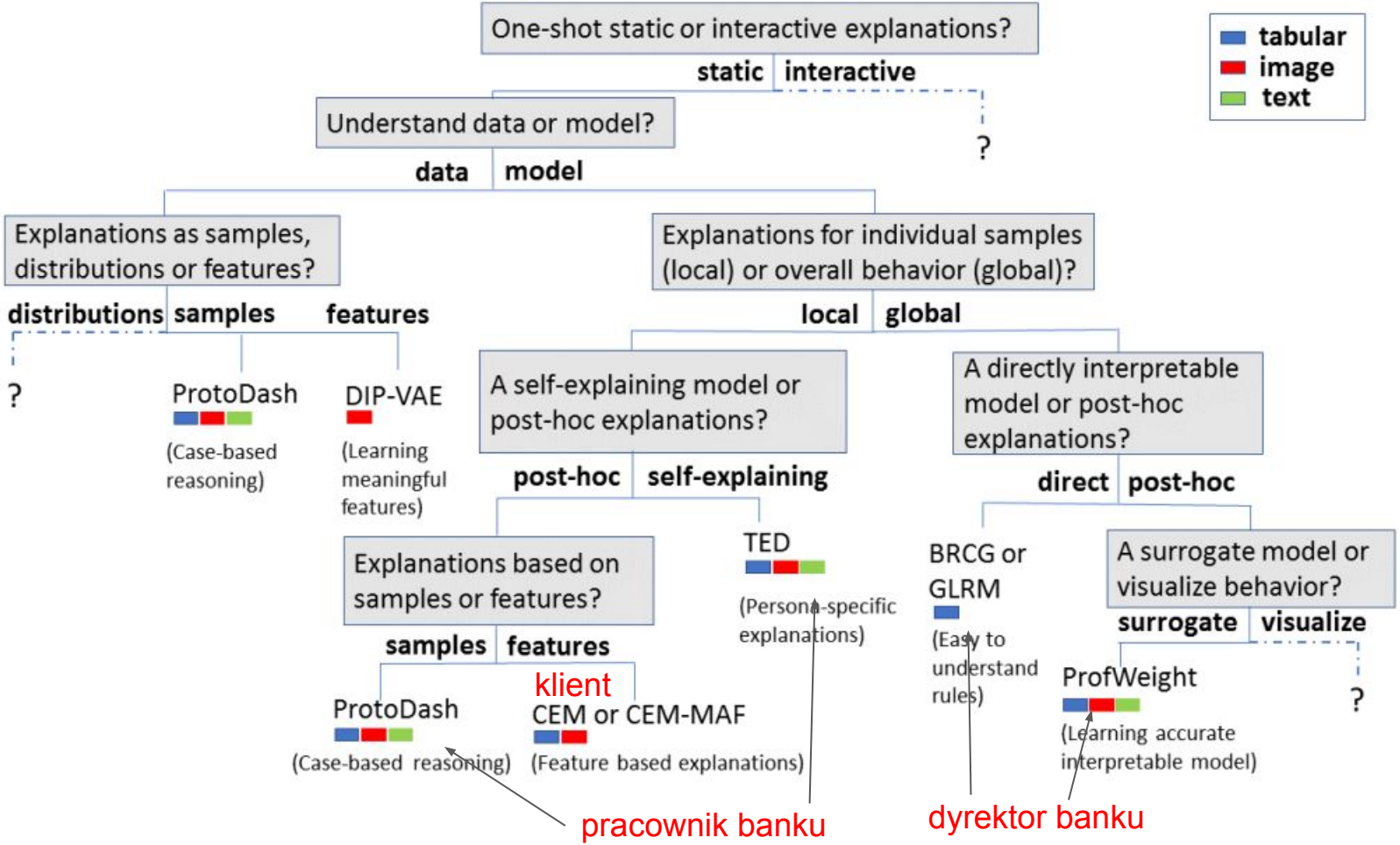
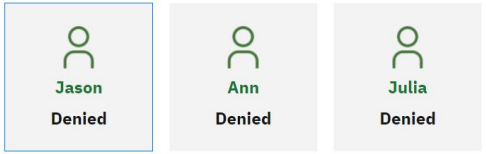
One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques


One-shot static or interactive explanations?
static
interactive
?
tabular
image
text
Understand data or model?
data
model
Explanations as samples, distributions or features?
distributions
samples
features
?
ProtoDash
(Case-based reasoning)
DIP-VAE
(Learning meaningful features)
Explanations for individual samples (local) or overall behavior (global)?
local
global
A self-explaining model or post-hoc explanations?
post-hoc
self-explaining
Explanations based on samples or features?
samples
features
ProtoDash
(Case-based reasoning)
CEM or CEM-MAF
(Feature based explanations)
TED
(Persona-specific explanations)
A directly interpretable model or post-hoc explanations?
direct
post-hoc
BRCG or GLRM
(Easy to understand rules)
A surrogate model or visualize behavior?
surrogate
visualize
ProfWeight
(Learning accurate interpretable model)
?
klient
pracownik banku
dyrektor banku

CEM

- SHAP, LIME
-



Several features in Jason's application fall outside the acceptable range. All would need to improve before acceptance was recommended.

Factors contributing to Jason's application denial

1. The value of **Consolidated risk markers** is **65**. It needs to be around **72** for the application to be approved.
2. The value of **Average age of accounts in months** is **52**. It needs to be around **68** for the application to be approved.
3. The value of **Months since most recent credit inquiry not within the last 7 days** is **2**. It needs to be around **3** for the application to be approved.

$$\theta_j^{pp} = 1 - \exp \frac{-\left|x_{[j]}^{pp}\right|}{\sigma_j}$$

$$\theta_j^{pn} = 1 - \exp \frac{-\left|x_{[j]} - x_{[j]}^{pn}\right|}{\sigma_j}$$

ProtoDash i TED (Teaching Explanations for Decisions)

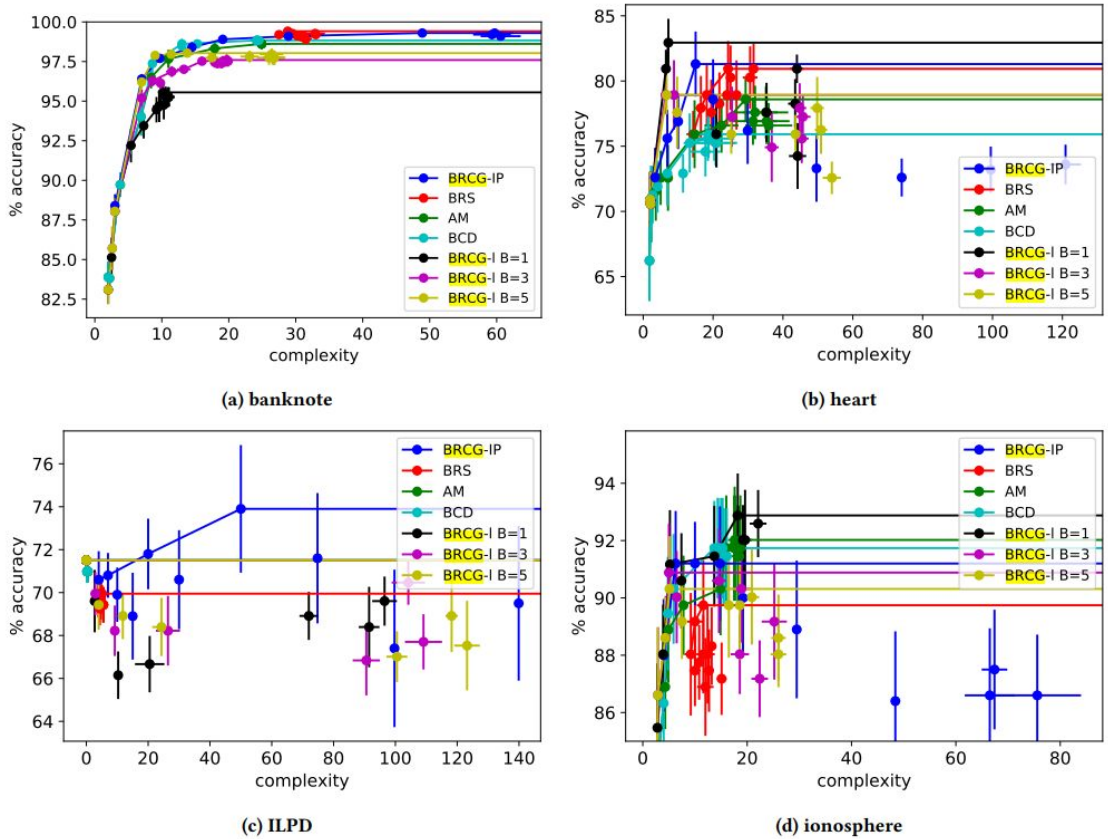
	Robert	James	Danielle	Franklin
Outcome	-	Defaulted	Defaulted	Defaulted
Similarity to Robert (from 0 to 1)	-	0.690	0.114	0.108
ExternalRiskEstimate	78	71	72	69
MSinceOldestTradeOpen	82	95	166	193
MSinceMostRecentTradeOpen	5	1	12	12
AverageMInFile	54	43	74	167
NumSatisfactoryTrades	33	33	37	36
NumTrades60Ever2DerogPubRec	0	0	1	0
NumTrades90Ever2DerogPubRec	0	0	1	0
PercentTradesNeverDelq	100	100	95	100
MSinceMostRecentDelq	0	0	7	0
MaxDelq2PublicRecLast12M	7	7	4	7
MaxDelqEver	8	8	4	8
NumTotalTrades	41	41	41	8
NumTradesOpeninLast12M	2	4	0	0
PercentInstallTrades	15	17	15	6
MSinceMostRecentInqexcl7days	0	0	0	0
NumInqLast6M	3	4	1	0
NumInqLast6				
NetFractionRe				

$$\theta_j = \exp \frac{- \left| x_{[j]} - x'_{[j]} \right|}{\sigma_j}$$

TED:

- tworzenie reguł wyjaśnienia przez specjalistów,
- model zwraca predykcję i wyjaśnienie swojej predykcji.

ProfWeight i BRCG



Co nam daje GLRM? (Generalized Linear Rule Models)

Odpowiada na pytania:

- jaka jest ogólna logika modelu przy podejmowaniu decyzji?
- czy ta logika jest rozsądna abyśmy z pewnością mogli ją wdrożyć?

AIX360 - metryki do oceniania wyjaśnień

Faithfulness

Monotonicity

$$\phi = -\rho(\theta, \mathbf{p})$$