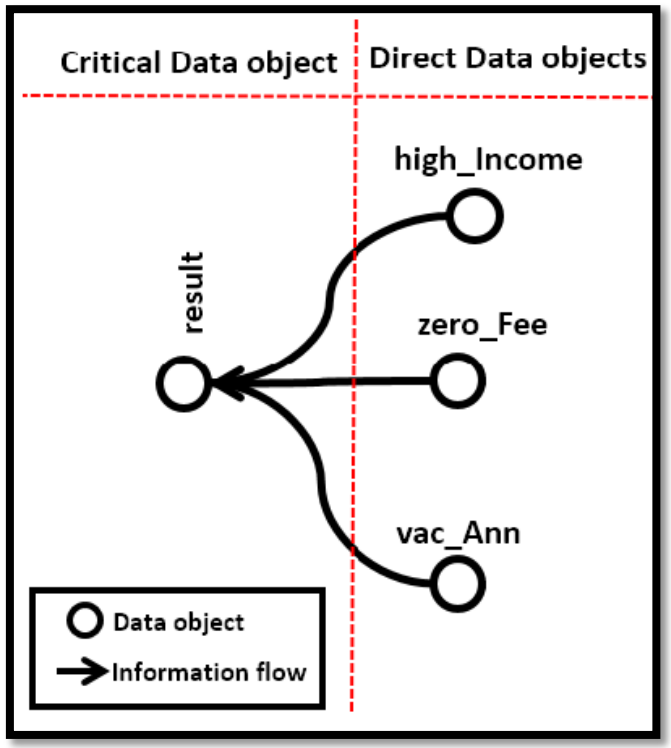
Business Analyst



High-level overview of a model-based discrimination analysis framework.

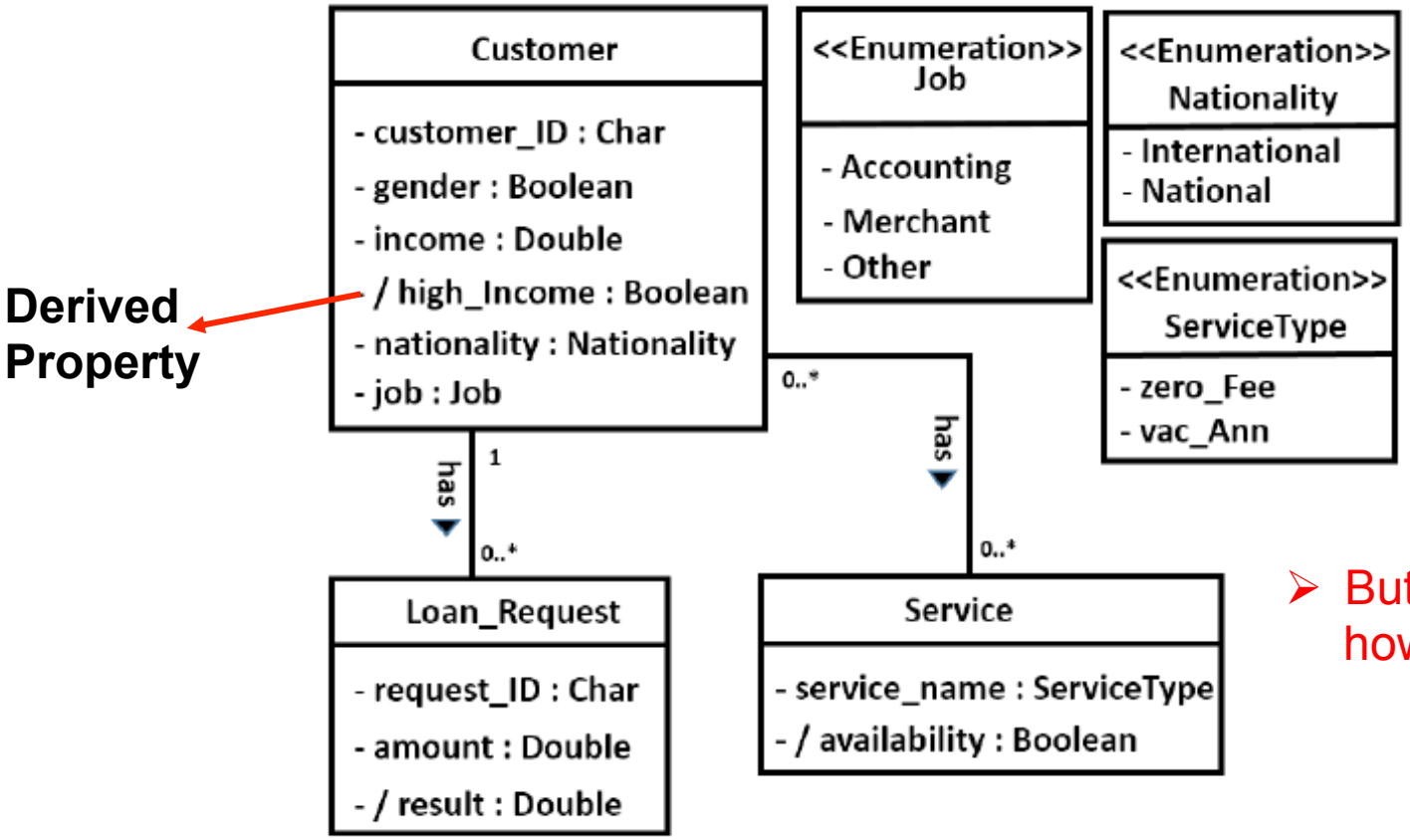
Information Flow Analysis





Database Schema

- Information about whether a data object is derived or not can be represented in the database schema.

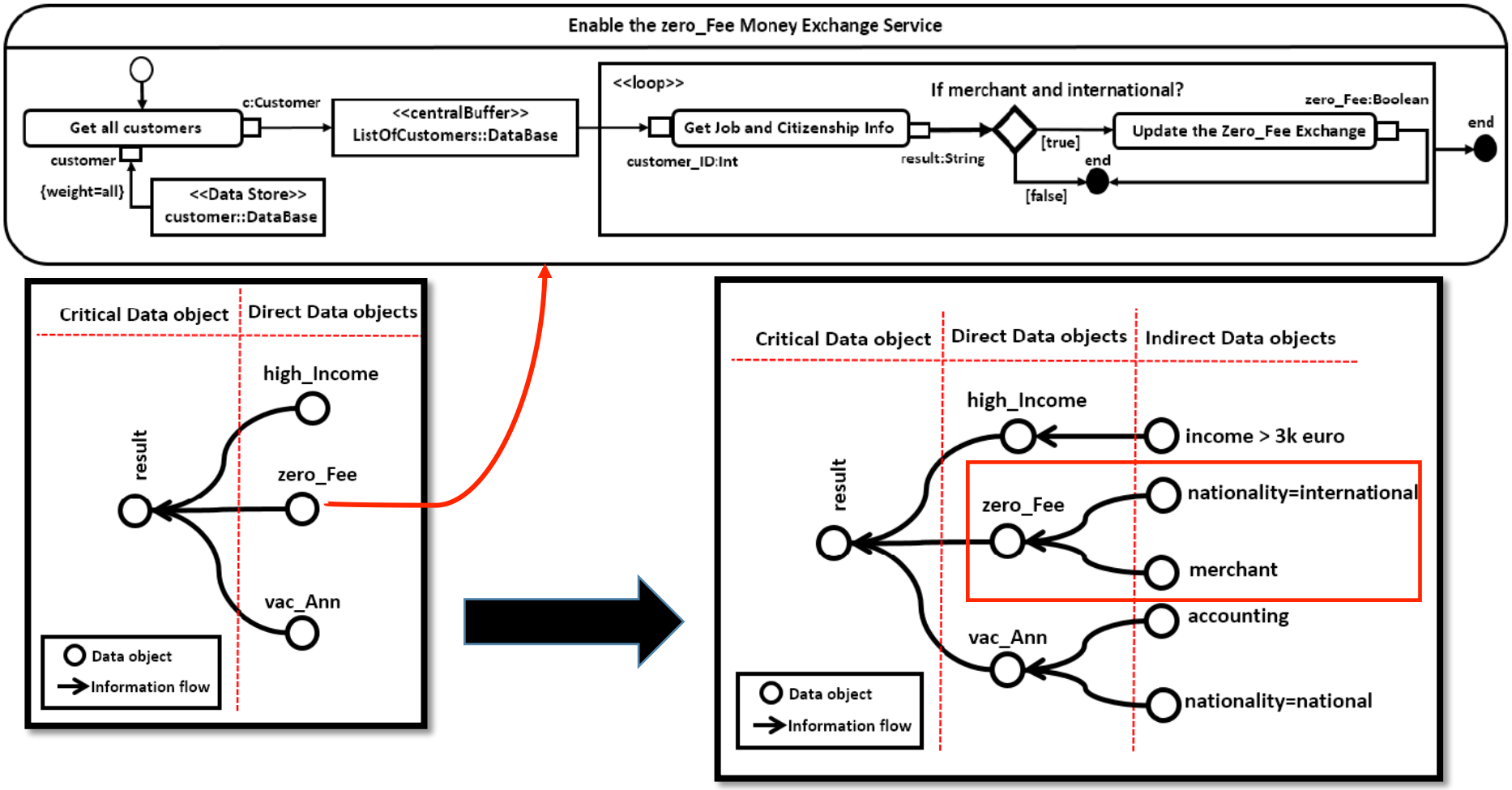


➢ But it does not tell how it is derived.

Database schema, specified using a class diagram

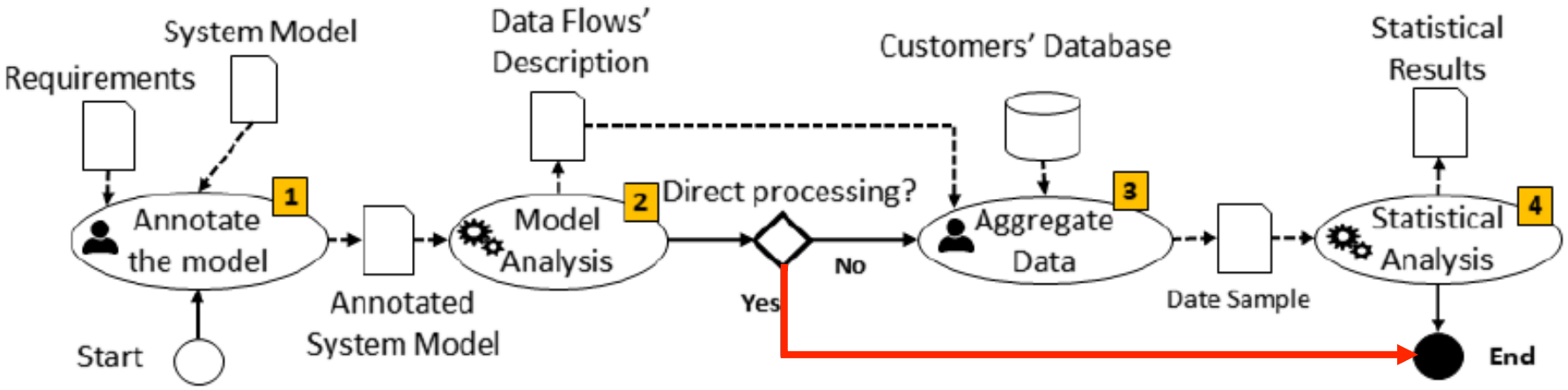
Information Flow Analysis

Activity diagram describing how the value of the *zero_Fee* data object is derived.



There is indirect leakage of the citizenship data to the result data object.

Roadmap: Model-based Analysis

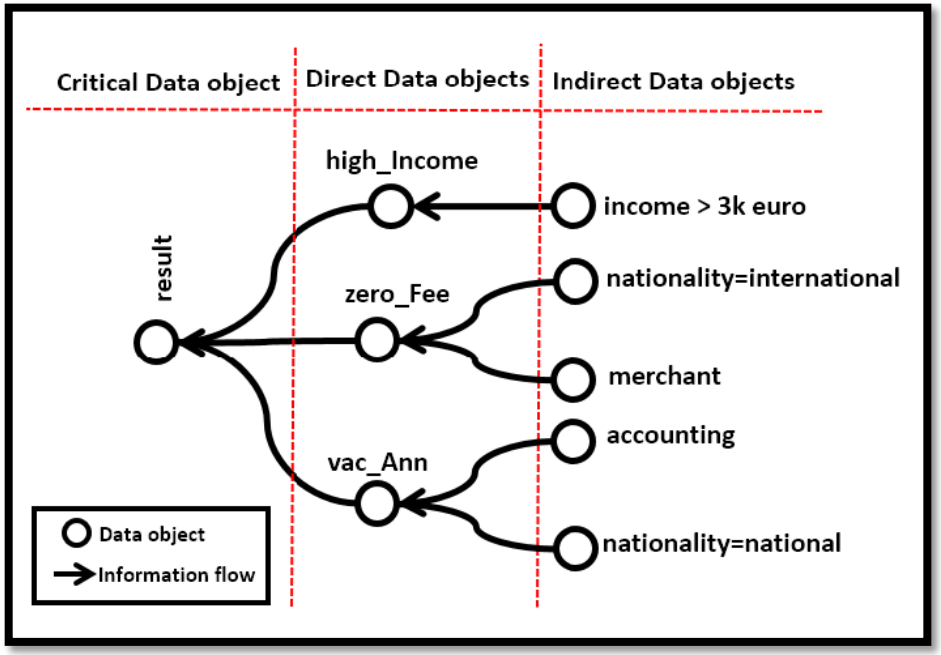


High-level overview of a model-based discrimination analysis framework.

Motivating Example 2

Business Analyst

- What about dependencies with the **gender**?

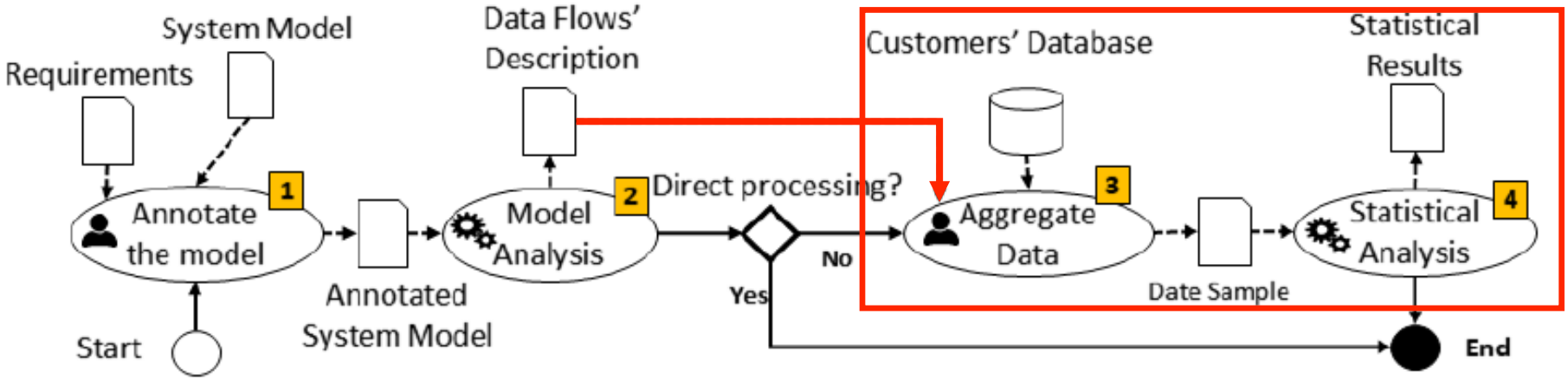


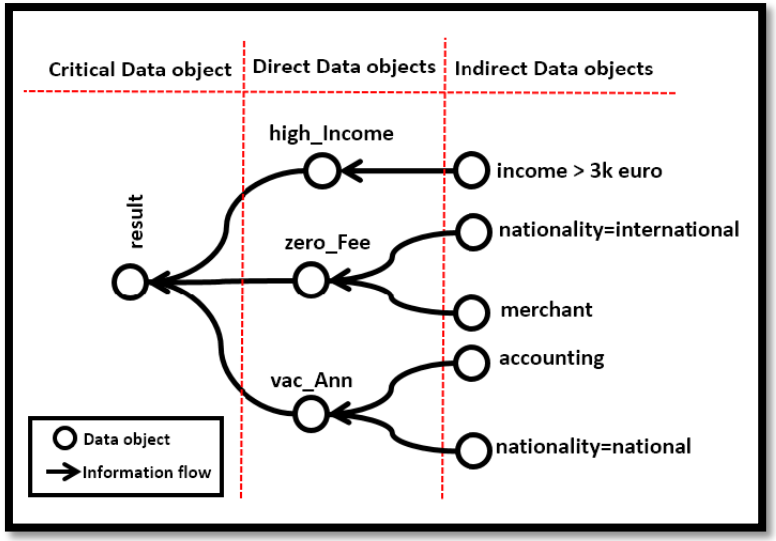
P(Female | ¬(">3k" €∩ international ∩ merchant) ∩ accounting) = 66.67%

(i.e., given a national customer with educational background in accounting and the other data objects are not true)

➢ The nationality and the education background in this context can act as a proxy for the gender.

Roadmap: Model-based Analysis



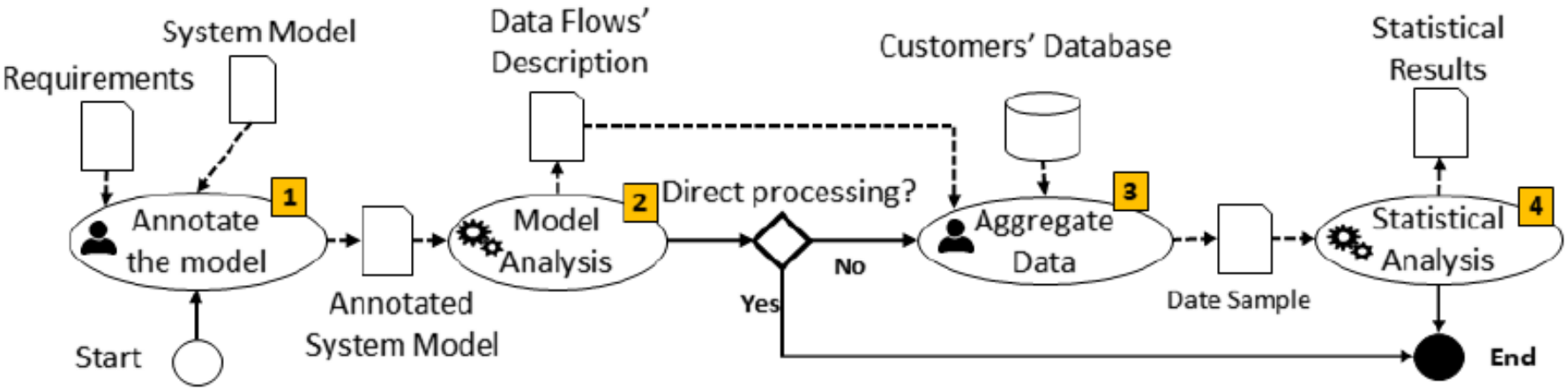


P(Female)	">3k" €	nationality	merchant	accounting
50.00%	1	1	1	1
66.67%	0	1	0	1
0.00%	0	0	1	0
100.00%	0	0	1	1
100.00%	1	1	0	1
0.00%	1	0	1	1

Conclusion

A Domain Expert and a Business Analyst

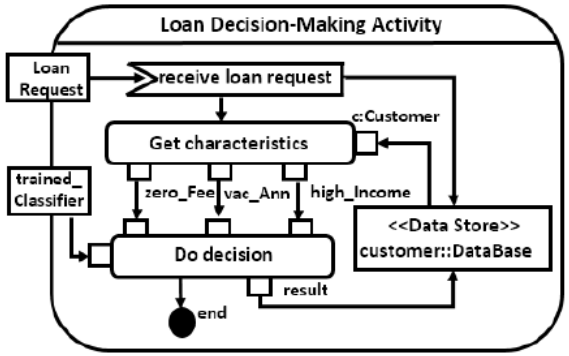
Business Analyst

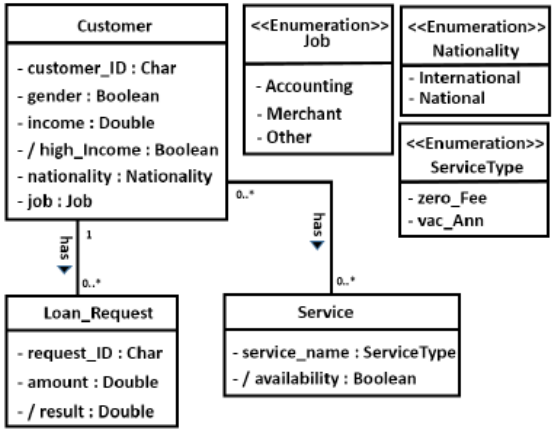


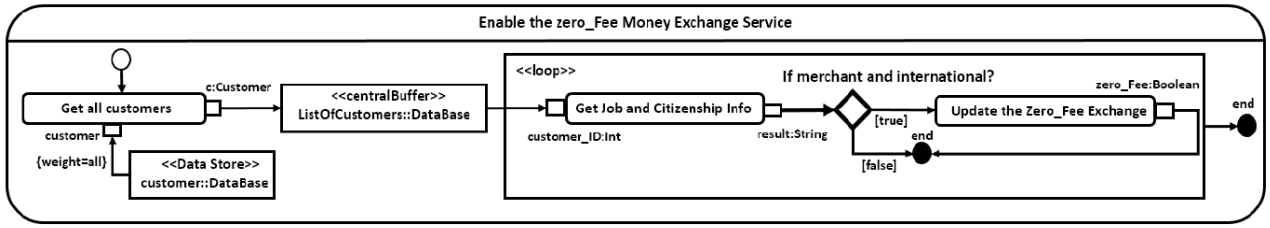
High-level overview of a model-based discrimination analysis framework.

Challenges

- Information about the derived data are distributed in multiples diagrams.

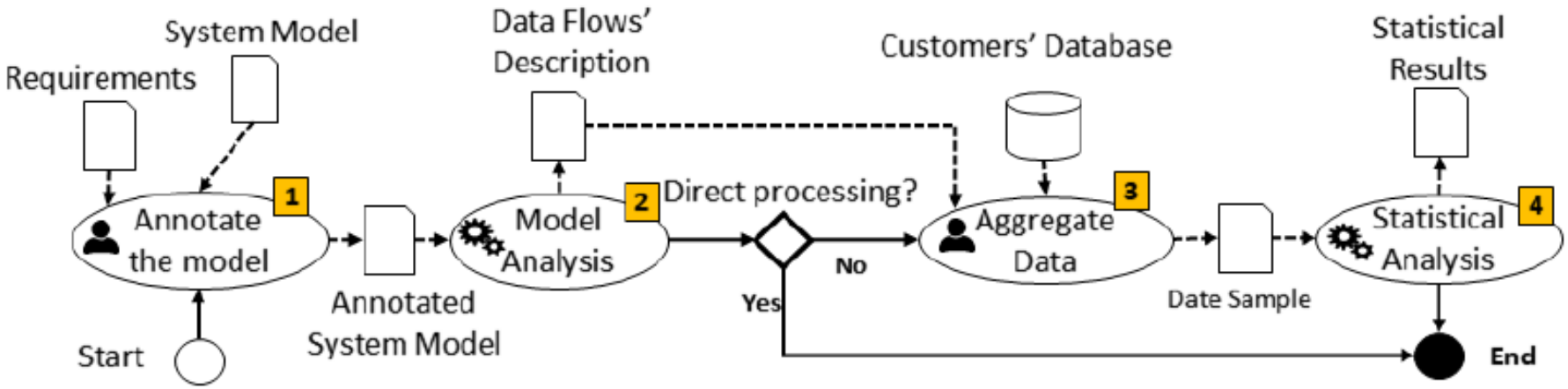






- How to measure the discrimination by proxy? (e.g., information gain)

Conclusion



High-level overview of a model-based discrimination analysis framework.

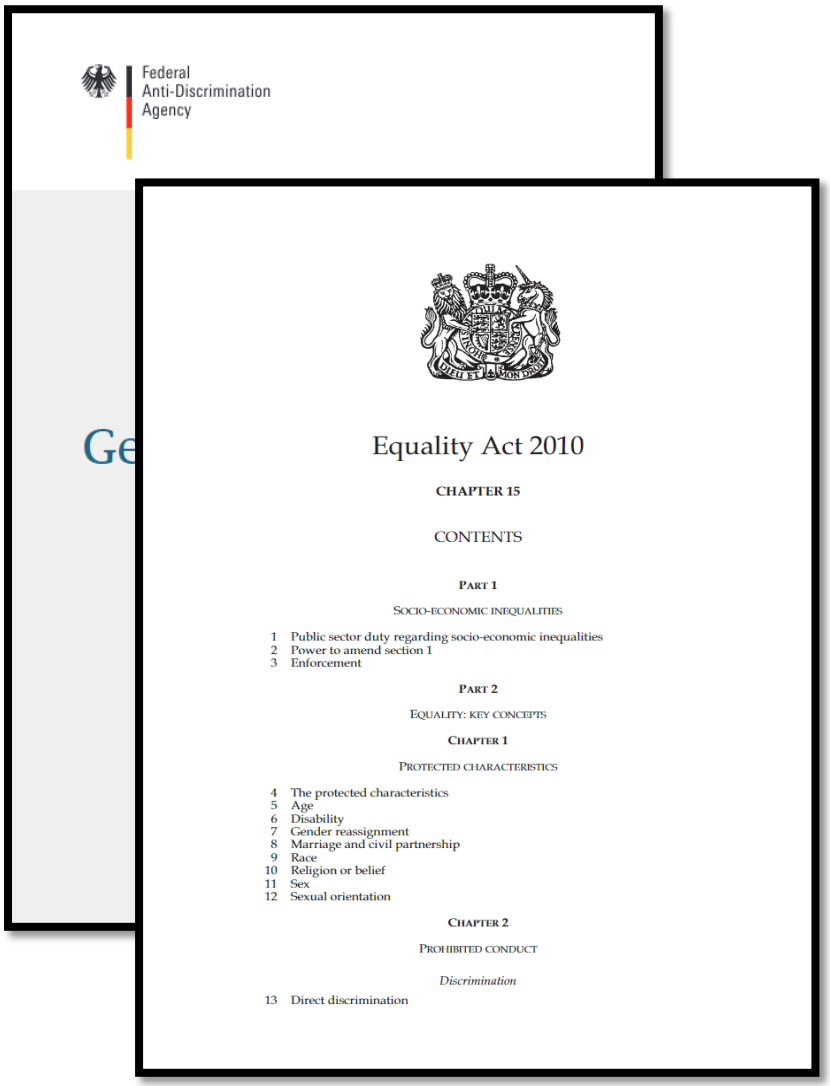
Backup slides

Protected Characteristics

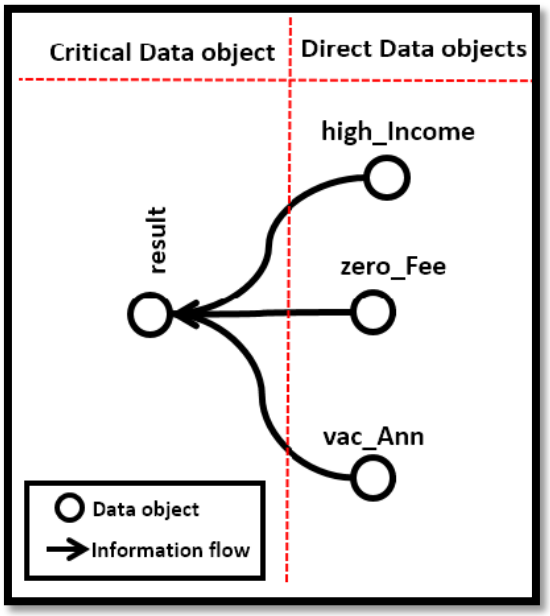
- legally protected characteristics (e.g., age, gender, race, …).

- But not limited to those listed by the laws and regulations.

  - **Example**: a bank may disallow discriminating between the loans applicants based on their citizenship.



Initial Statistical Analysis



P(international | zero_Fee ∩ vac_Ann ∩ high_Income)

P(international)	zero_Fee	vac_Ann	high_Income
66.67%	0	0	0
100.00%	1	0	0
0.00%	1	1	0
100.00%	1	0	1
0.00%	0	1	0
0.00%	0	1	1
100.00%	0	0	1

- a **societal fact** (e.g., a taxi driver in Saudi Arabia).
- They could be **derived** from processing the citizenship information.

Table 1: Personal Data

customer_ID	gender	income	high_Income	nationality	merchant	accounting
BA01	0	3000	0	0	1	1
BA02	0	4500	1	1	1	1
BA03	0	2500	0	1	0	1
BA04	0	3200	1	1	0	1
BA05	0	2900	0	1	0	1
BA06	1	5000	1	0	1	1
BA07	1	2450	0	0	1	0
BA08	1	3600	1	1	1	1
BA09	1	3100	1	0	1	1
BA10	1	1800	0	1	0	1

Table 2. Services Data

customer_ID	zero_Fee	vac_Ann
BA01	0	1
BA02	1	0
BA03	0	0
BA04	0	0
BA05	0	0
BA06	0	1
BA07	0	0
BA08	1	0
BA09	0	1
BA10	1	0