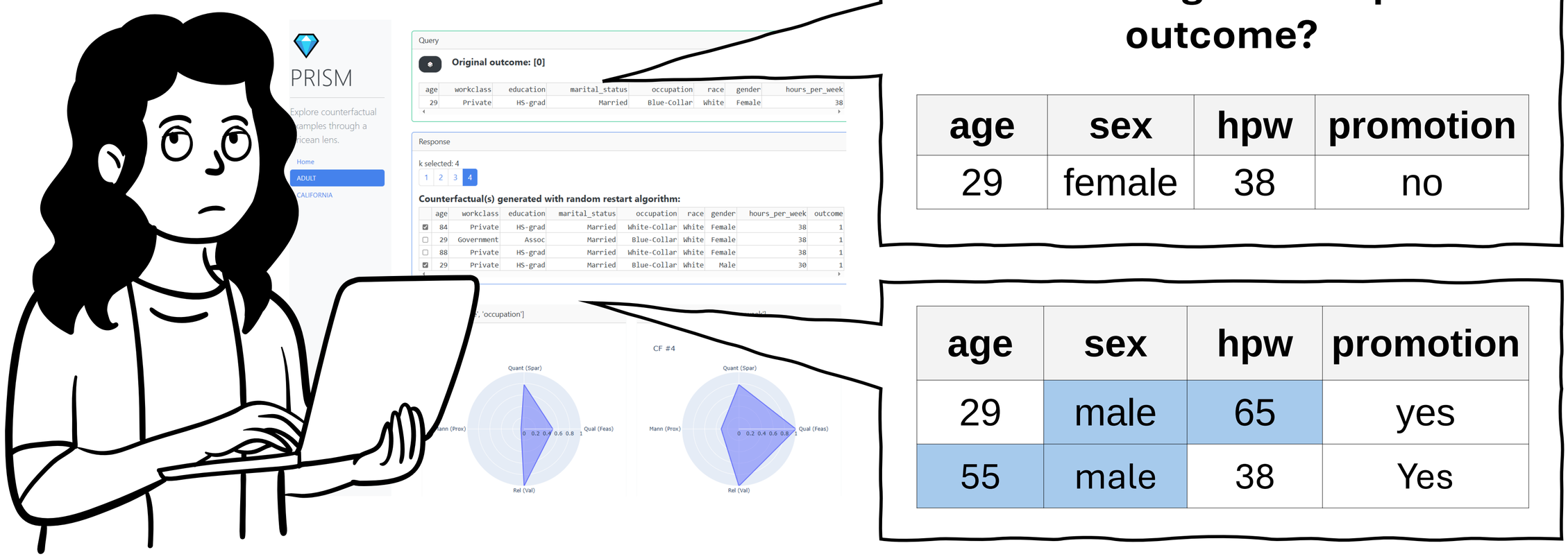
PRISM

A Pragmatic Framework for Evaluating Counterfactual Explanations

Leila Methnani, Virginia Dignum, Andreas Theodorou
leila.methnani@umu.se, virginia@cs.umu.se, andreas.theodorou@upc.edu

Jane Doe is a **29-year-old female** employee who works **38 hours per week (hpw)**. This year, she was **not up for a promotion**.



Are these **poor counterfactual (CF) examples** because they suggest inactionable feature changes? Or do they **create implicatures that reveal model vulnerabilities** we should investigate further?

Though inactionable, the examples offered to Jane hold meaning – they reveal potential bias towards male employees.

Grice's conversational model [1] suggests that implicatures are made when people blatantly violate any of four maxims.

Metric	Quantity	Quality	Relation	Manner
Validity			✓	
Proximity				✓
Sparsity	✓			
Feasibility		✓		
Actionability		✓		
Diversity	✓			
Efficiency				✓
Stability				✓

In this work, we:

1. mapped CF metrics of evaluation to Grice's four maxims of conversation;
2. built a framework called **P**ragmatics, **I**nferences, and **S**ubtext analysis through **M**axims (**PRISM)**; and
3. demonstrated with an interactive dashboard.

Read more here 👉

UMEÅ UNIVERSITY
UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH
AI Policy Lab at Umeå University
EXPLAIN