Explainable agents for less Bias in Human-Agent Decision Making

Avleen Malhi[1], Samanta Knapič[2], Kary Främling[1,2]

[1]Aalto University, Finland

[2]Umeå University, Sweden

Aalto University

UMEÅ UNIVERSITET

Agenda

Motivation

- Effect of explanations on human-agent decision making

Method

- Design non-explainable agents and explainable agents

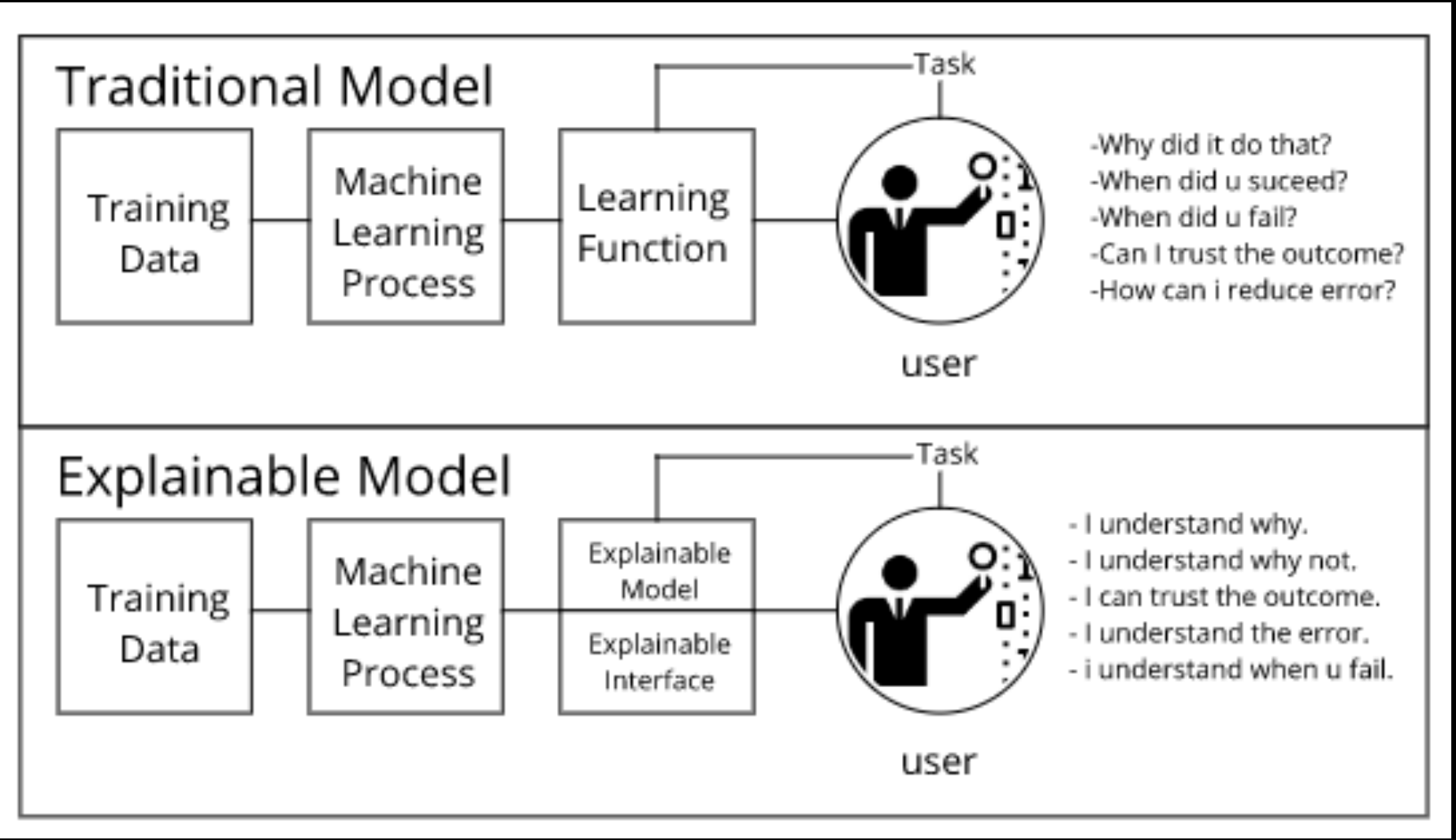
Empirical Assessment

- User study on 65 participants
- Quantitative analysis

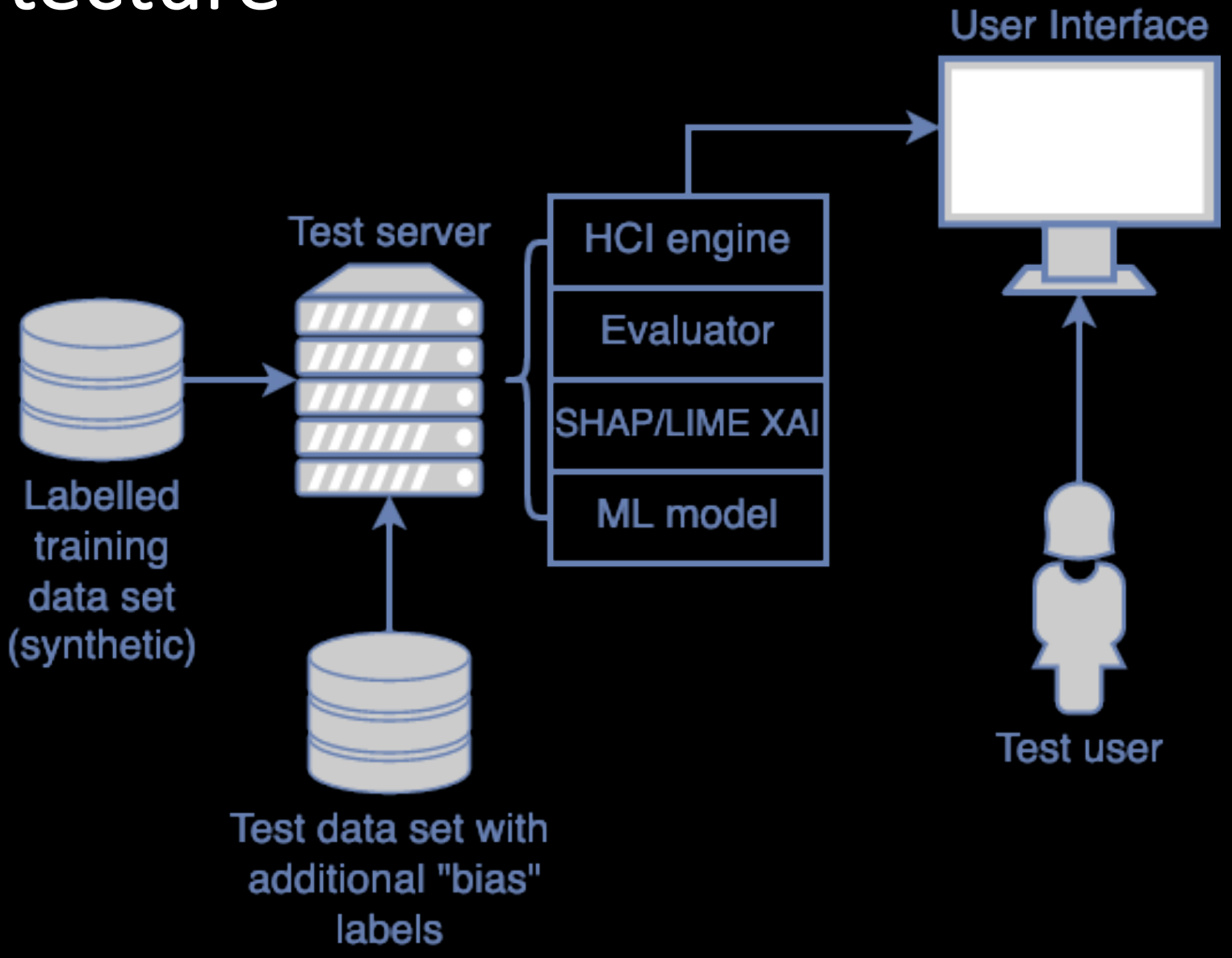
Human-agent decision making

- User understanding of AI system
- Understand machine learning decisions
- Satisfaction with model’s decisions
- User’s Trust on AI
- Human-agent system behaviour

Explainable Artificial Intelligence



Architecture



Human-agent interaction Method

Case Study: Loan Application

- Generate Test Data
- Add Bias
- Training model- Random Forest

Explanation types

- No explanation
- Explanation I - LIME
- Explanation II - SHAP

Data Generation

Decision Rules

If age<18: reject
Else if income<20000 and loan>10000: reject
Else if assets<20000 and loan>10000: reject
Else if assets<0: reject
Else if employment != permanent and loan>400000: reject
Else if loan>500000: reject
Else: accept

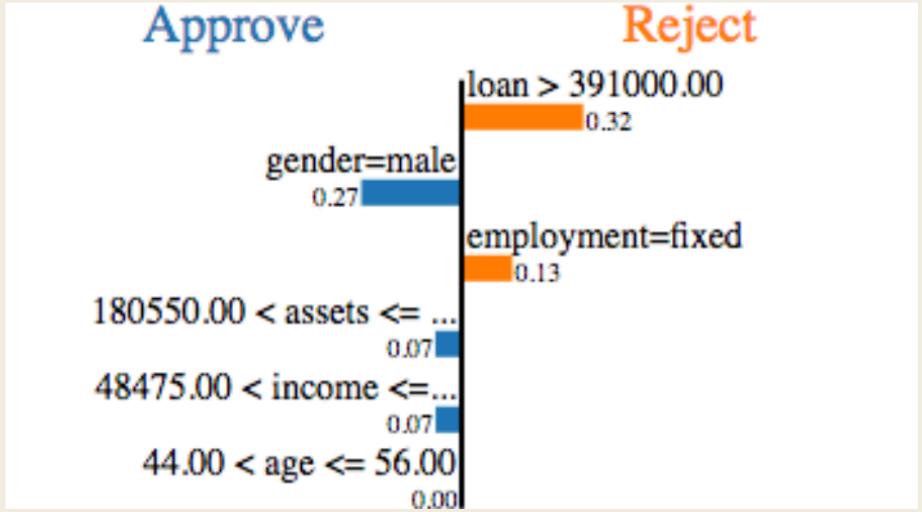
Bias Added

If Gender == ’female’ or ’other’: reject with probability of 80%
Else: call the first set of rules

Dataset variables

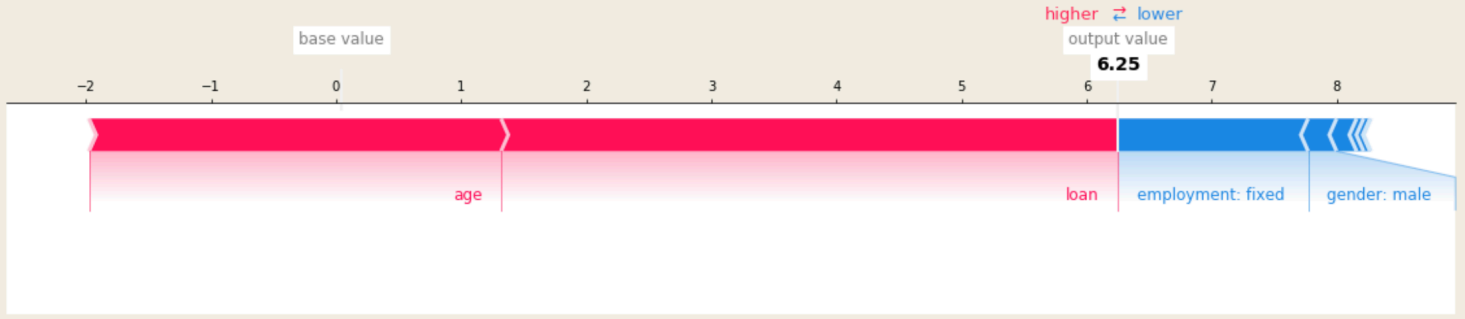
Age	17 - 70 years;
Income	0 - 200000€;
Assets	100000 - 1000000€;
Employment type	Fixed-term or permanent;
Gender	female, male or other;
Loan amount	5000 - 520000€

Explanation Methods



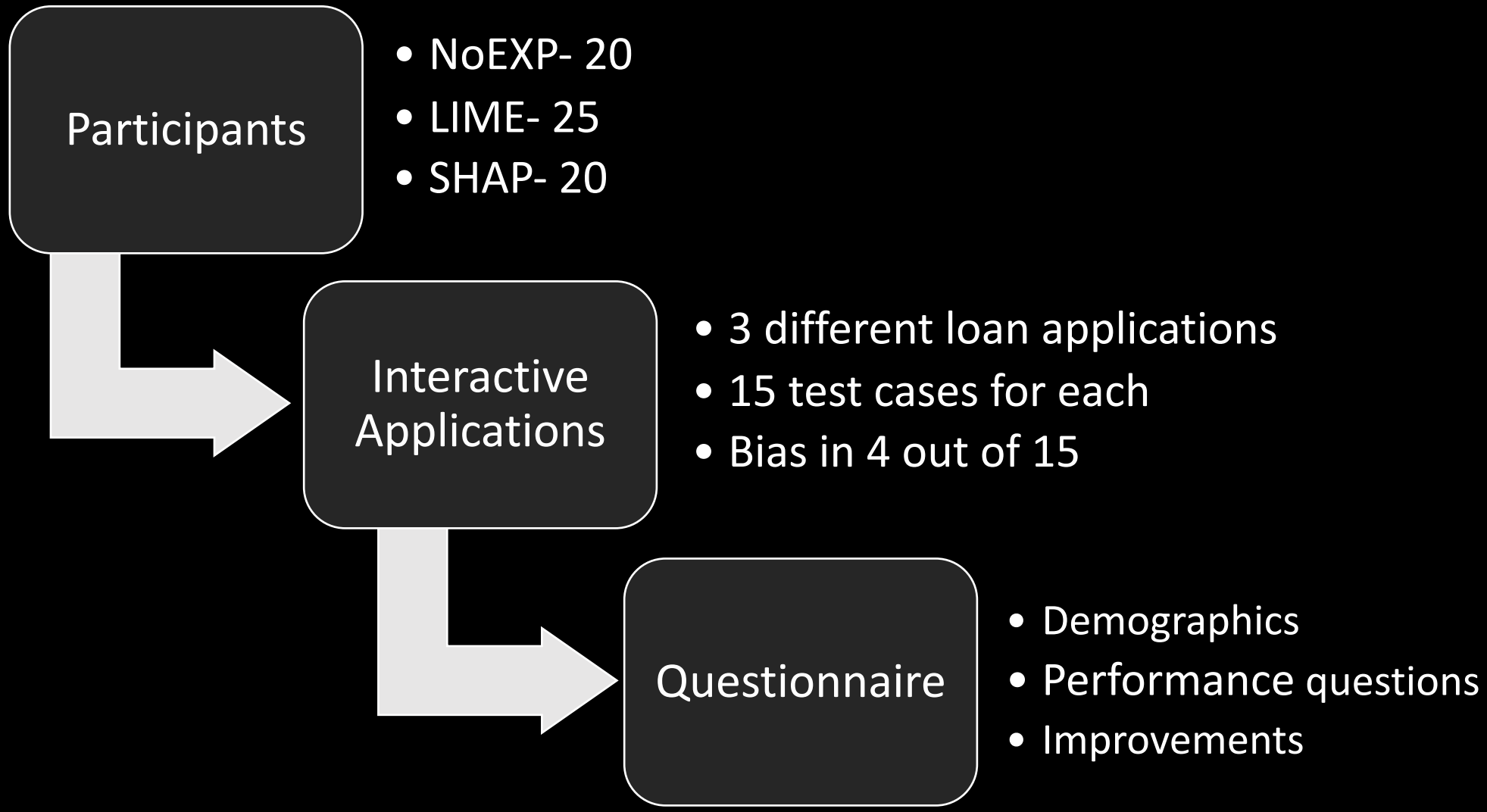
Recommendation: Reject

Explanation I: LIME



Explanation II: SHAP

Study Description



Loan Application

Loan Application Approval

Case Data

Income	46800
Gender	male
Employment	permanent
Loan	86000
Assets	922000
Age	25

Recommendation: Approve

Approve | Reject

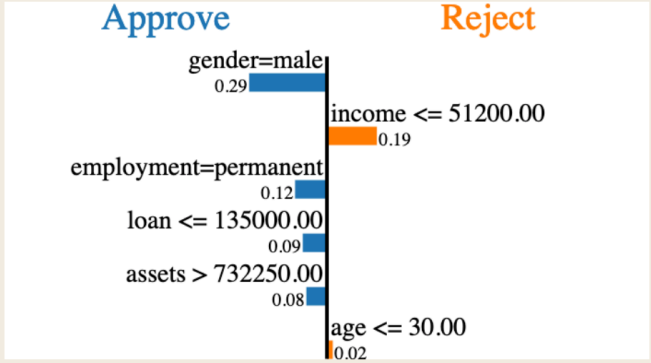
Application I: noEXP

Loan Application Approval

Case Data

Income	46800
Gender	male
Employment	permanent
Loan	86000
Assets	922000
Age	25

Explanation

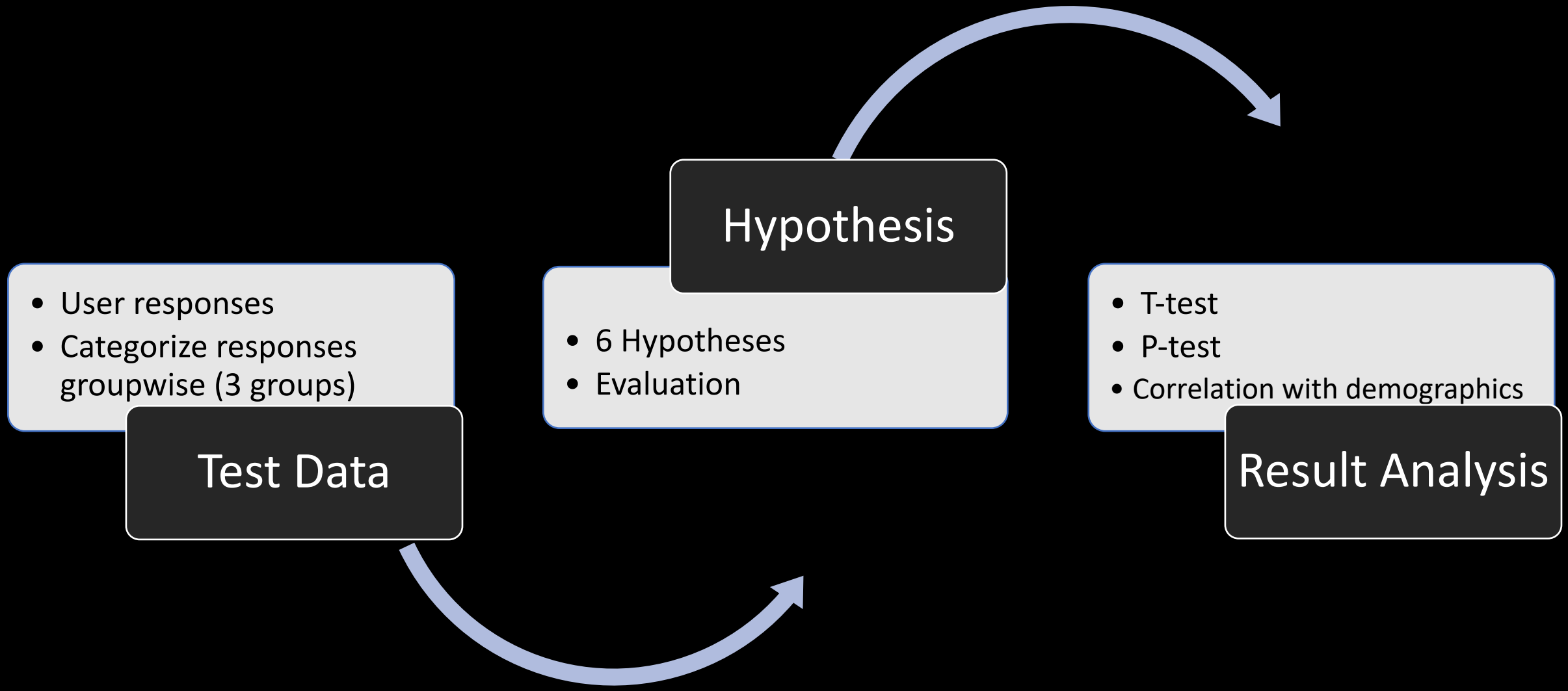


Recommendation: Approve

Approve | Reject

Application II: LIME

Analysis Method



Study Participants

Methods	Total	Gender			Highest Degree				STEM Background		Age (years)
		Male	Female	OTH	Ph.D. (or higher)	Master	Bachelor	High school	Yes	No	
noEXP	**20**	10	9	0	1	5	7	6	13	7	21 (2), 23, 24 (2), 26(2), 27(2), 28(3), 30(3), 31(2), 34, 50, 57
LIME	**25**	18	5	1	4	12	6	2	24	1	20, 24(3), 25(2), 28, 29(4), 30(4), 32(3), 33(2), 37(2), 38, 51, 53
SHAP	**20**	11	7	1	7	9	3	1	18	2	21, 23, 24, 25(2), 26, 27(3), 28, 29, 32, 33(2), 34, 35, 36(2), 38, 41

Hypotheses

Number of ”overridden” recommendations that are:

Ha: biased SHAP > noEXP

Hb: not biased SHAP < noEXP

Hc: biased LIME > noEXP

Hd: not biased LIME < noEXP

He: biased LIME > SHAP

Hf: not biased LIME < SHAP

Result Analysis

		No XAI	LIME	SHAP
Overrides biased Recommendations	TP	48	63	57
Overrides non-biased Recomm.	FP	81	87	61
Supports non-biased Recomm.	TN	139	188	159
Supports biased Recomm.	FN	32	37	23
	Σ	300	375	300

Hypotheses analysis

	Hypothesis	P-value (2-tailed)	P-value (1-tailed)
TP (SHAP, noEXP)	Ha	0.18	0.09
FP (SHAP, noEXP)	Hb	0.13	0.06
TP (LIME, noEXP)	Hc	0.71	0.35
FP (LIME, noEXP)	Hd	0.35	0.17
TP (LIME, SHAP)	He	0.36	0.18
FP (LIME, SHAP)	Hf	0.39	0.19

Discussion

1. The study supports Ha, Hb, … , He
2. Significant results to support Hb and Hc as overriding of
   a) **bias** recommendations **LIME > noEXP**.
   b) **non-biased** recommendations **SHAP < noEXP**.
3. Not significant but notable results for Ha, Hd and He
4. **Hf** fails as non-biased overridden recommendations LIME >= SHAP

Conclusion

1. **Users prefer explanations** compared with noEXP
2. **Explanations** helps in **detecting bias** better than noEXP because:
   a) less overriding of non biased recommendations
   b) more overriding of biased recommendations
3. Results **can not** be **generalized** because of **small sample** size
4. **Income** is considered important by users.
5. More than **50% users are pro for** explanations in decision making.

Future work

To scale the study to other XAI tools

Evaluate the study applicability with domain experts

Extend the scope to real-life case study

Thank You
For Your Attention

Please email to avleen.malhi@aalto.fi