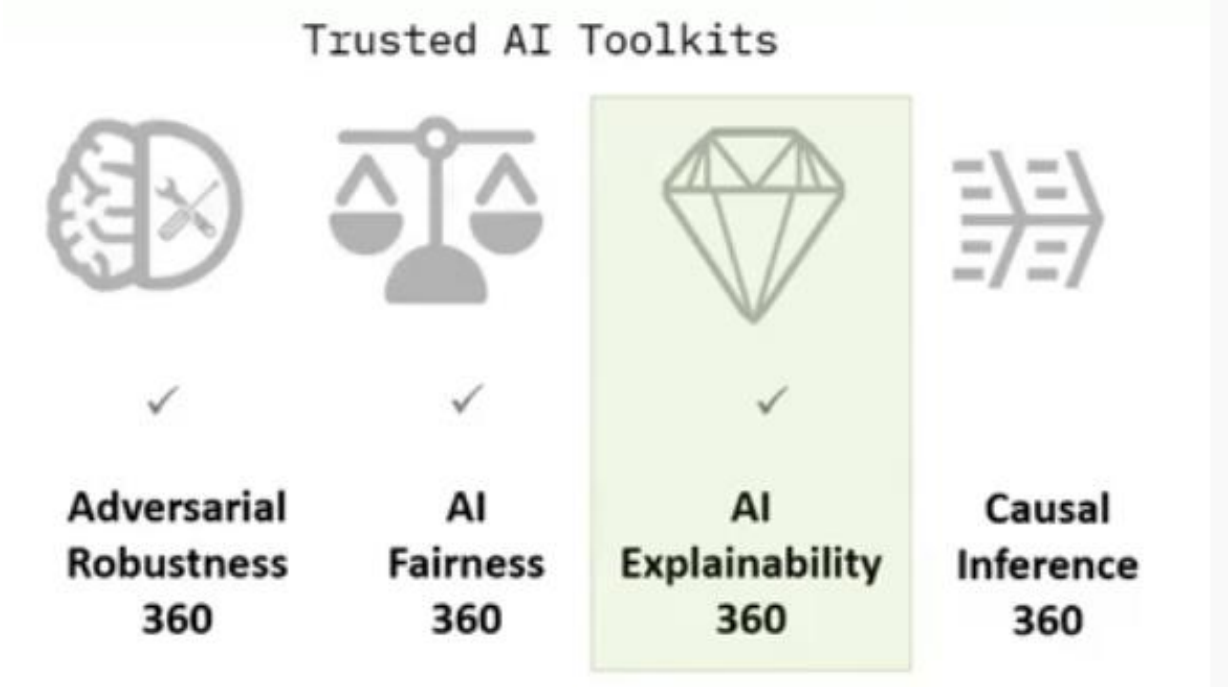
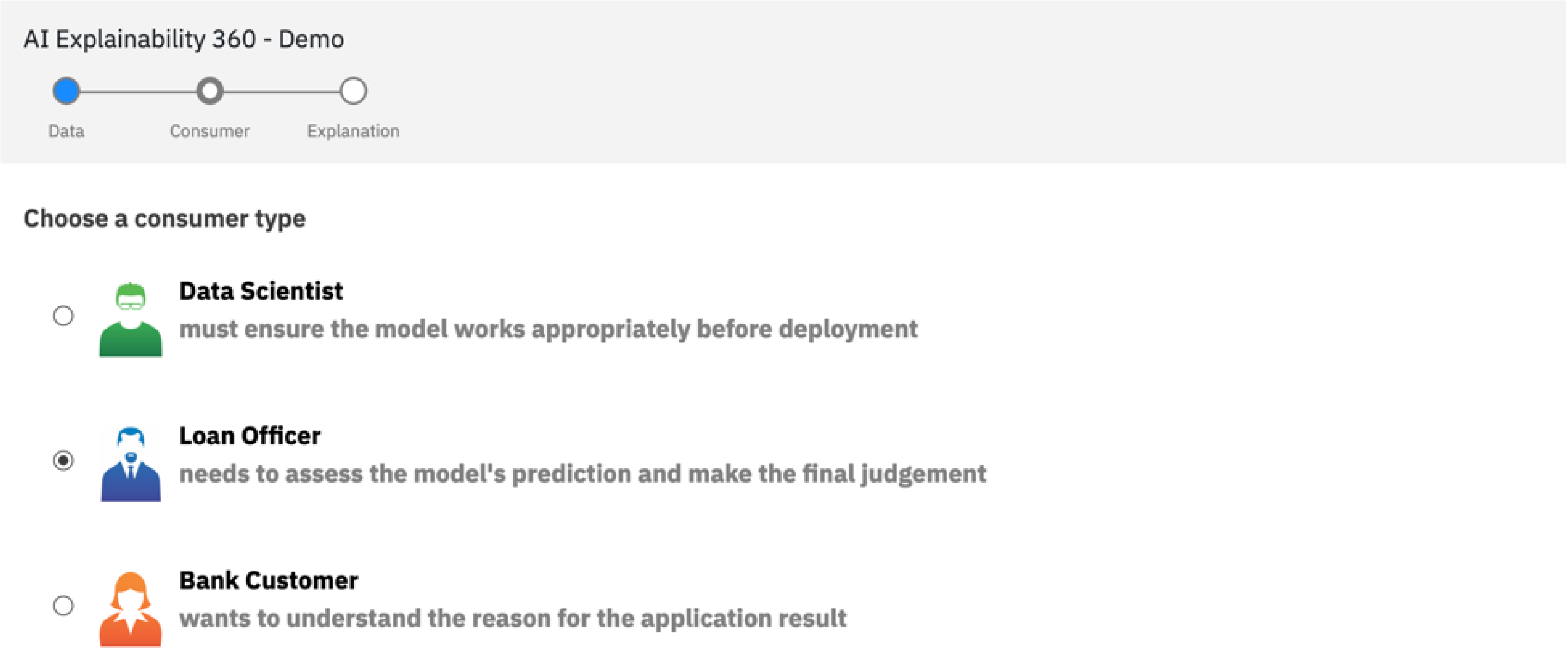


AIX360 by IBM Research

Michał Kuźba

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

Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind
Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović
Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri
Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, Yunfeng Zhang
IBM Research

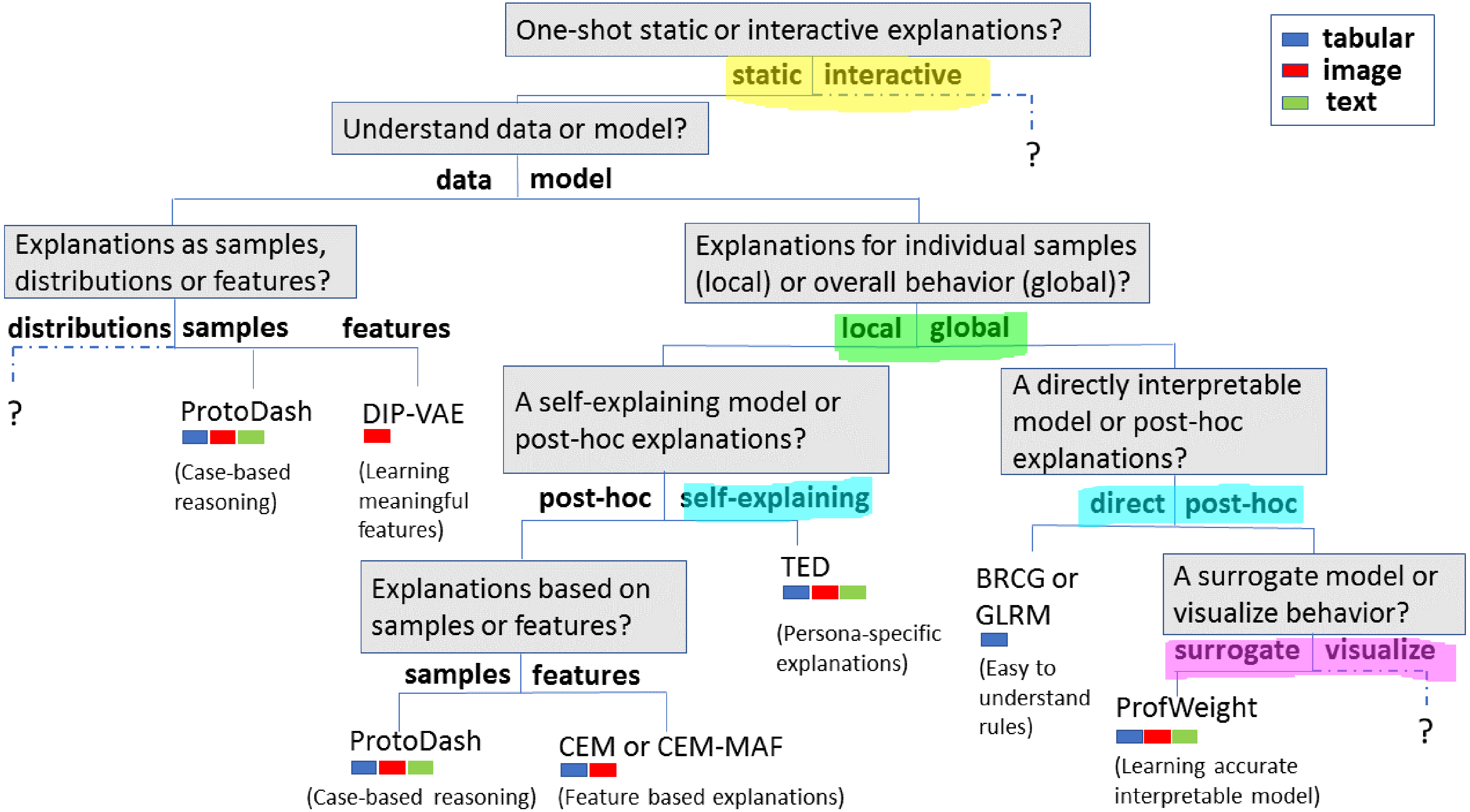


Co zrobili?

- AI Explainability 360 - toolkit
- XAI taxonomy
- Interaktywne demo
- Algorytmiczne ulepszenia do istniejących metod
- Implementacja metryk wyjaśnialności
- Tutoriale

Słownictwo

- explainability = interpretability (tutaj)
- people interacting with AI system – consumers
- types of consuments – personas

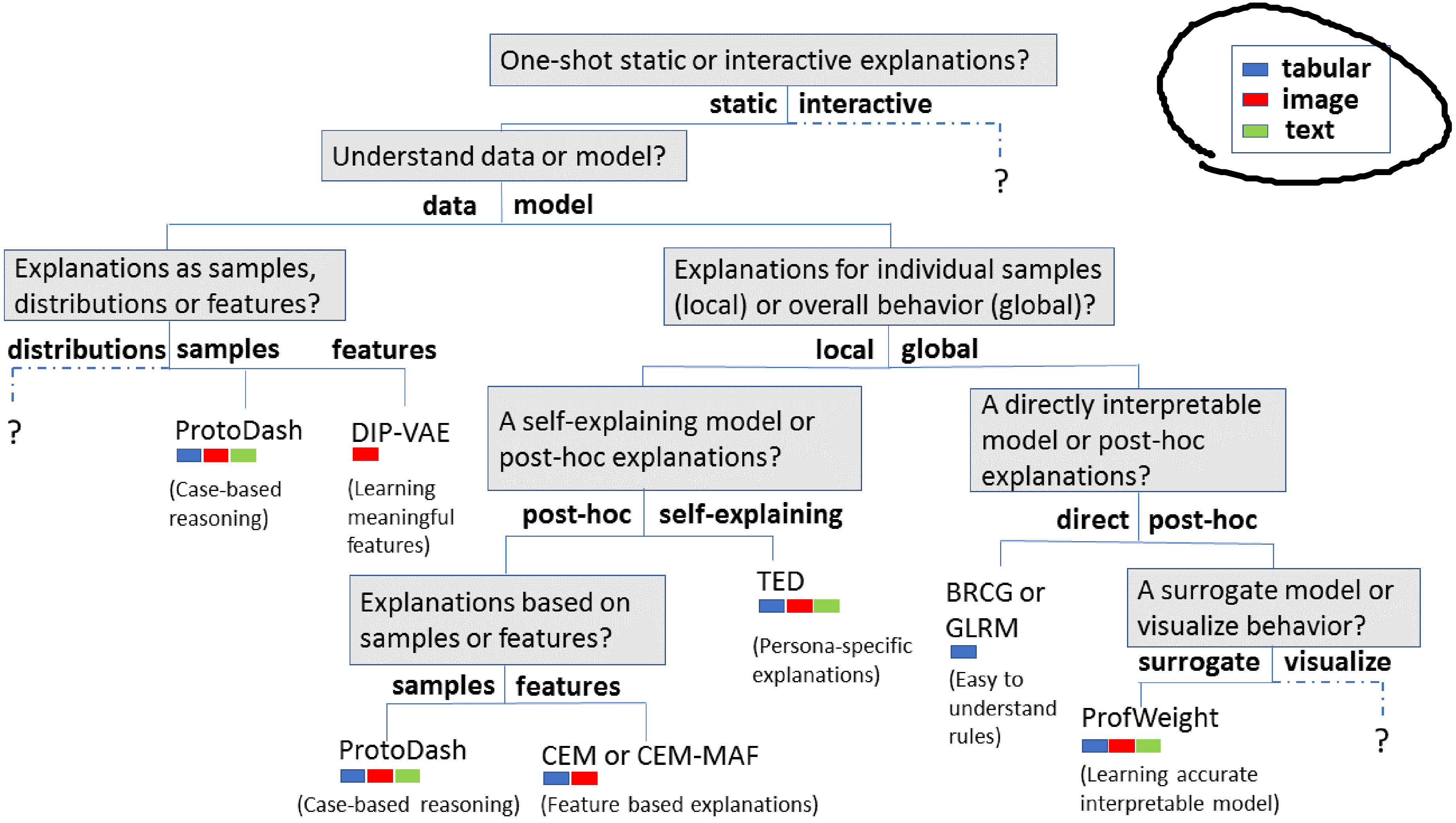


- Statyczne – nie zmienia się względem feedbacku od konsumenta
- interaktywne – pozwala na dalszą eksplorację ("drążenie") lub pytanie (dialog)

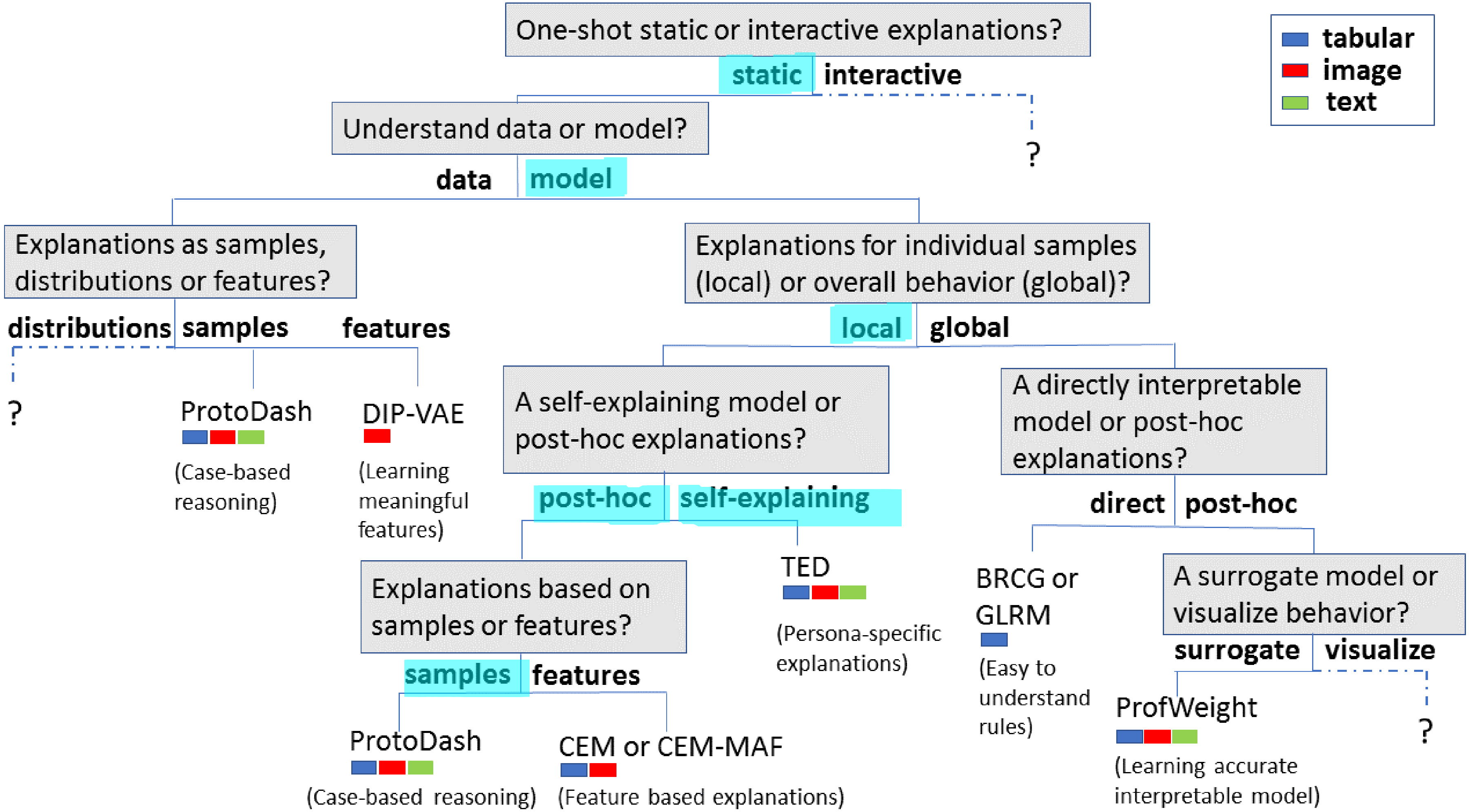
- Lokalne vs globalne (pojedyncza predykcja vs model)

- Bezpośrednio interpretowalne vs metody post-hoc vs samowyjaśnialny (generujący wyjaśnienia np. tekstowe)

- Surrogate model (interpretowalny). Wizualizacja (części modelu), nie jest modelem.



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)
?

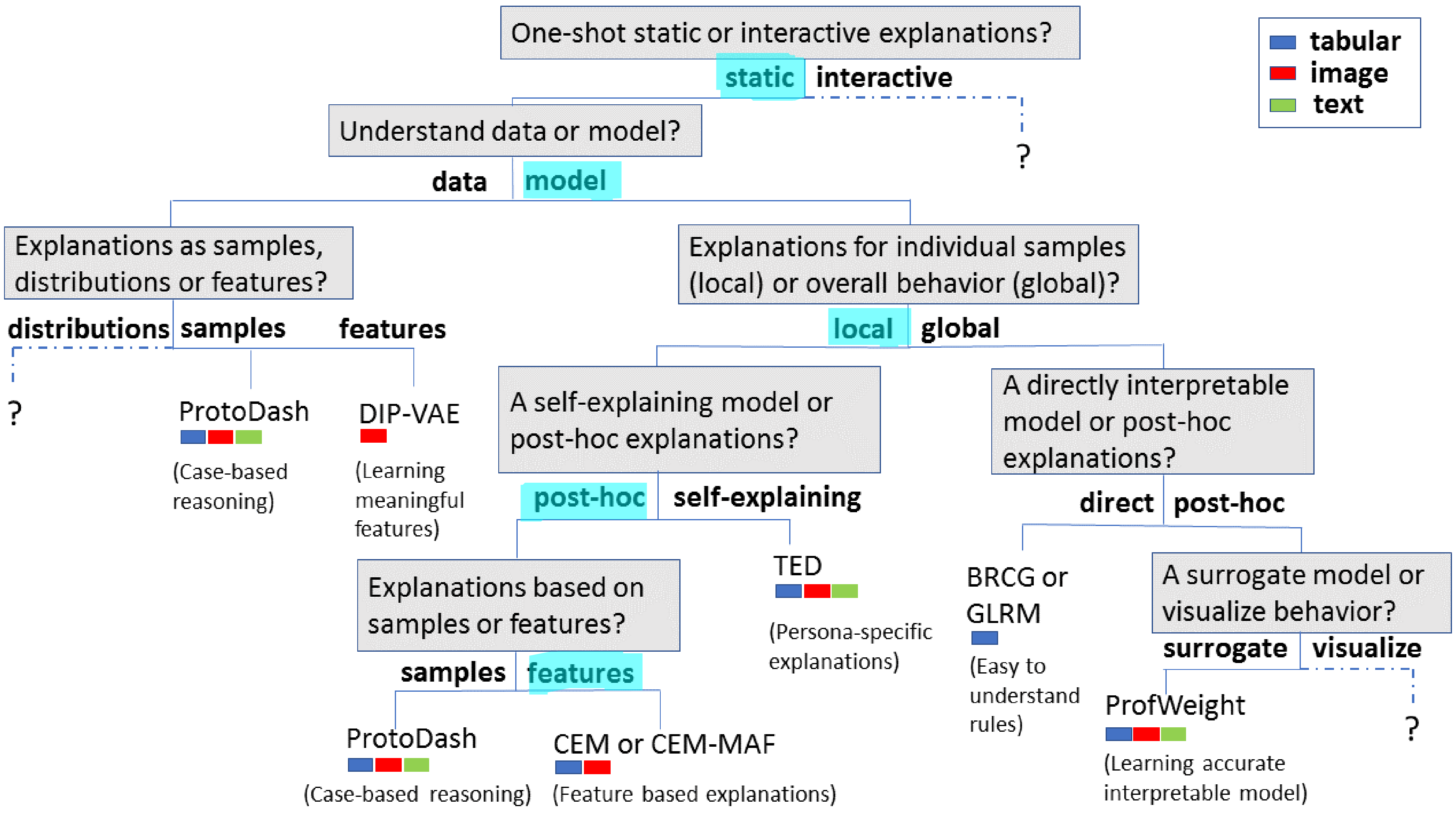


Usecase – wniosek kredytowy

Pracownik banku

Walidacja czy oceny są uzasadnione na podstawie porównania z podobnymi obserwacjami.

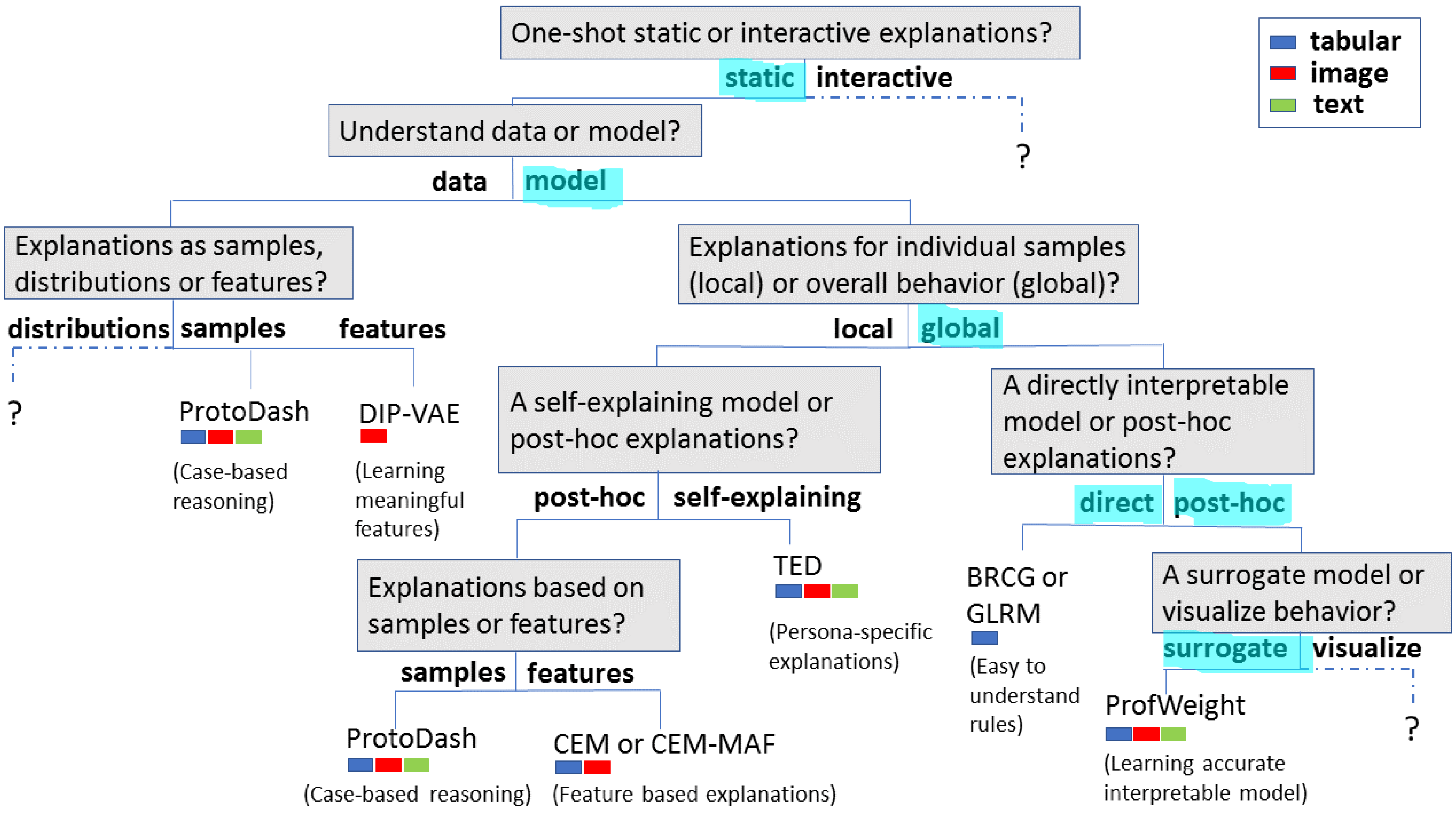
Podobnie lekarz?



Usecase – wniosek kredytowy

Klient

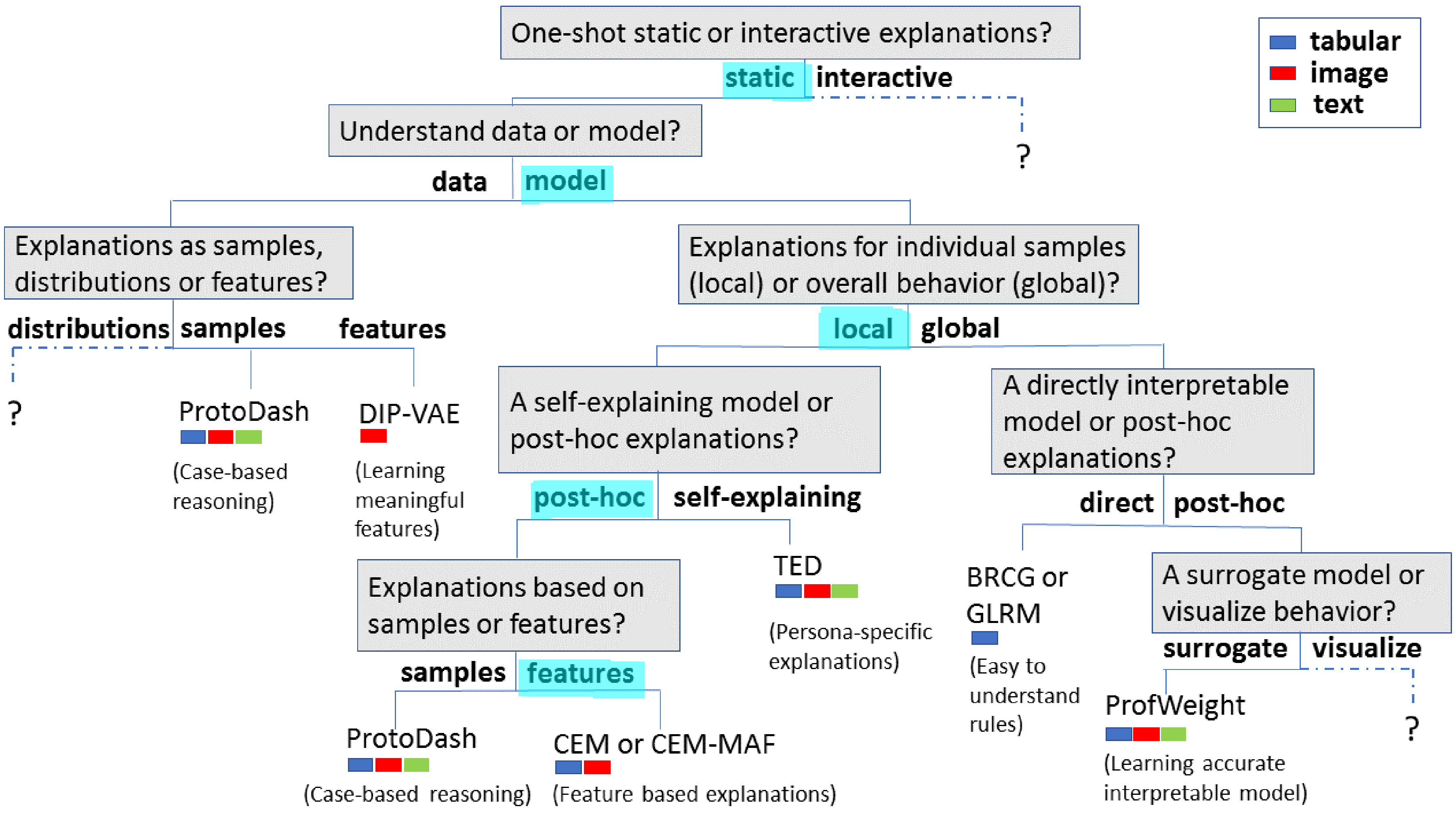
"Jak poprawić swoje szanse?" - zmienne



Usecase – wniosek kredytowy

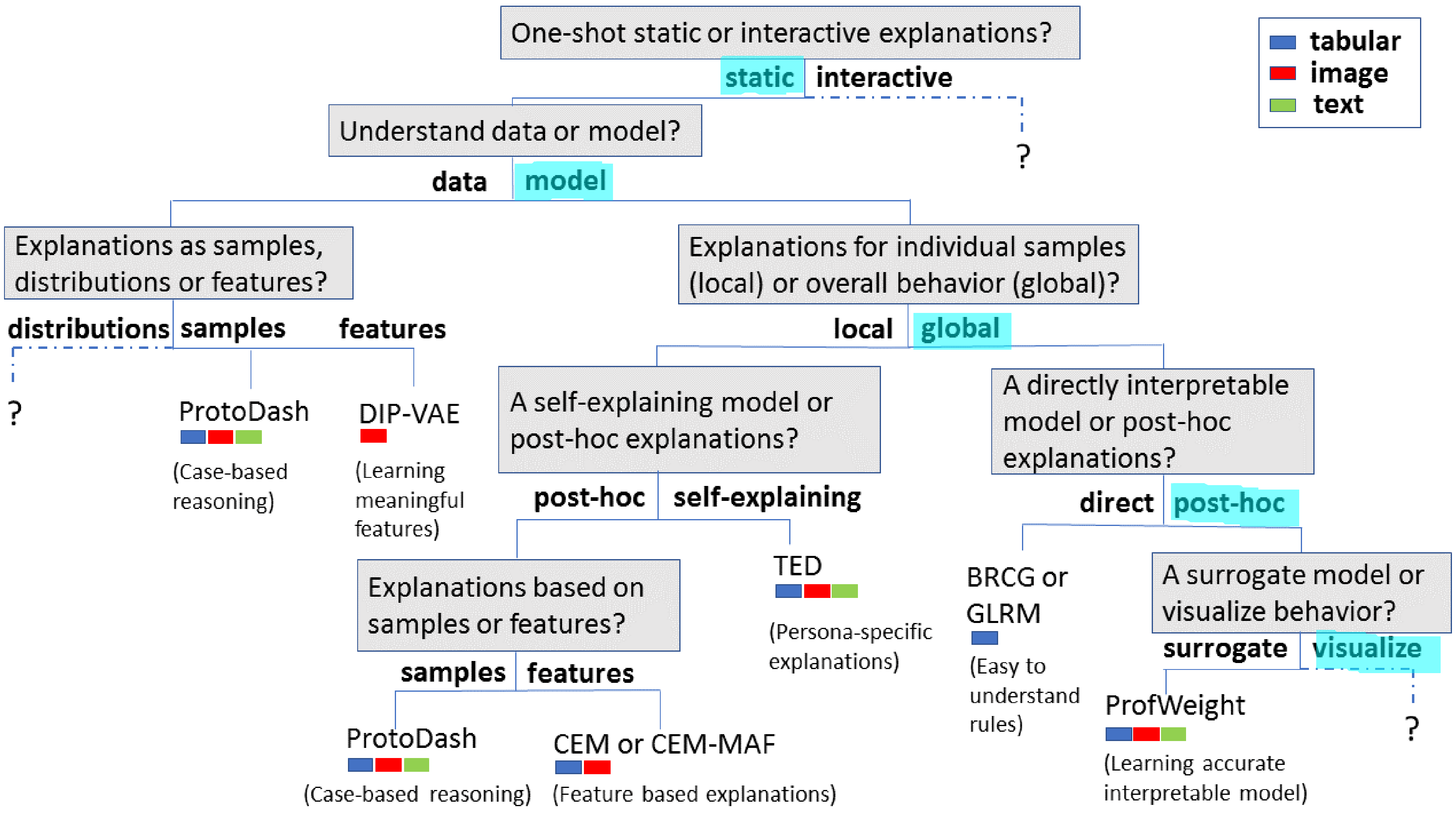
Bank executive

Ogólna ocena modelu. Surrogate model albo destylacja wiedzy do interpretowalnego



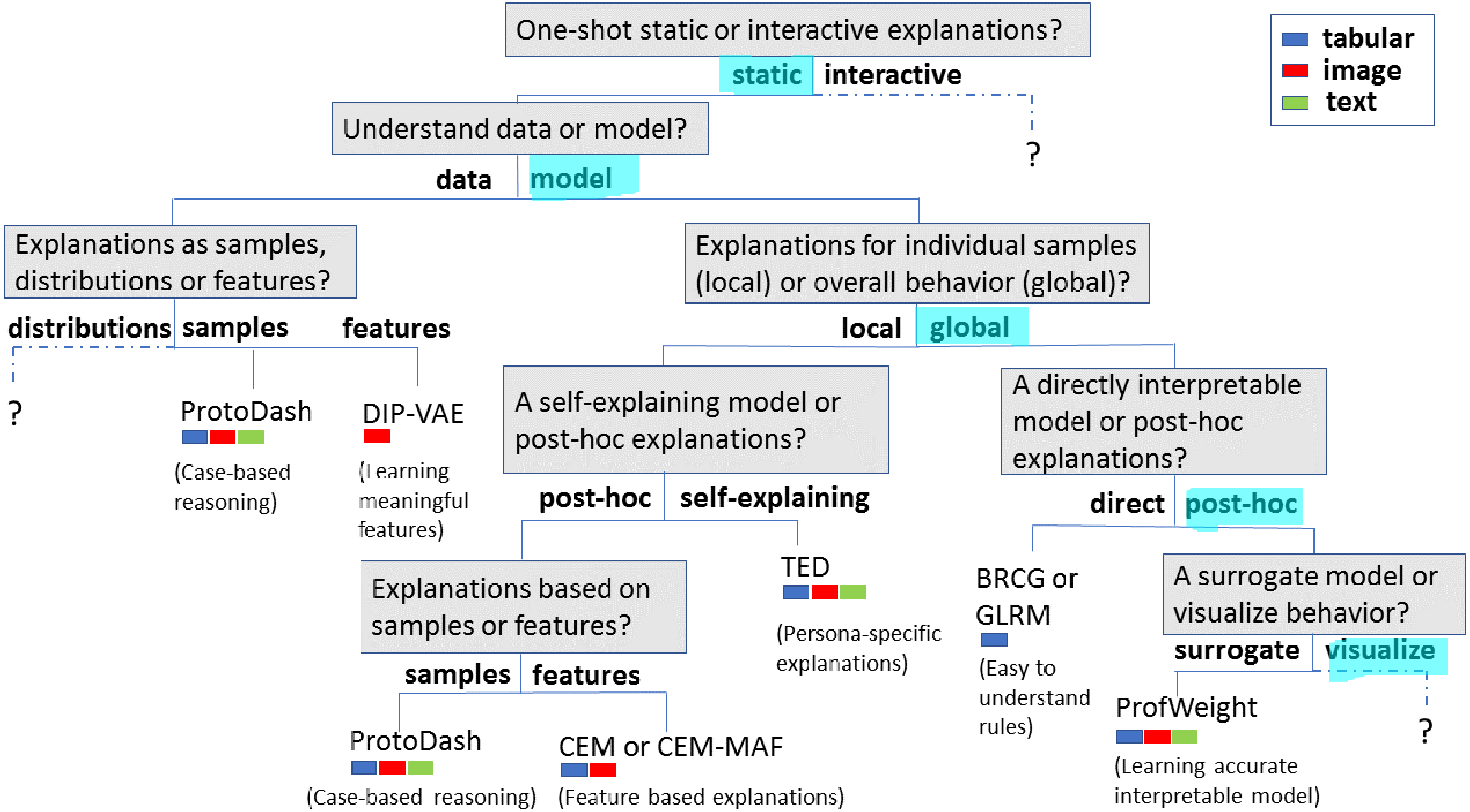
Popularne klasy wyjaśniaczy

Saliency methods, LIME, SHAP, counterfactual explanations



Popularne klasy wyjaśniaczy

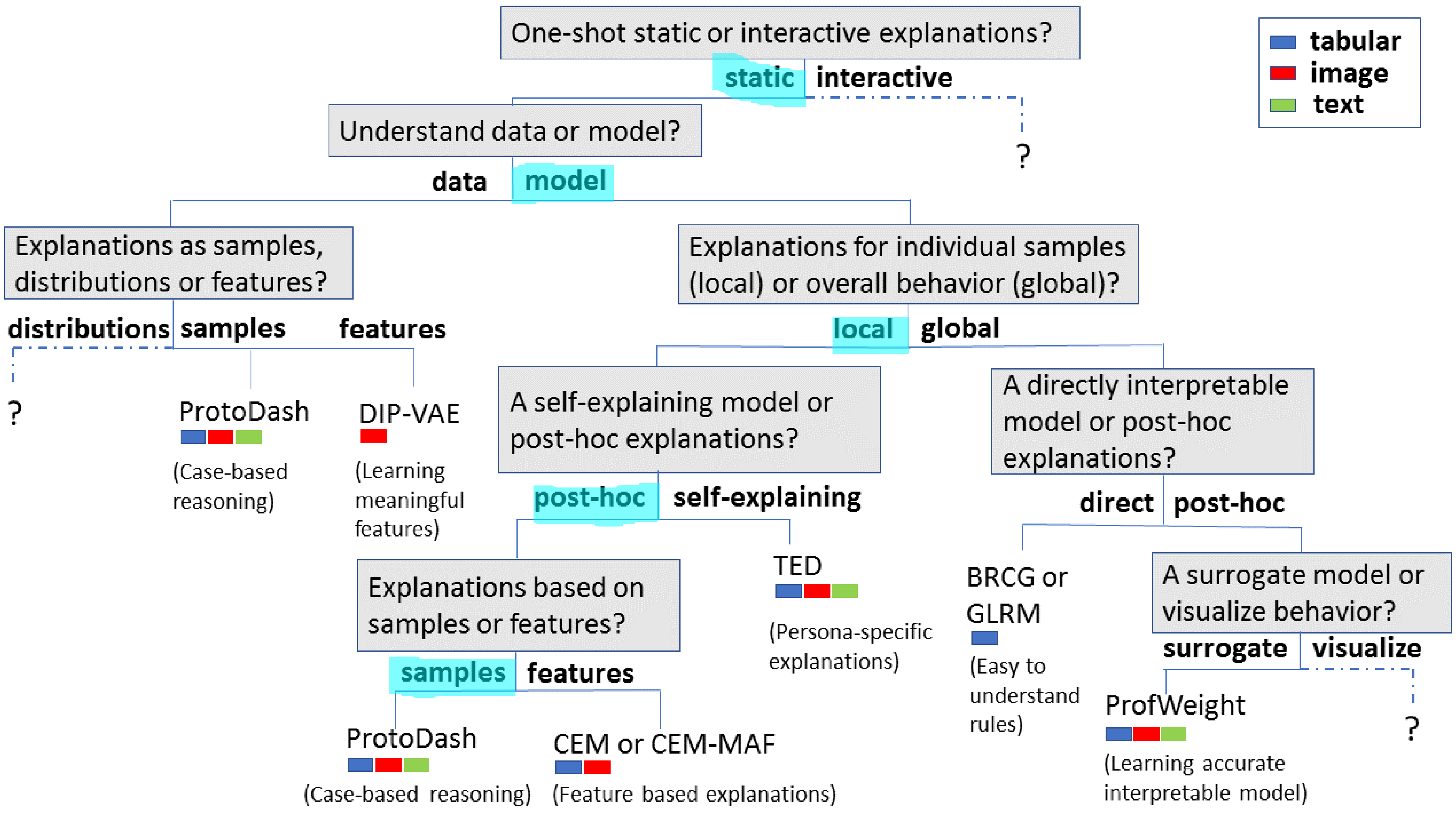
Wizualizacje sieci neuronowych - pośrednie reprezentacje warstw.



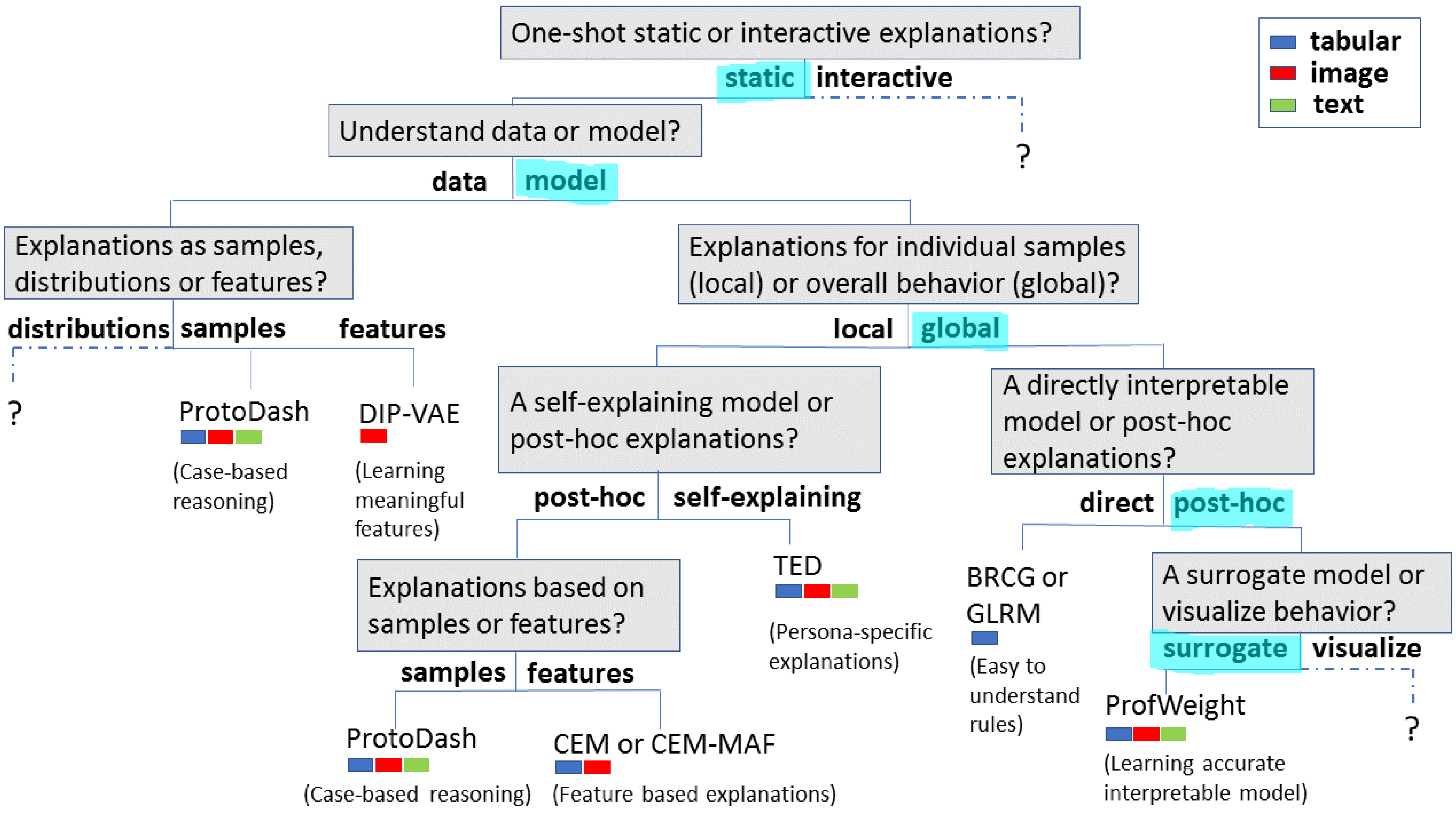
Popularne klasy wyjaśniaczy

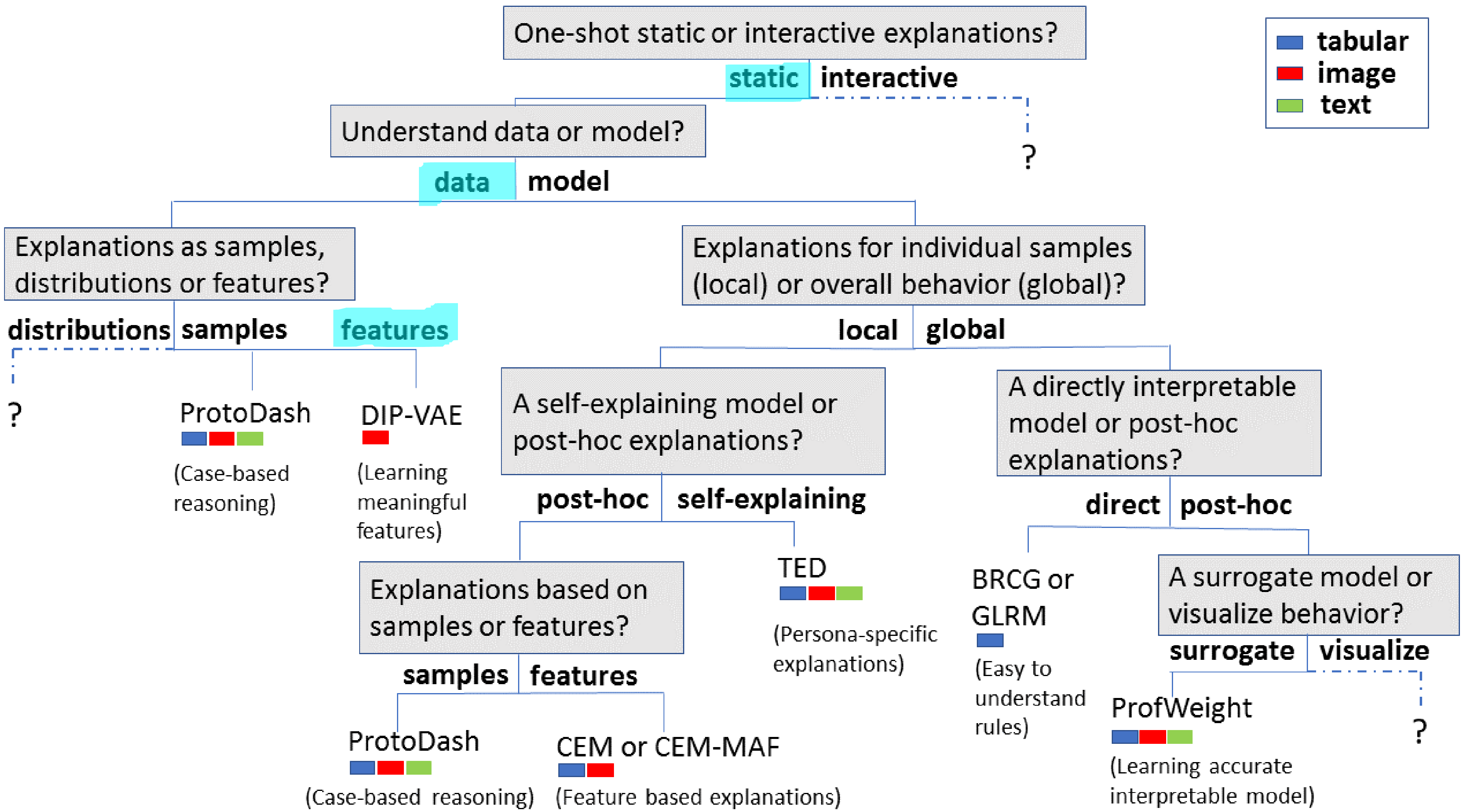
Ważność zmiennych, PDP plots

Popularne klasy wyjaśniaczy
Exemplar methods - wyjaśnienia na podstawie podobnych przykładów



Popularne klasy wyjaśniaczy Knowledge distillation methods

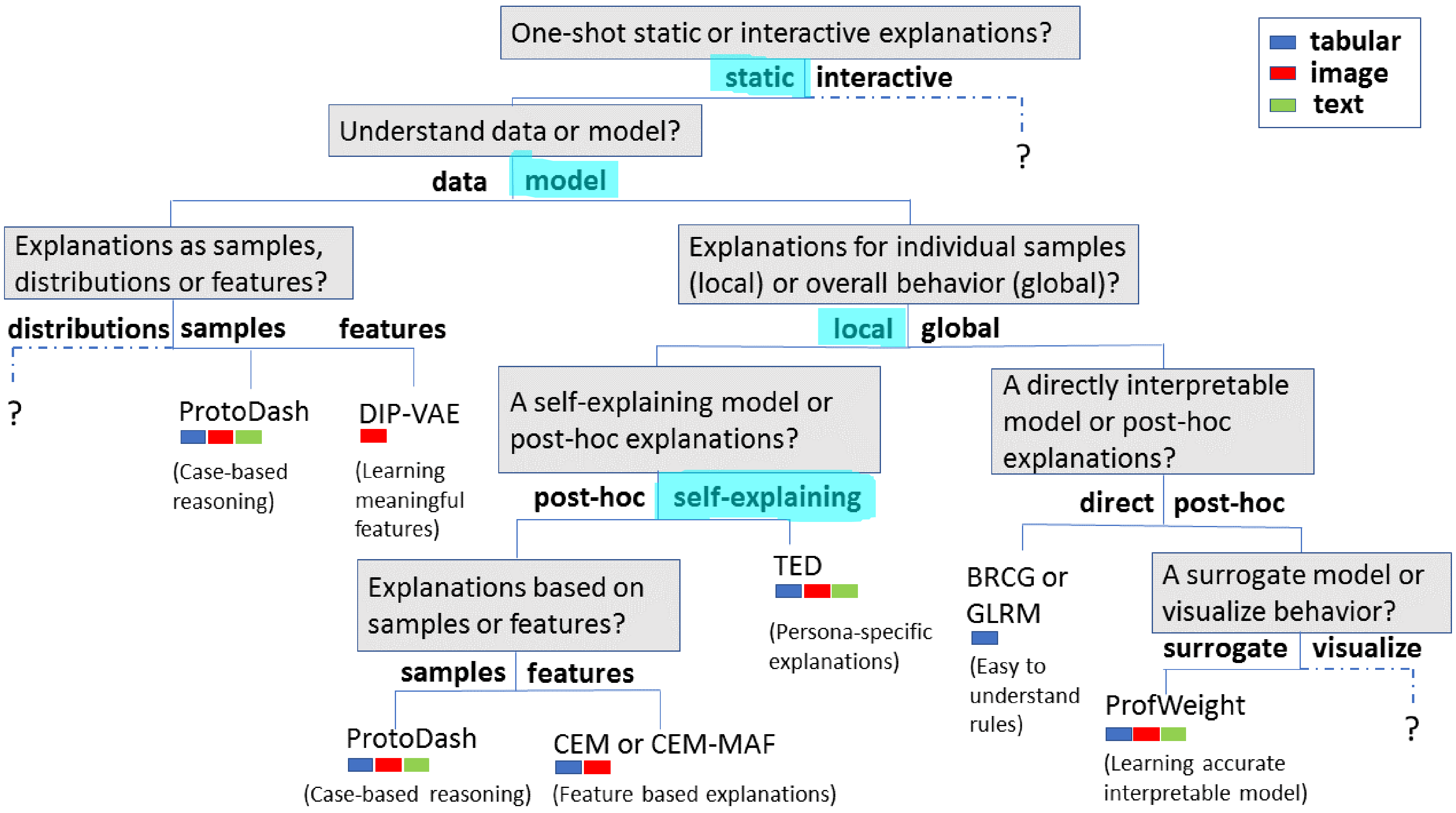


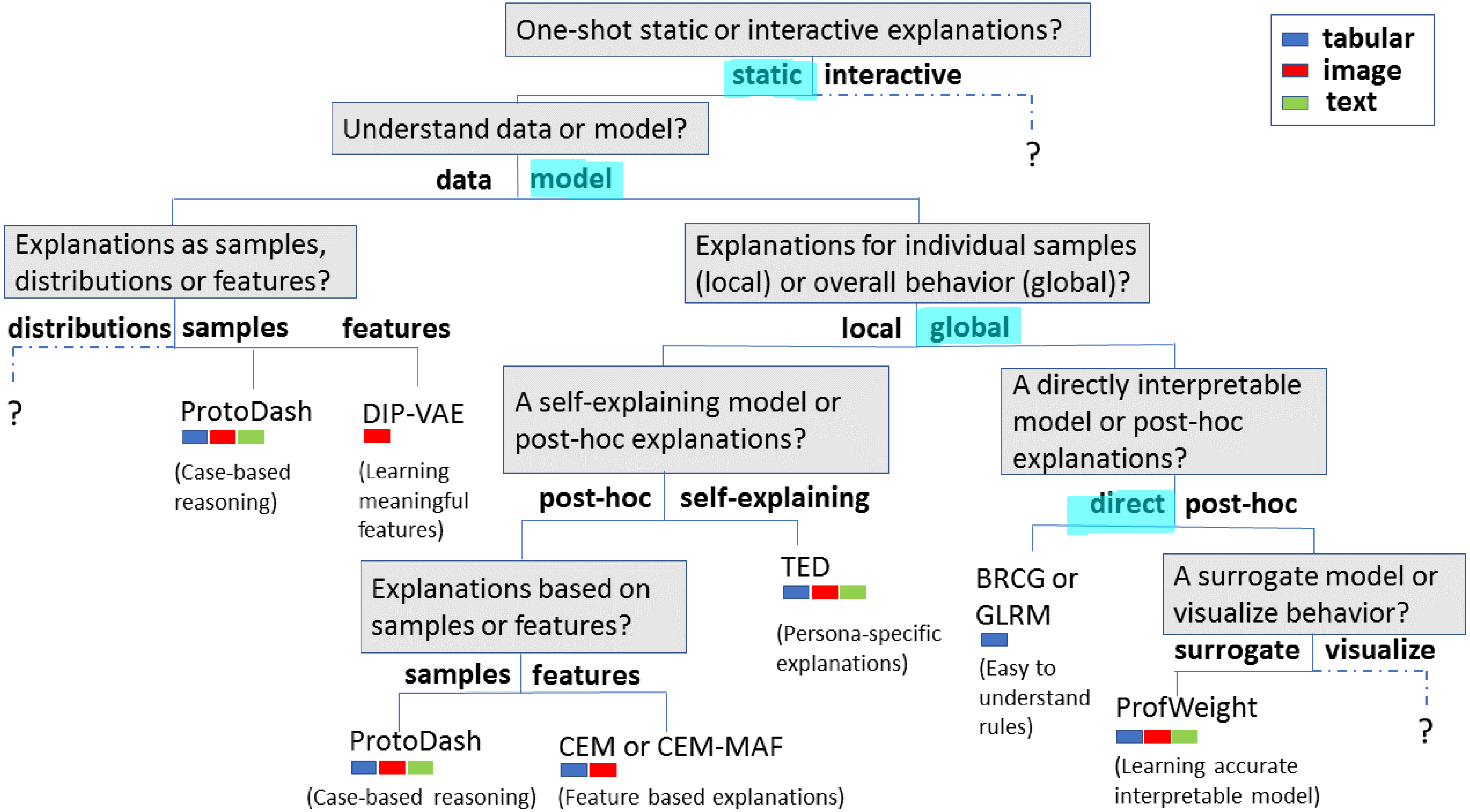


Popularne klasy wyjaśniaczy
High-level feature learning methods.

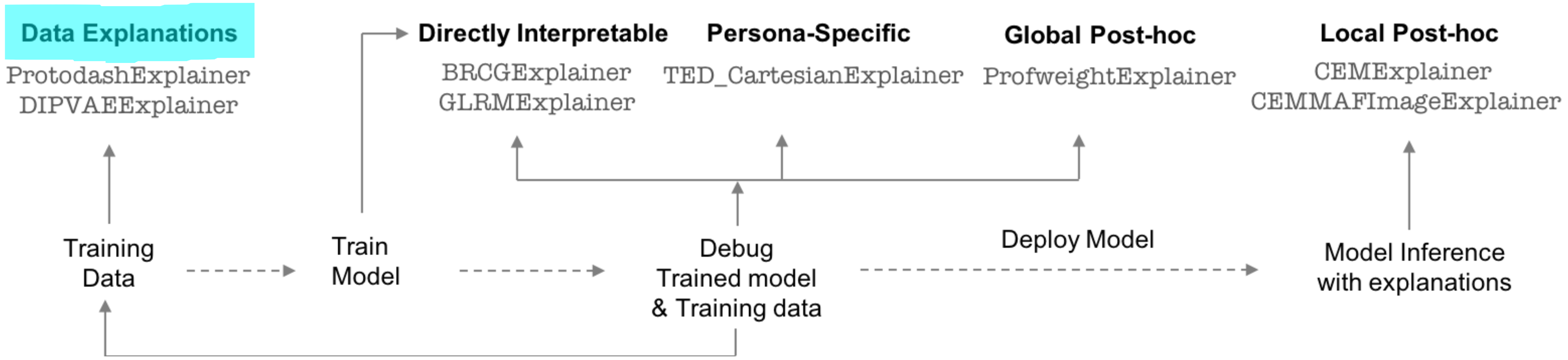
Metody unsupervised (variational autoencoder, GANs) lub metody nadzorowane?

Popularne klasy wyjaśniaczy
Methods that provide rationales.

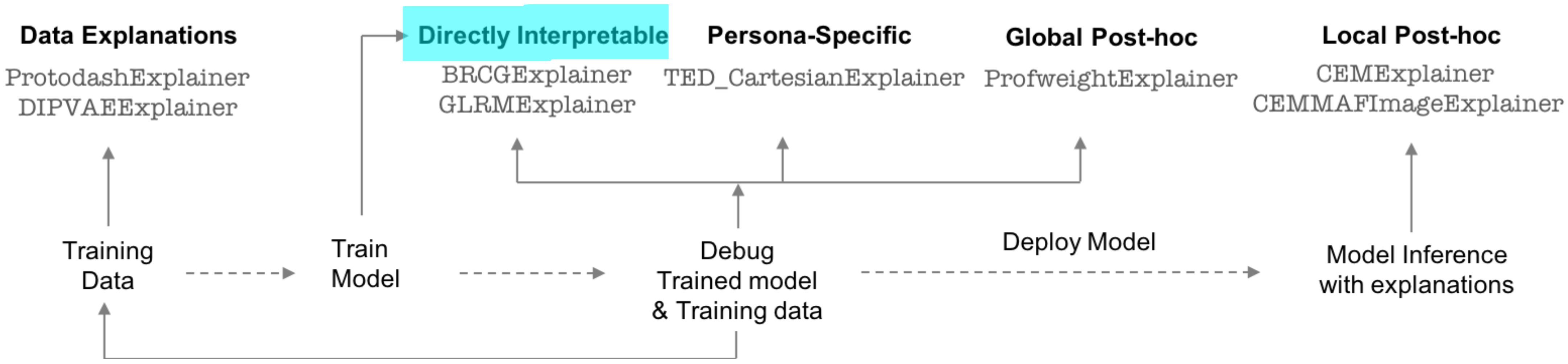




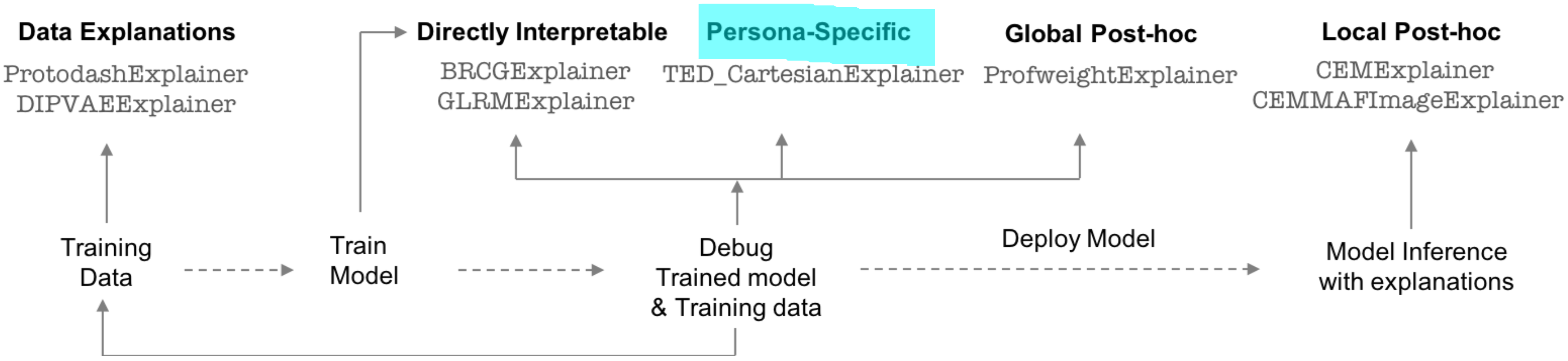
Popularne klasy wyjaśniaczy Restricted NN architectures – metody ograniczające architektury, tak aby były interpretowalne.



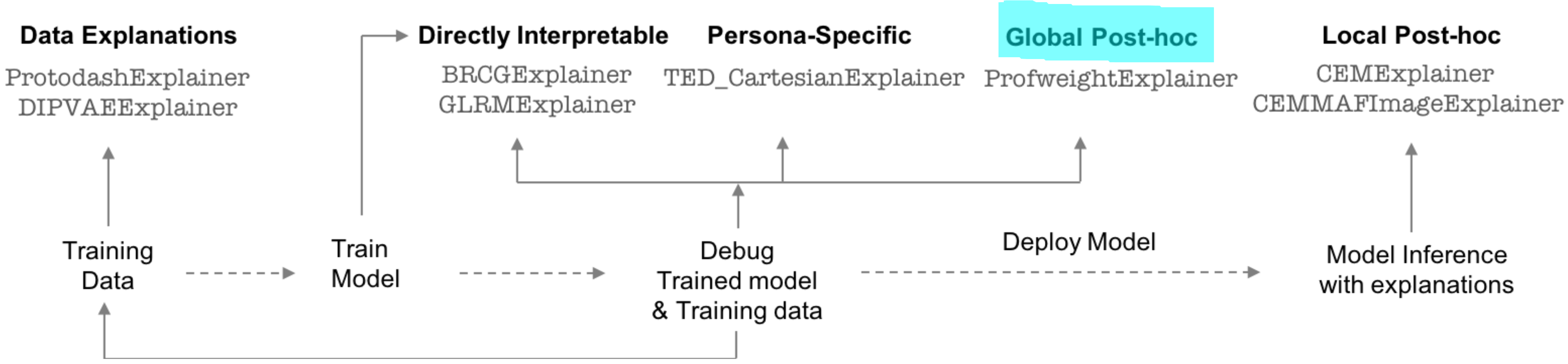
- ProtodashExplainer - wybiera reprezentatywną próbkę podsumowującą zbiór danych lub wyjaśnia przypadek testowy. Także, pokazuje outliery.
- DIPVAEExplainer - uczy się wysokopoziomowych cech z obrazków, które mogą mieć semantyczną interpretację



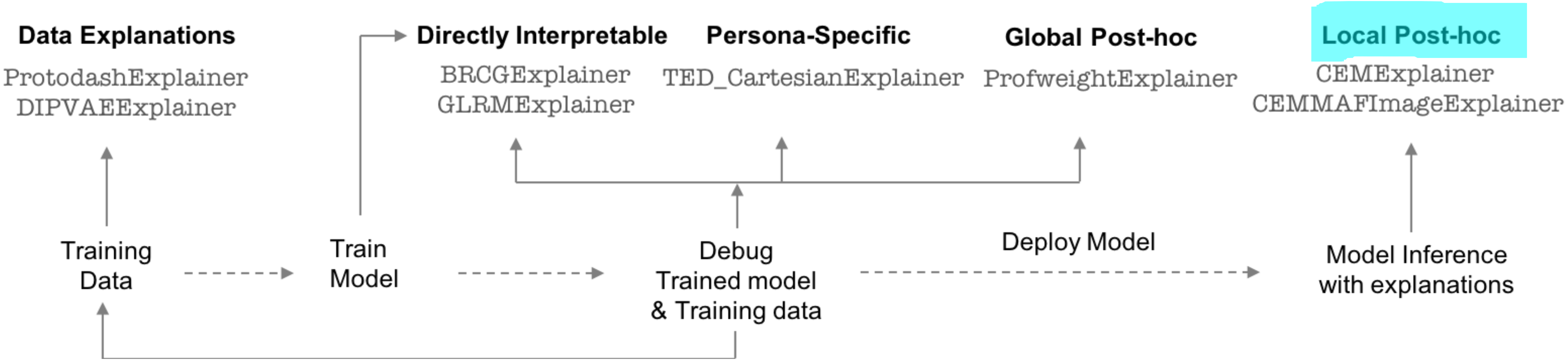
- BRCGExplainer - uczy się prostej, interpretowalnej reguły logicznej w postaci DNF dla klasyfikacji binarnej
- GLRMExplainer – ważona reguła logiczna



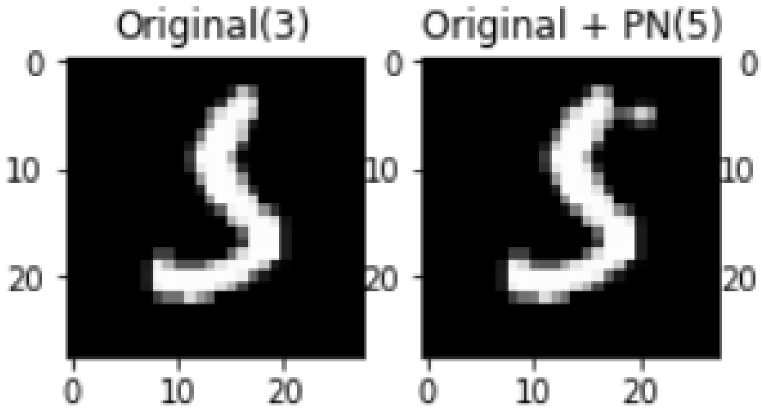
- TED_CartesianExplainer - uczy się odpowiedzi + wyjaśnień (wymaga dostarczenia wyjaśnień dla zbioru uczącego)

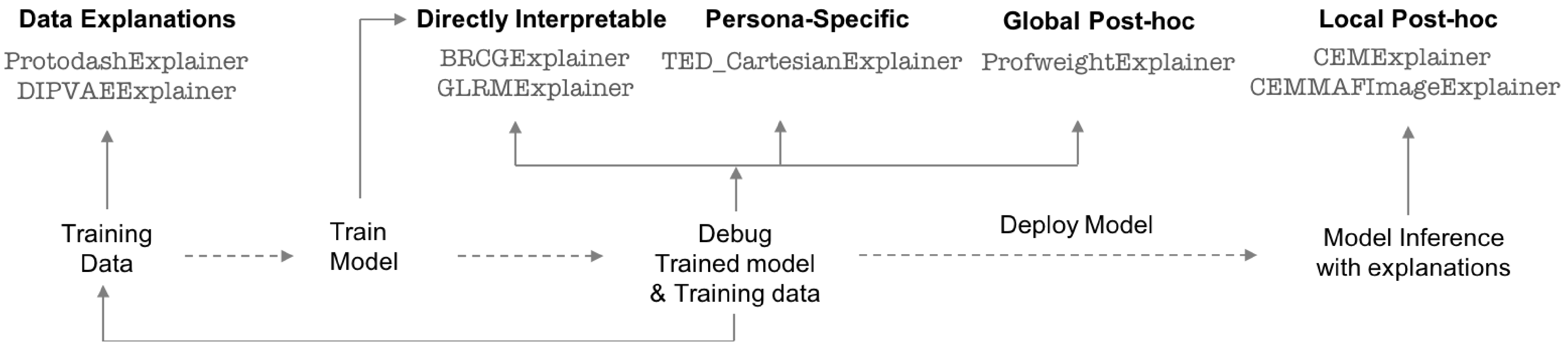


- ProfweightExplainer – na podstawie warstw w sieci neuronowej uczy się nadawać wagi obserwacjom treningowym, tak żeby prosty, interpretowalny model miał dobrą skuteczność. Wagi nadajemy w zależności od tego jak "łatwo" nauczyć się danego przykładu.



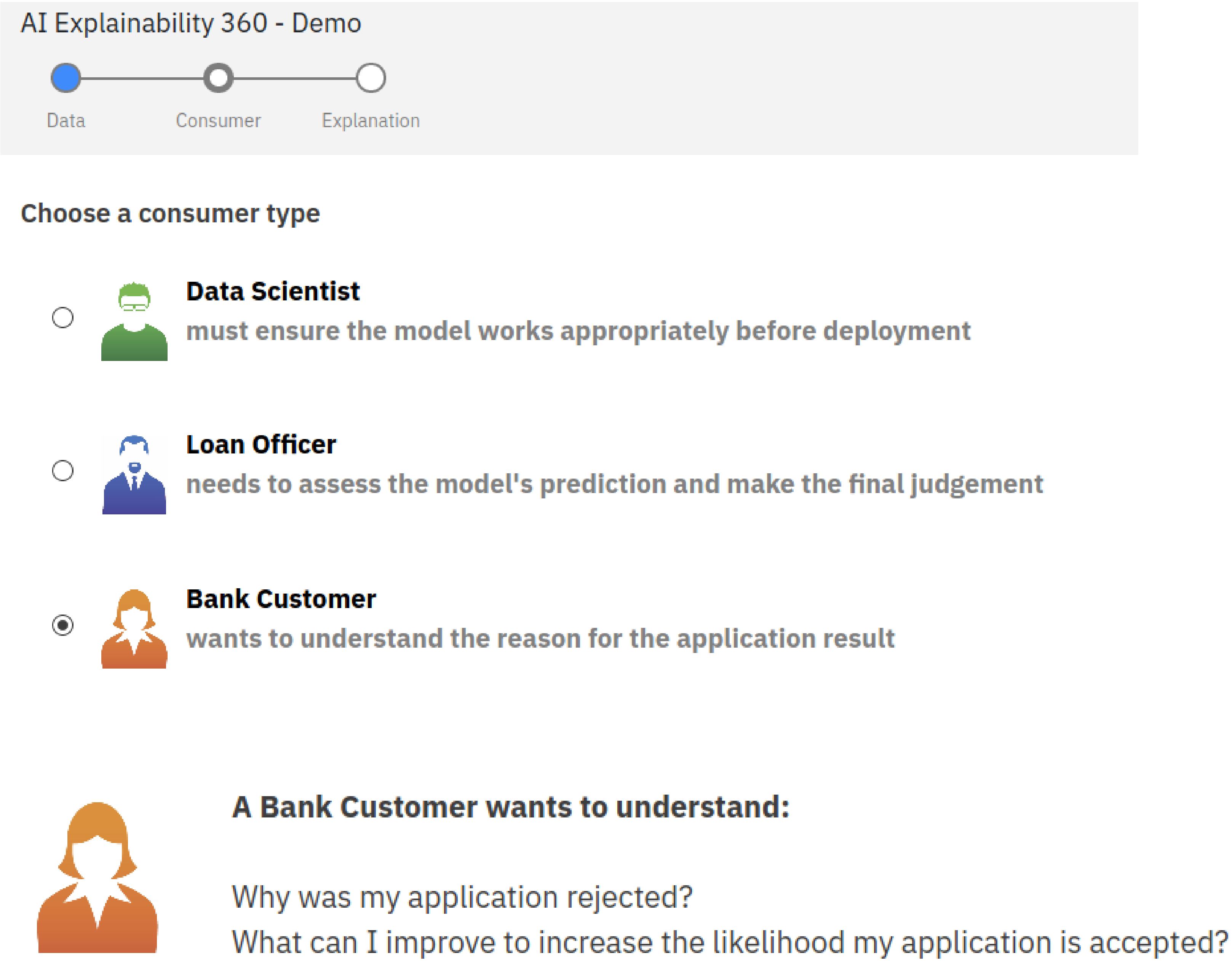
- CEMExplainer – generuje lokalne wyjaśnienie, które mówi jakie minimum trzeba zachować aby utrzymać odpowiedź modelu, a co zmieniłoby odpowiedź
- CEMMAFImageExplainer – j.w. ale na wysokopoziomowych cechach dla obrazków





Toolkit	Data Explanations	Directly Interpretable	Local Post-Hoc	Global Post-Hoc	Persona-Specific Explanations	Metrics
AIX360	✓	✓	✓	✓	✓	✓
Alibi [1]			✓			
Skater [7]		✓	✓	✓		
H2O [4]		✓	✓	✓		
InterpretML [6]		✓	✓	✓		
EthicalML-XAI [3]				✓		
DALEX [2]			✓	✓		
tf-explain [8]			✓	✓		
iNNvestigate [5]			✓			

Table 1: Comparison of AI explainability toolkits.


AI Explainability 360 - Demo
Data
Consumer
Explanation
Choose a consumer type
Data Scientist
must ensure the model works appropriately before deployment
Loan Officer
needs to assess the model's prediction and make the final judgement
Bank Customer
wants to understand the reason for the application result
A Bank Customer wants to understand:
Why was my application rejected?
What can I improve to increase the likelihood my application is accepted?

Jason

Denied

Ann

Denied

Julia

Denied

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

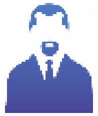
1. The value of **Consolidated risk markers** is **65**. It needs to be around **72** for the application to be approved.
2. The value of **Average age of accounts in months** is **52**. It needs to be around **68** for the application to be approved.
3. The value of **Months since most recent credit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

AI Explainability 360 - Demo

Data — Consumer — Explanation

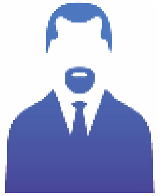
Choose a consumer type

- ( ) **Data Scientist**
  must ensure the model works appropriately before deployment



- (•) **Loan Officer**
  needs to assess the model's prediction and make the final judgement

- ( ) **Bank Customer**
  wants to understand the reason for the application result



A Loan Officer wants to understand:

Why is the model recommending this person’s credit be approved or denied?
How can I inform my decision to accept or reject a line of credit by looking at similar individuals?

Alice

Approved

Robert

Denied

	Alice	Mia	Kate	Cala
Outcome	-	Paid	Paid	Paid
Similarity to Alice (from 0 to 1)	-	0.765	0.081	0.065
ExternalRiskEstimate	82	85	80	89
MSinceOldestTradeOpen	280	223	382	379
MSinceMostRecentTradeOpen	13	13	4	156
AverageMInFile	102	87	90	257
NumSatisfactoryTrades	22	23	21	3
NumTrades60Ever2DerogPubRec	0	0	0	0
NumTrades90Ever2DerogPubRec	0	0	0	0
PercentTradesNeverDelq	91	91	95	100

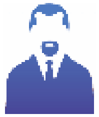
	Robert	James	Danielle	Franklin
Outcome	-	Defaulted	Defaulted	Defaulted
Similarity to Robert (from 0 to 1)	-	0.690	0.114	0.108
ExternalRiskEstimate	78	71	72	69
MSinceOldestTradeOpen	82	95	166	193
MSinceMostRecentTradeOpen	5	1	12	12
AverageMInFile	54	43	74	167
NumSatisfactoryTrades	33	33	37	36
NumTrades60Ever2DerogPubRec	0	0	1	0
NumTrades90Ever2DerogPubRec	0	0	1	0
PercentTradesNeverDelq	100	100	95	100
MSinceMostRecentDelq	0	0	7	0
MaxDelq2PublicRecLast12M	7	7	4	7

AI Explainability 360 - Demo

Data Consumer Explanation

Choose a consumer type

Data Scientist
must ensure the model works appropriately before deployment



Loan Officer
needs to assess the model's prediction and make the final judgement

Bank Customer
wants to understand the reason for the application result

A Data Scientist wants to understand:

What is the overall logic of the model in making decisions?
Is the logic reasonable, so that we can deploy the model with confidence?

ExternalRiskEstimate

- For every increase of 10 in ExternalRiskEstimate, increase score by 0.266.
- If ExternalRiskEstimate > 69, increase score by an additional 0.035.
- If ExternalRiskEstimate > 72, increase score by an additional 0.108.
- If ExternalRiskEstimate > 75, increase score by an additional 0.263.