Università degli Studi di Padova

Dipartimento Matematica

Interpretabilità nel Machine Learning

Mirko Polato, PhD

mpolato@math.unipd.it

Cosa vedremo oggi

Cosa

- Cos'è l'interpretabilità?
- Cosa comporta avere un metodo interpretabile?

Perché

- Perché vogliamo l'interpretabilità?

Quando

- Quando è utile un metodo interpretabile

Come

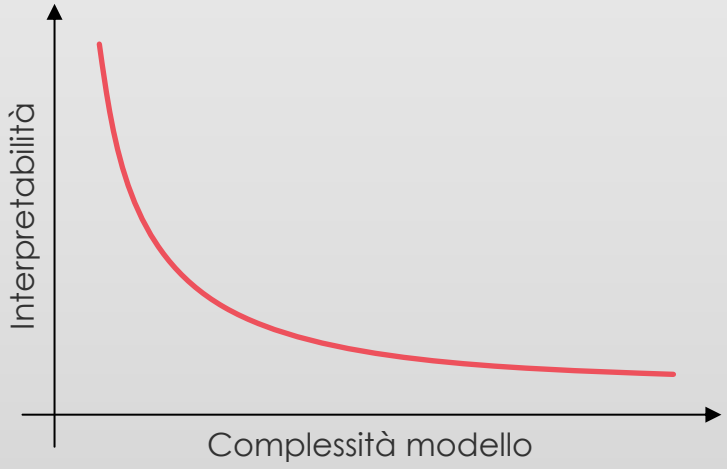
- Come si rende un metodo interpretabile?
- Come possiamo spiegare l'output di un modello?

Alcuni dei contenuti di questa presentazione sono stati estrapolati da ***Molnar (2018)***

Cos'è l'interpretabilità?

"L'interpretabilità è il grado con cui un umano comprende la causa di una decisione"

Miller (2017)



Interpretabilità “in action”

- **Applicazioni con conseguenze significative:** giustizia, sanità, self-driving cars…
- **Applicazioni mediche**: una diagnosi deve poter essere spiegata
- **Controllo su eventuali bias:** modelli ML che discriminano minoranze (fairness)
- **Recommender System:** spiegare la raccomandazione aiuta nel mantenere fiducia nel sistema
- **GDPR**: decisioni prese da machine con conseguenze legali devono essere spiegate

Quando è utile?

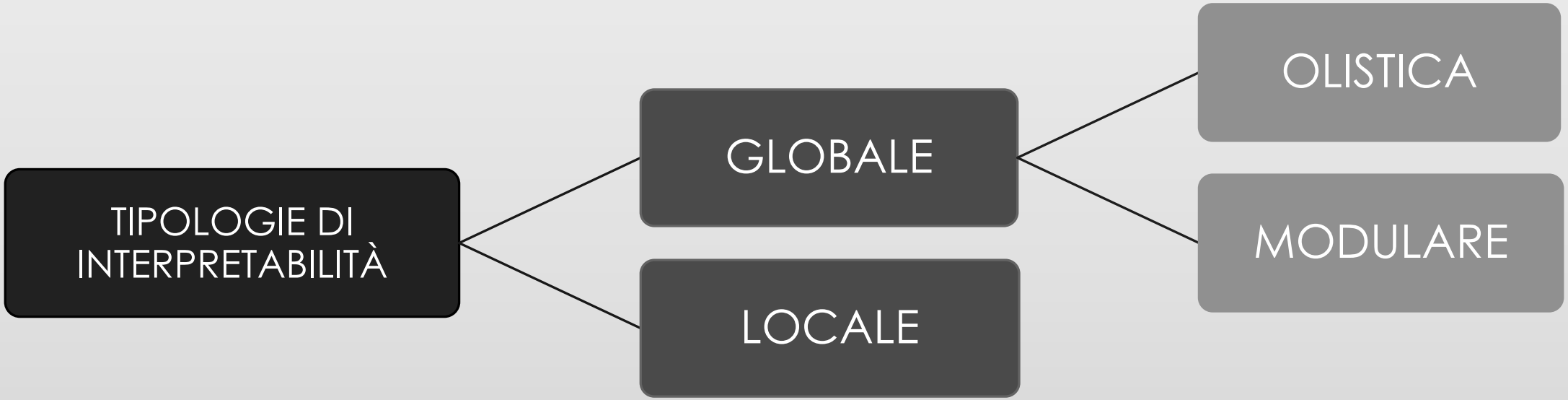
UTILE/DESIDERABILE

- Conoscere il 'perché' è parte fondamentale del problema
- **Debugging**
- Incrementare l'**accettazione sociale**
- **Fairness** check
- **Trust**
- **Privacy**

- Il modello **non** ha un particolare **impatto**
- Il problema è **ben conosciuto**
- Evitare **manipolazioni**

NON NECESSARIA

Tipi di interpretabilità



Properietà

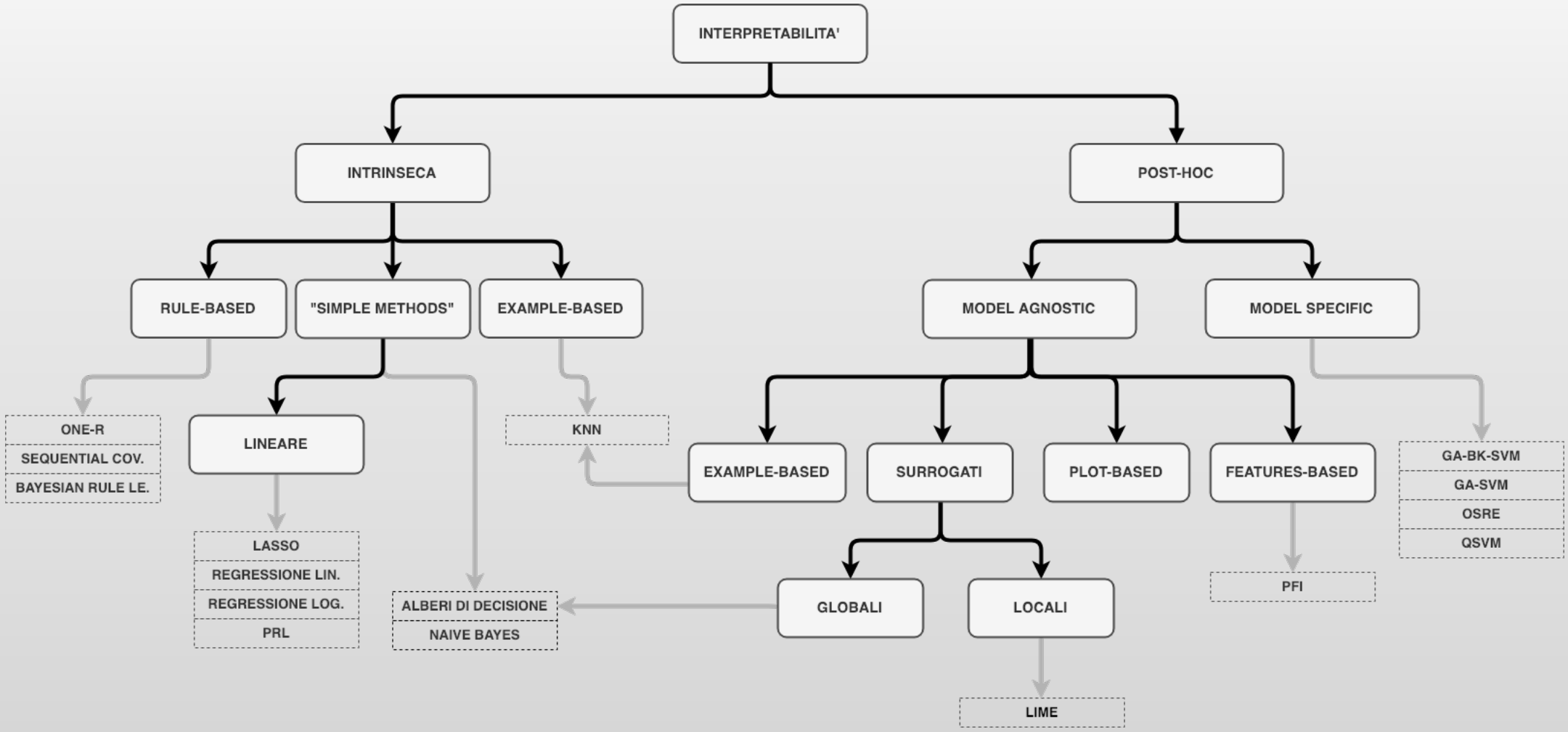
Explanation methods

- Espressività
- Translucenza
- Portabilità
- Complessità computazionale

Individual explanation

- Accuratezza
- Fedeltà
- Stabilità
- Comprensibilità

Tassonomia


INTERPRETABILITA'
INTRINSECA
POST-HOC
RULE-BASED
"SIMPLE METHODS"
EXAMPLE-BASED
MODEL AGNOSTIC
MODEL SPECIFIC
ONE-R
SEQUENTIAL COV.
BAYESIAN RULE LE.
LINEARE
KNN
EXAMPLE-BASED
SURROGATI
PLOT-BASED
FEATURES-BASED
GA-BK-SVM
GA-SVM
OSRE
QSVM
LASSO
REGRESSIONE LIN.
REGRESSIONE LOG.
PRL
ALBERI DI DECISIONE
NAIVE BAYES
GLOBALI
LOCALI
PFI
LIME

Come valutare un metodo interpretabile

Valutazione a livello applicazione (task reale)

La spiegazione è testata sul campo direttamente dagli utenti (esperti)

Valutazione a livello umano (task semplificato)

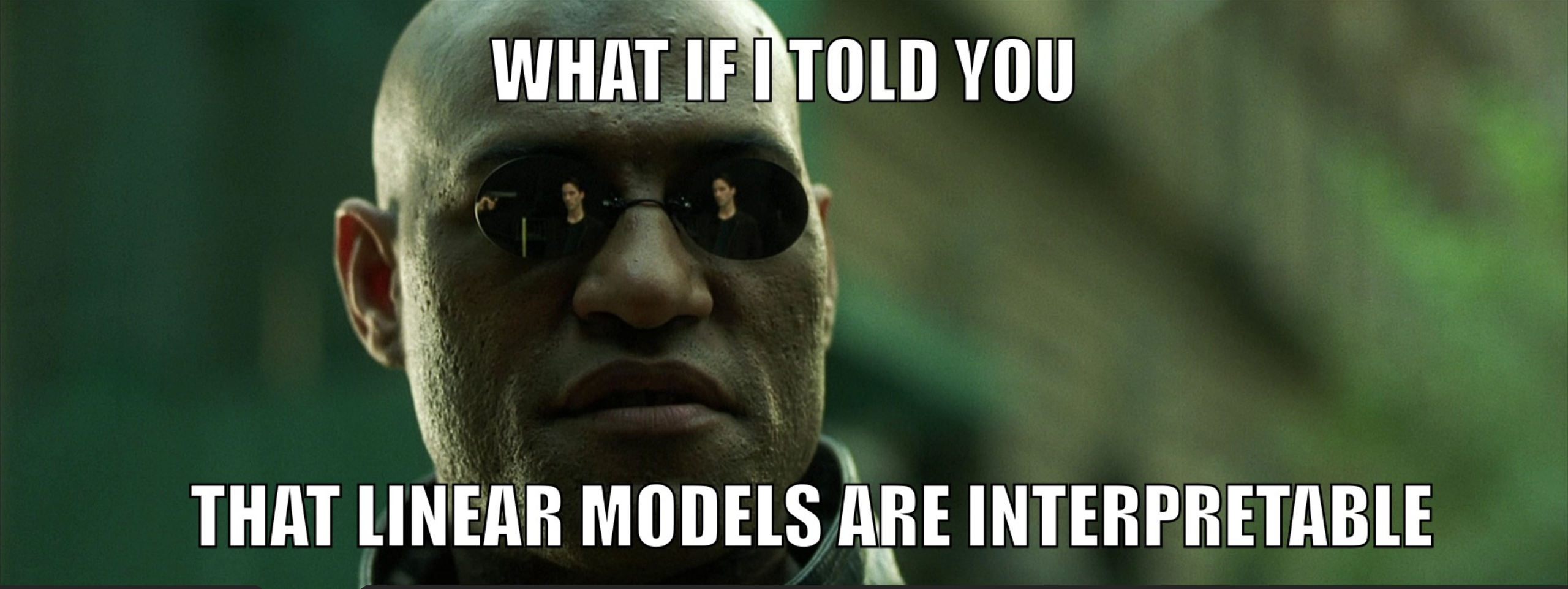
Semplificazione della valutazione a livello applicazione (utenti non esperti)

Valutazione a livello funzionale (task proxy)

Non richiede l'intervento umano. Utile a monte di una precedente valutazione a livello umano.

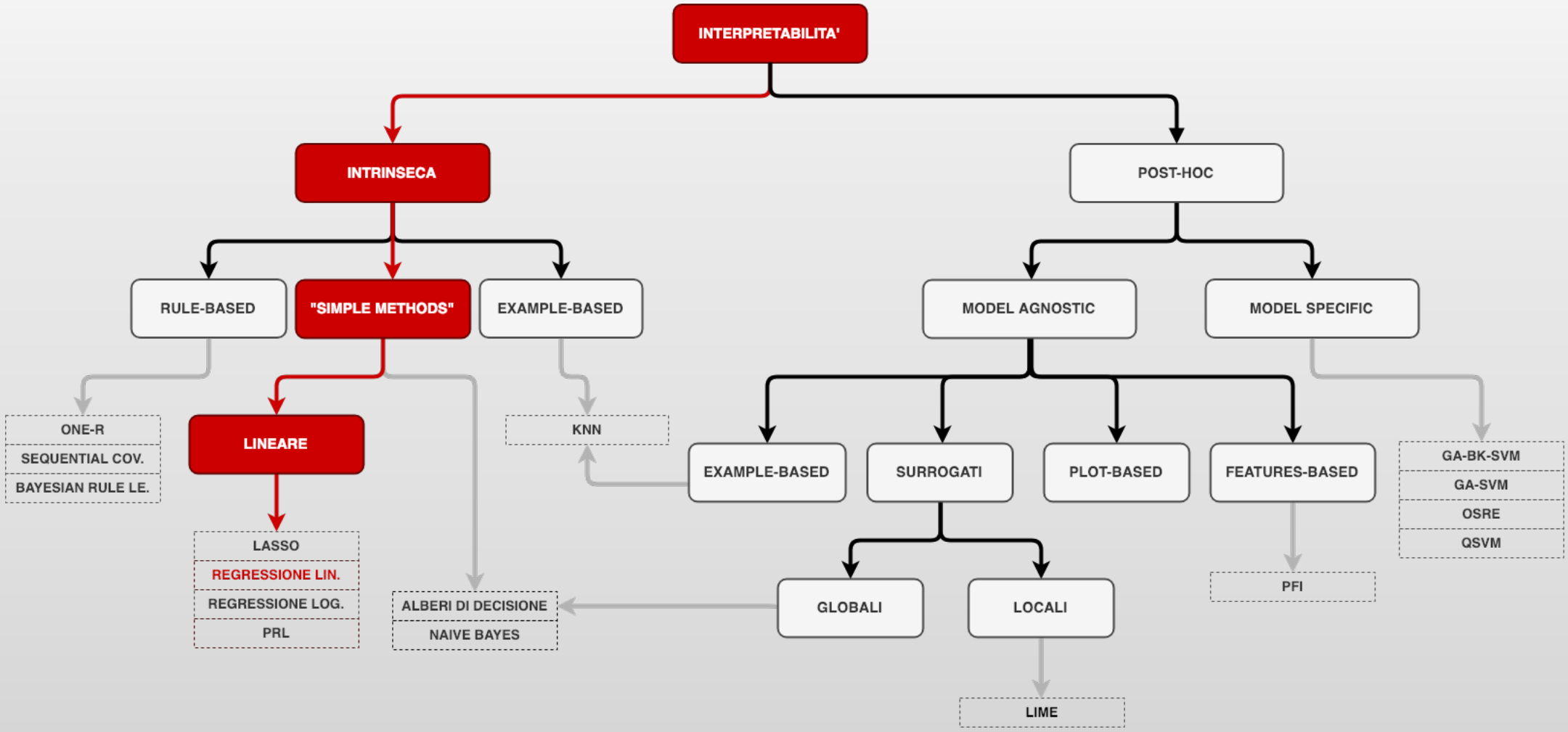
Metodi nel “dettaglio”

- Regressione Lineare
- Alberi di decisione
  - metodo intrinsecamente interpretabile
  - metodo surrogato
- PRL: Preference and rule learning [Polato (2019)]
- GA-BK-SVM [Polato (2018a, 2018b)]



Regressione lineare

Regressione Lineare

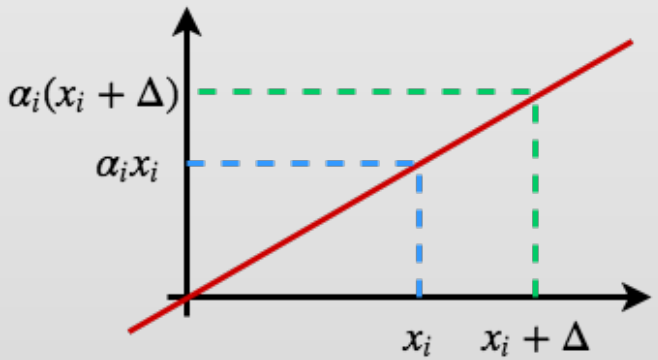


Interpretare un regressore lineare

Per un'istanza $(\boldsymbol{x}, y)$, con $\boldsymbol{x} \in \mathbb{R}^p$ e $y \in \mathbb{R}$ il modello appreso ha la seguente forma

$$y = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p$$

Feature a valori reali



Feature binaria (*)

$x_i \in \{0,1\}$

0 → Nessuna influenza

1 → $+\alpha_i$

Intercetta

Significativa su esempi standardizzati

y dell'esempio standard

(*) può essere generalizzato a feature categoriche

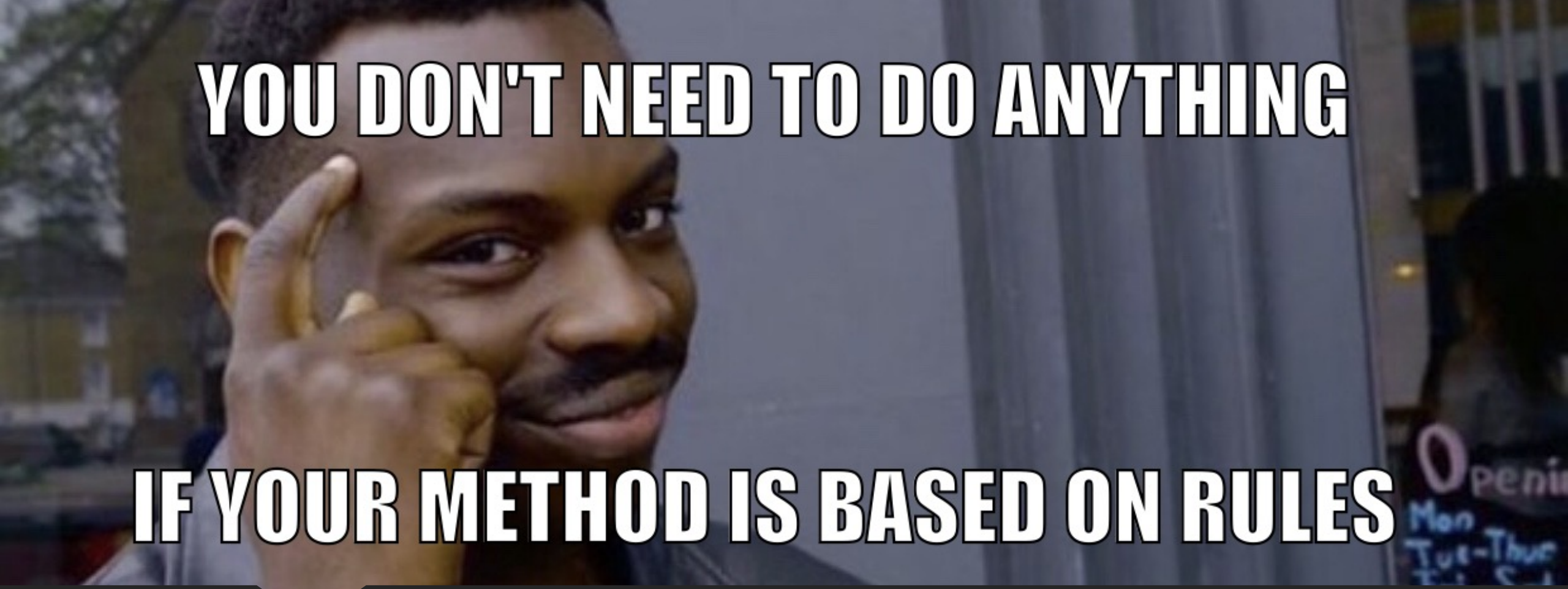
Regressione lineare: pro e contro

PROS

- **Trasparente**
- **Semplice**
- Tra i più conosciuti e implementato in molti framework
- Esistono molte **estensioni**/varianti: LASSO, GLM, SLM…

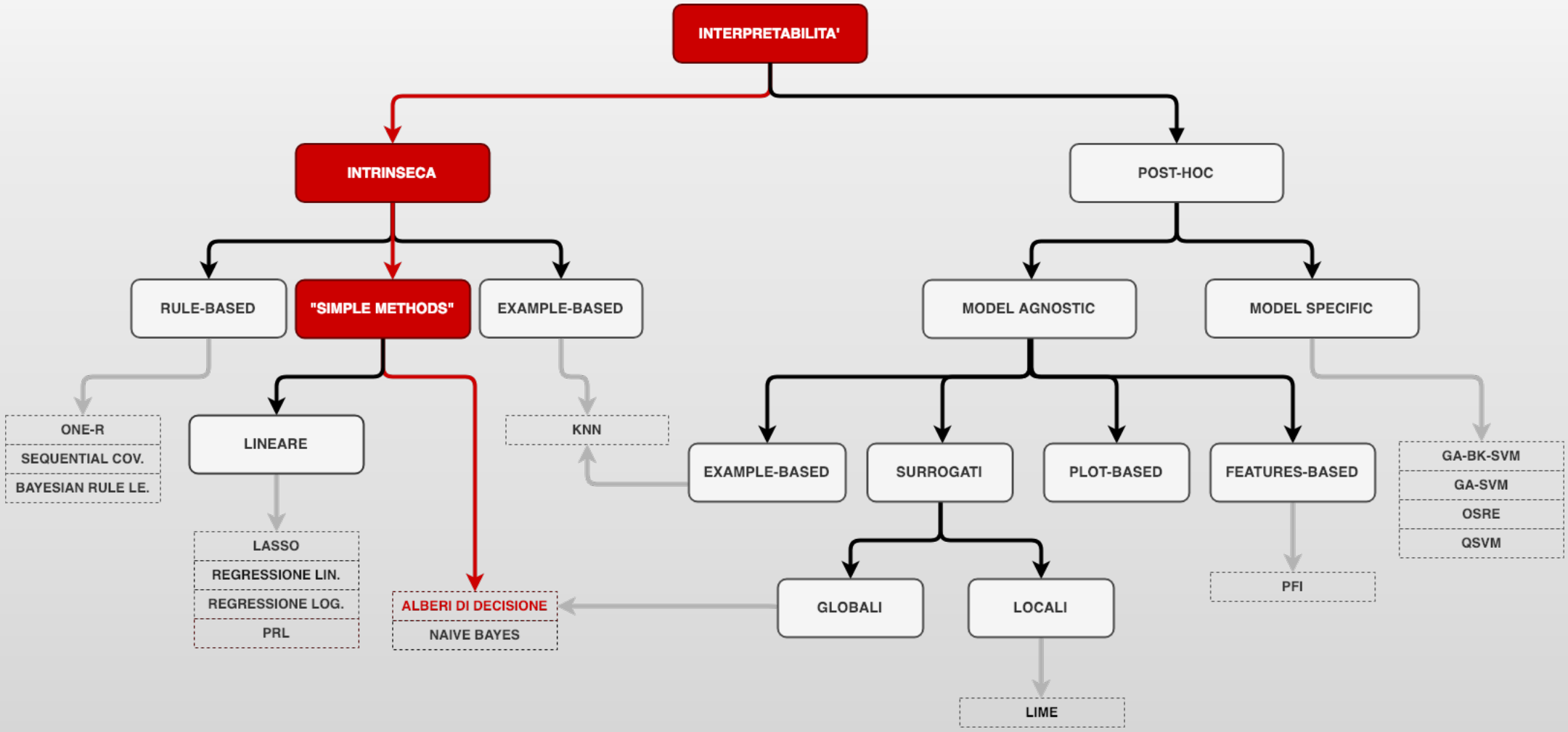
CONS

- **Solo relazioni lineari**
- Generalmente non ha performance allo stato dell'arte
- L'**interpretazione** potrebbe **non** essere **intuitiva**



Alberi di decisione

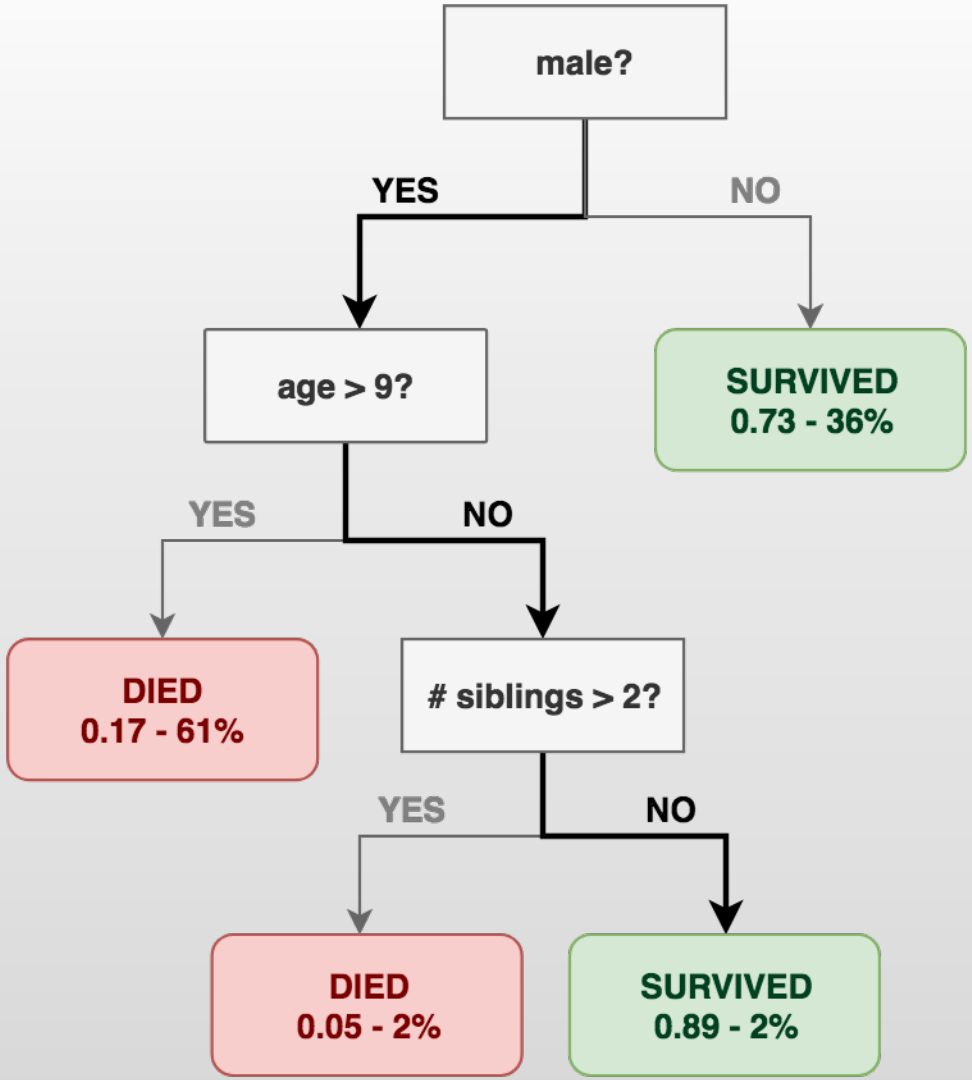
Alberi di decisione



Interpretare un albero di decisione

Maschio, 7 anni, figlio unico

SOPRAVVISSUTO perché maschio con meno di 9 anni e con meno di 2 fratelli



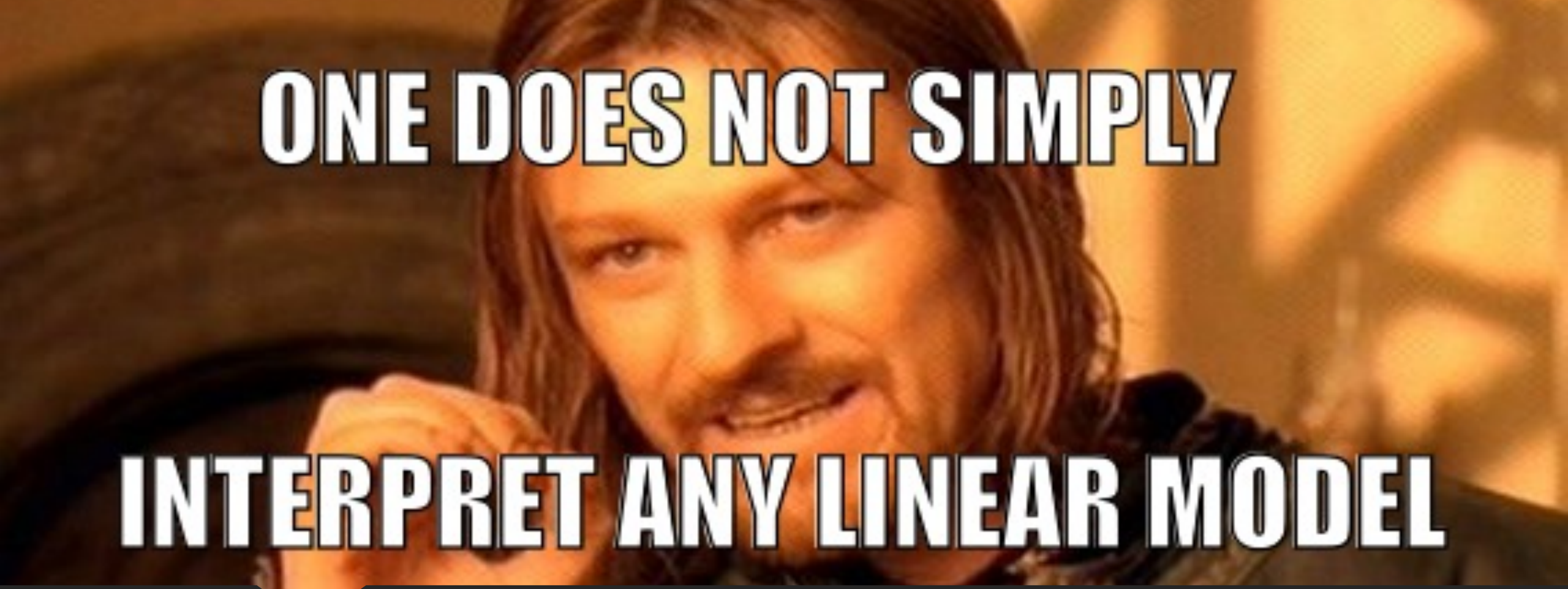
Alberi di decisione: pro e contro

PROS

- Ideale per catturare relazioni tra feature
- L'albero definisce implicitamente **buone spiegazioni**
- E' facile creare scenari **what-if**

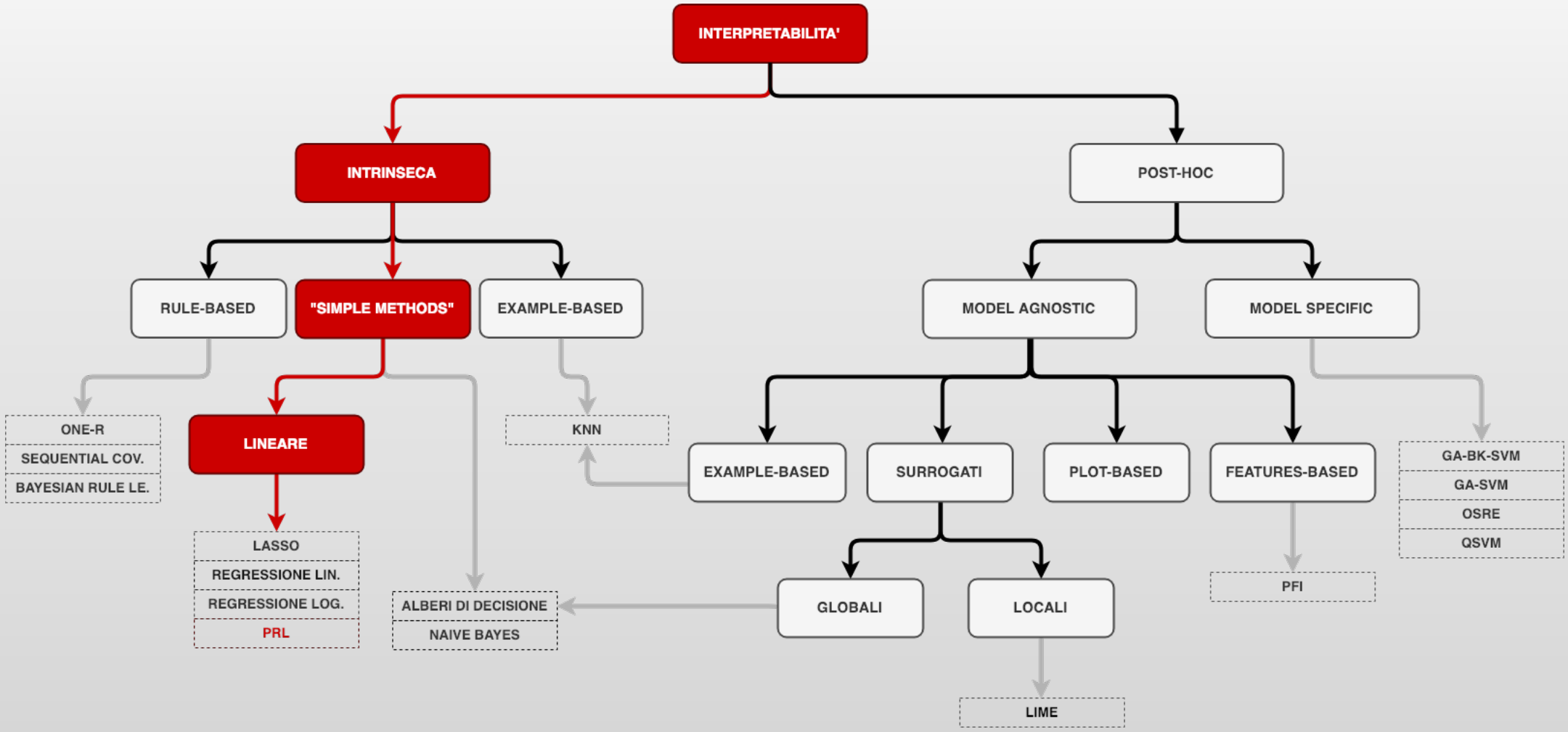
CONS

- **Falliscono** con **relazioni lineari**
- **Lack of smoothness**
- **Instabilità**
- Il numero di foglie incrementa velocemente rendendo difficile la comprensione delle regole



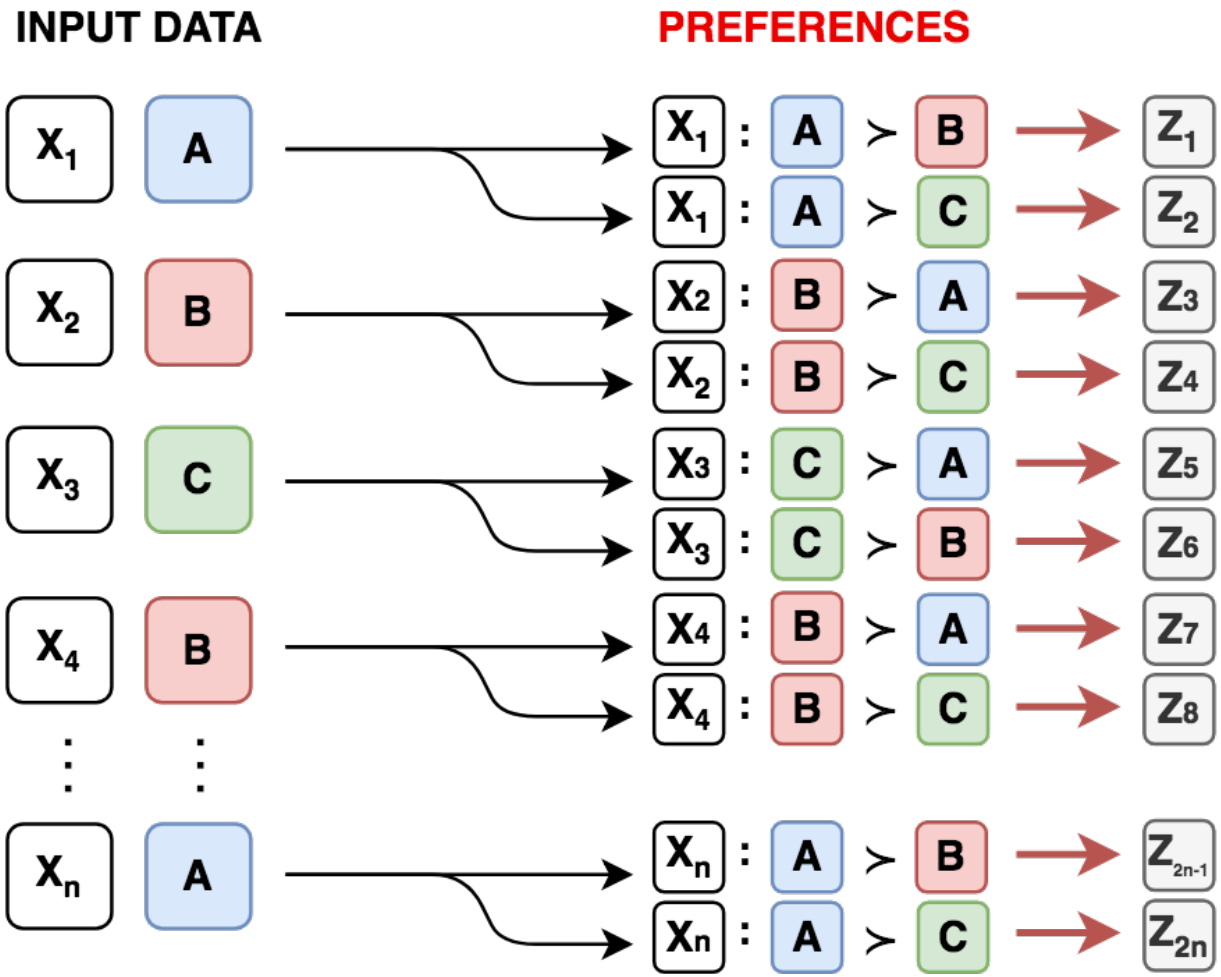
PRL: Preference and Rule Learning

PRL

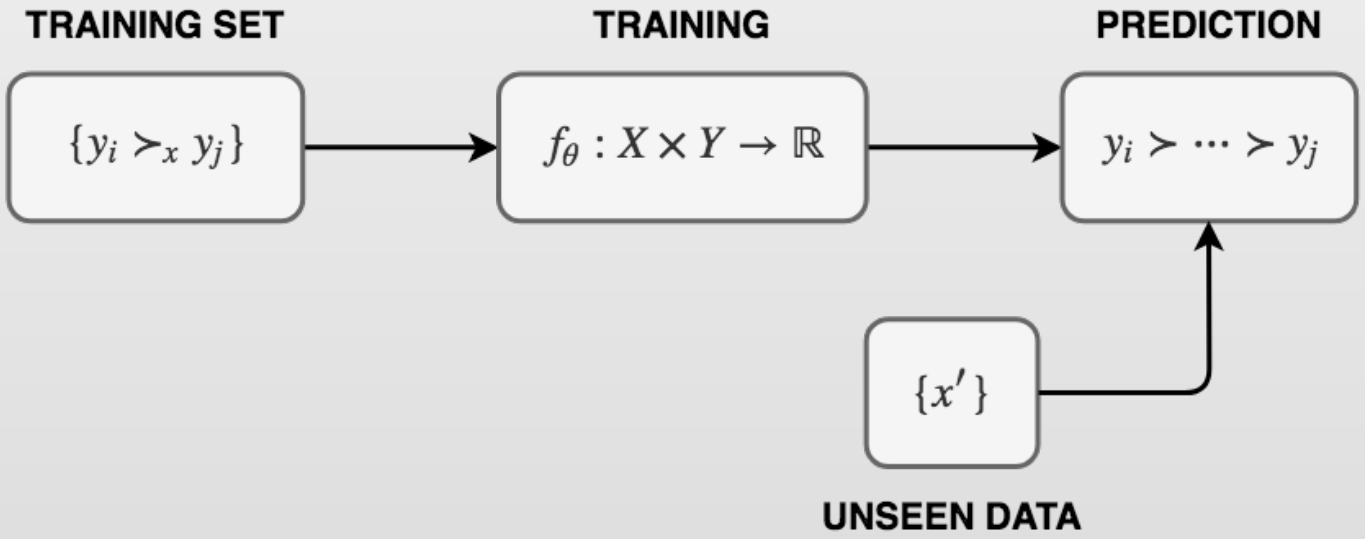

INTERPRETABILITA'
INTRINSECA
POST-HOC
RULE-BASED
"SIMPLE METHODS"
EXAMPLE-BASED
MODEL AGNOSTIC
MODEL SPECIFIC
ONE-R
SEQUENTIAL COV.
BAYESIAN RULE LE.
LINEARE
KNN
EXAMPLE-BASED
SURROGATI
PLOT-BASED
FEATURES-BASED
GA-BK-SVM
GA-SVM
OSRE
QSVM
LASSO
REGRESSIONE LIN.
REGRESSIONE LOG.
PRL
ALBERI DI DECISIONE
NAIVE BAYES
GLOBALI
LOCALI
PFI
LIME

Da classificazione a Preference Learning

- Ogni problema di classificazione può essere visto come un problema di **preference learning**



Preference Learning in PRL



PRL apprende una funzione di scoring **lineare** $f_\theta = \boldsymbol{w}^\top \boldsymbol{z}$

Come apprende PRL

PRL impara l'ipotesi che **massimizza il margine** tra preferenze

$$\boldsymbol{w} \propto \sum_j \alpha_j \boldsymbol{z}_j \longrightarrow \rho(\boldsymbol{z}) = \sum_j \alpha_j \boldsymbol{z}_j^\top \boldsymbol{z}$$

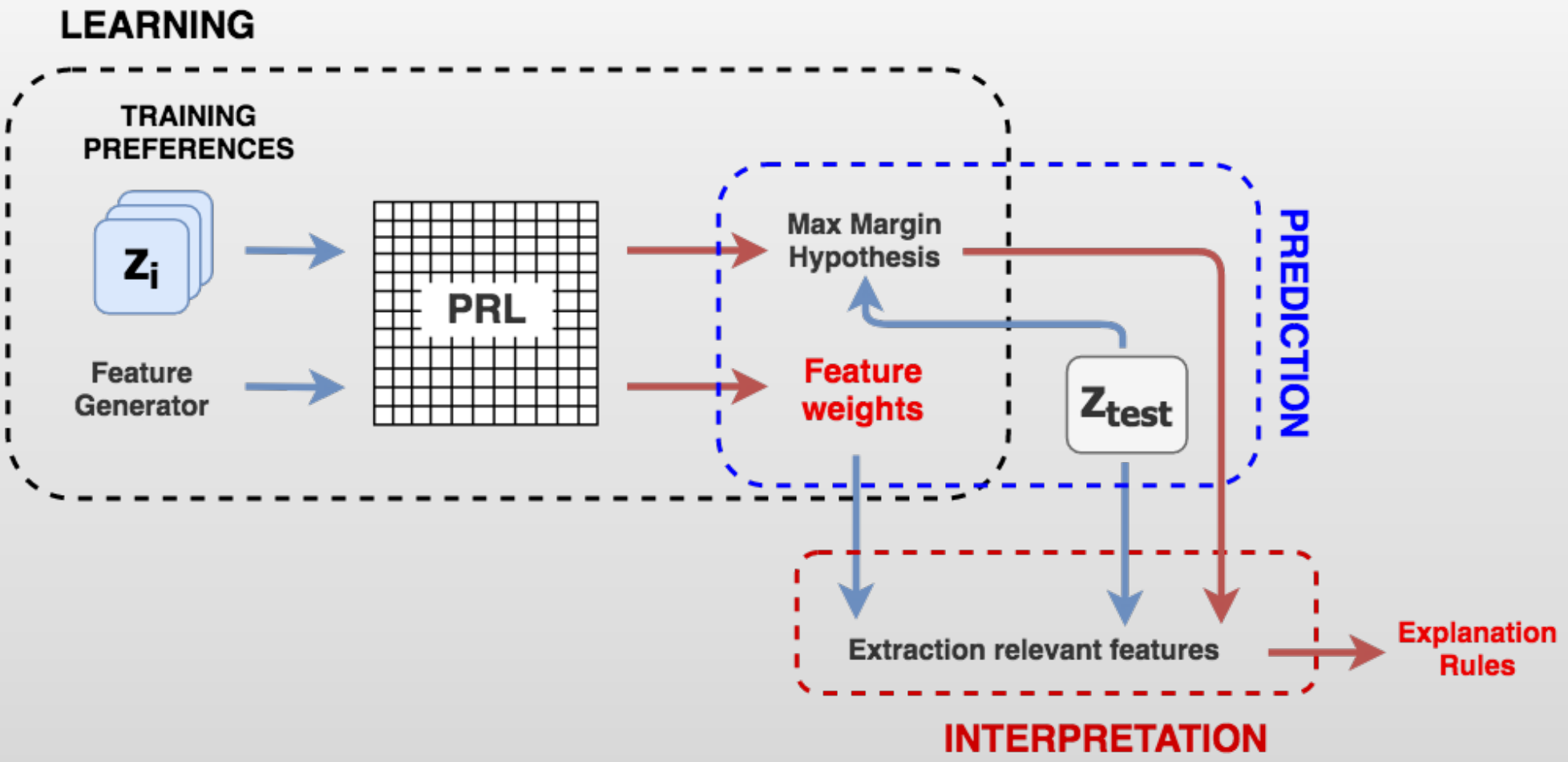
margin

Per scopi di feature selection PRL introduce una seconda **distribuzione $\boldsymbol{\mu}$** (da apprendere) **sulle feature**

$$\rho(\boldsymbol{z}) = \sum_j \alpha_j \sum_f \mu_f \, \boldsymbol{z}_j[f]^\top \boldsymbol{z}[f] = \sum_{j,f} q_{j,f} \, \boldsymbol{z}_j[f]^\top \boldsymbol{z}[\boldsymbol{f}]$$

PRL apprende la distribuzione **q** usando la teoria dei giochi.

Interpretabilità con PRL



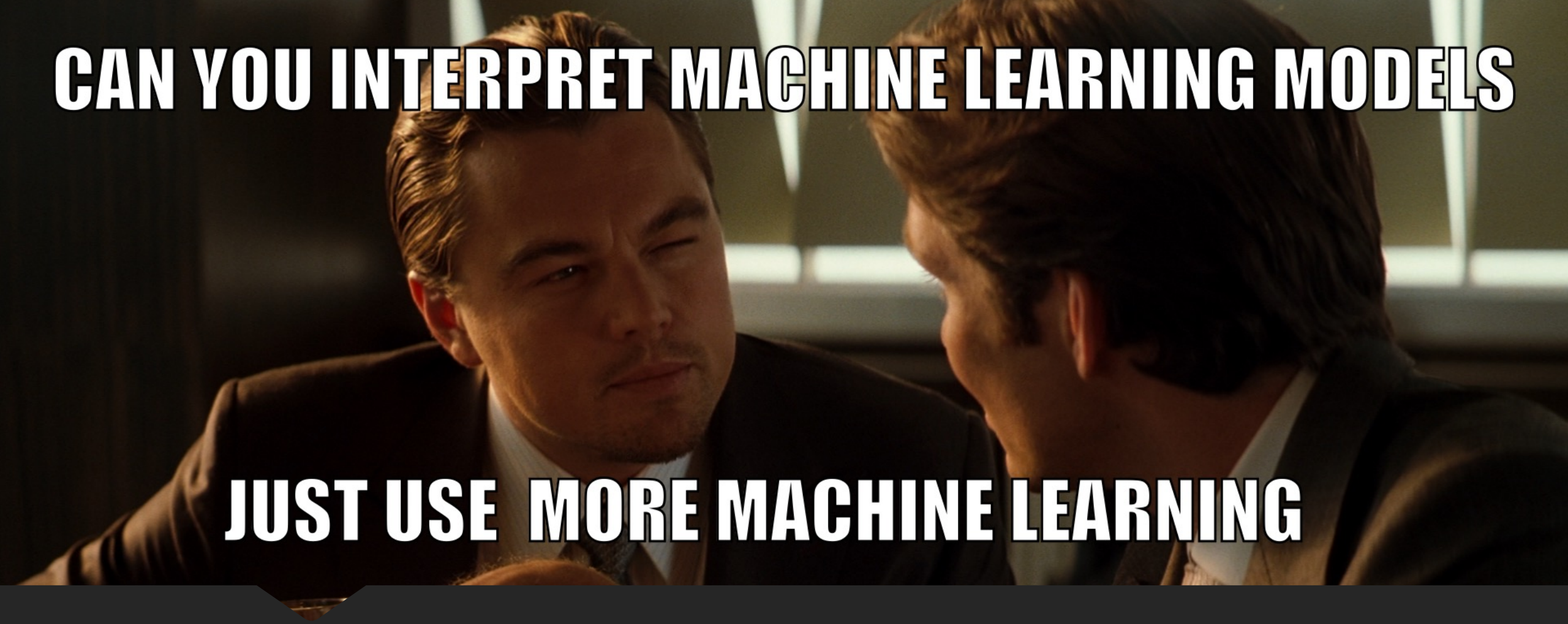
PRL: pro e contro

PROS

- **Garanzie teoriche**
- Feature selection nel feature space
- Ideale per problemi con moltissime feature

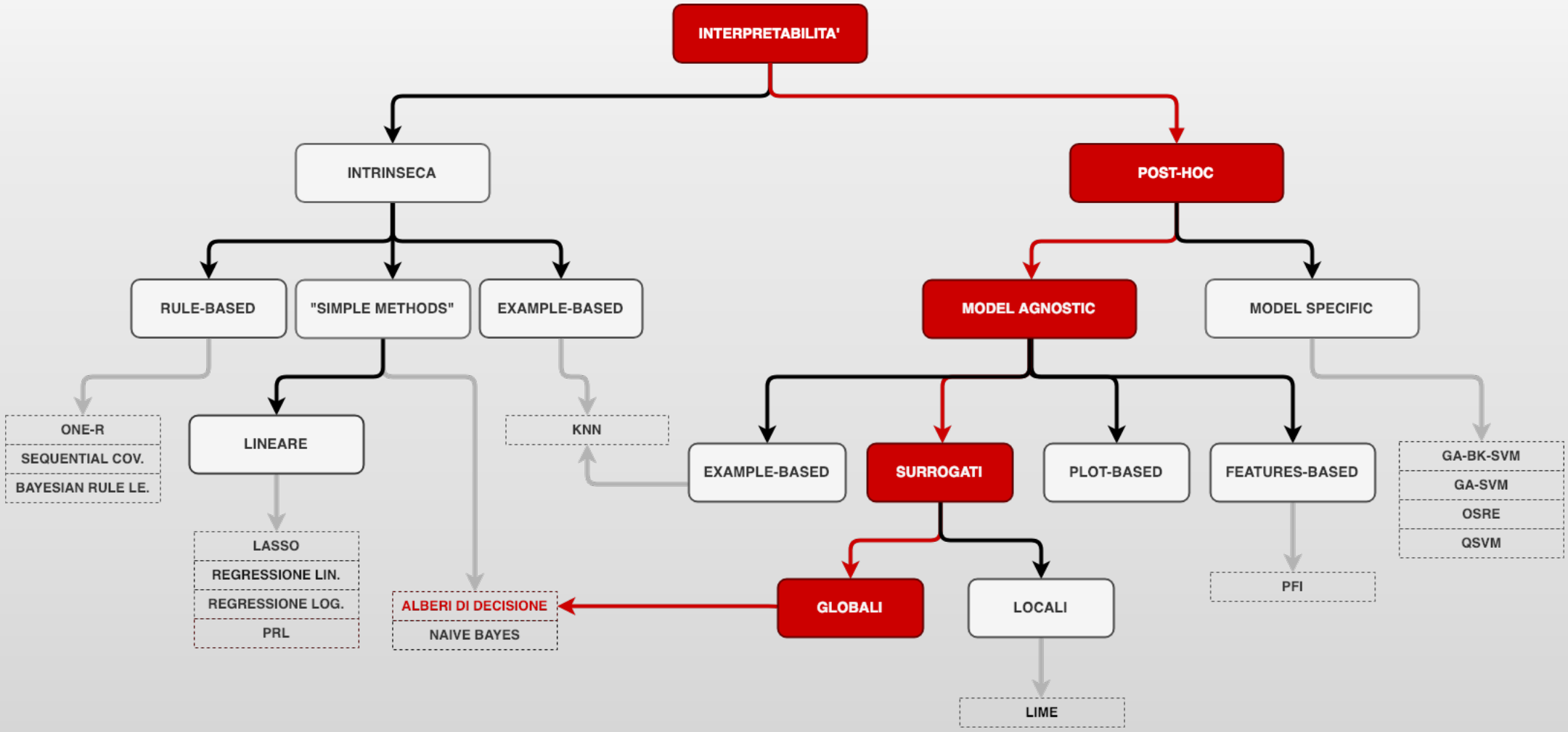
CONS

- Nella pratica, **non particolarmente efficiente**
- Con molti esempi il **numero di preferenze** può diventare **intrattabile**
- Se il numero di feature selezionate è elevato diventa difficile l'interpretazione
- L'interpretabilità dipende da che feature si utilizzano

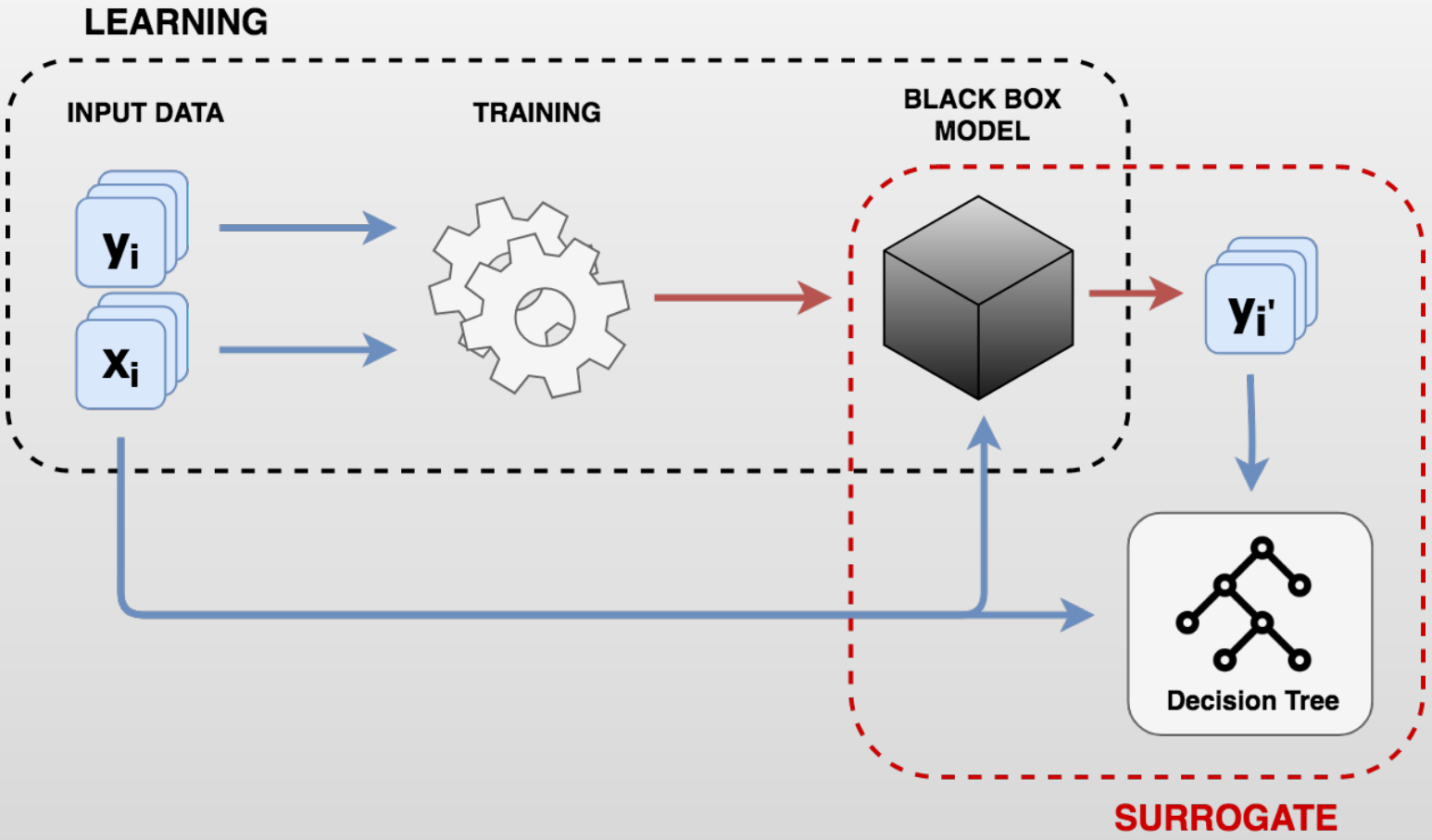


Metodi surrogati

Surrogati: Alberi di decisione



Surrogati: Alberi di decisione



EVALUATION

Valutare quanto il modello surrogato rispecchia le predizioni del modello black box

$$R^2 = 1 - \frac{\sum_i (y^* - y')^2}{\sum_i (y' - \bar{y}')^2}$$

- y^* output surrogato
- y' output black box
- $\bar{y}'$ output medio BB

Surrogati: pro e contro

PROS

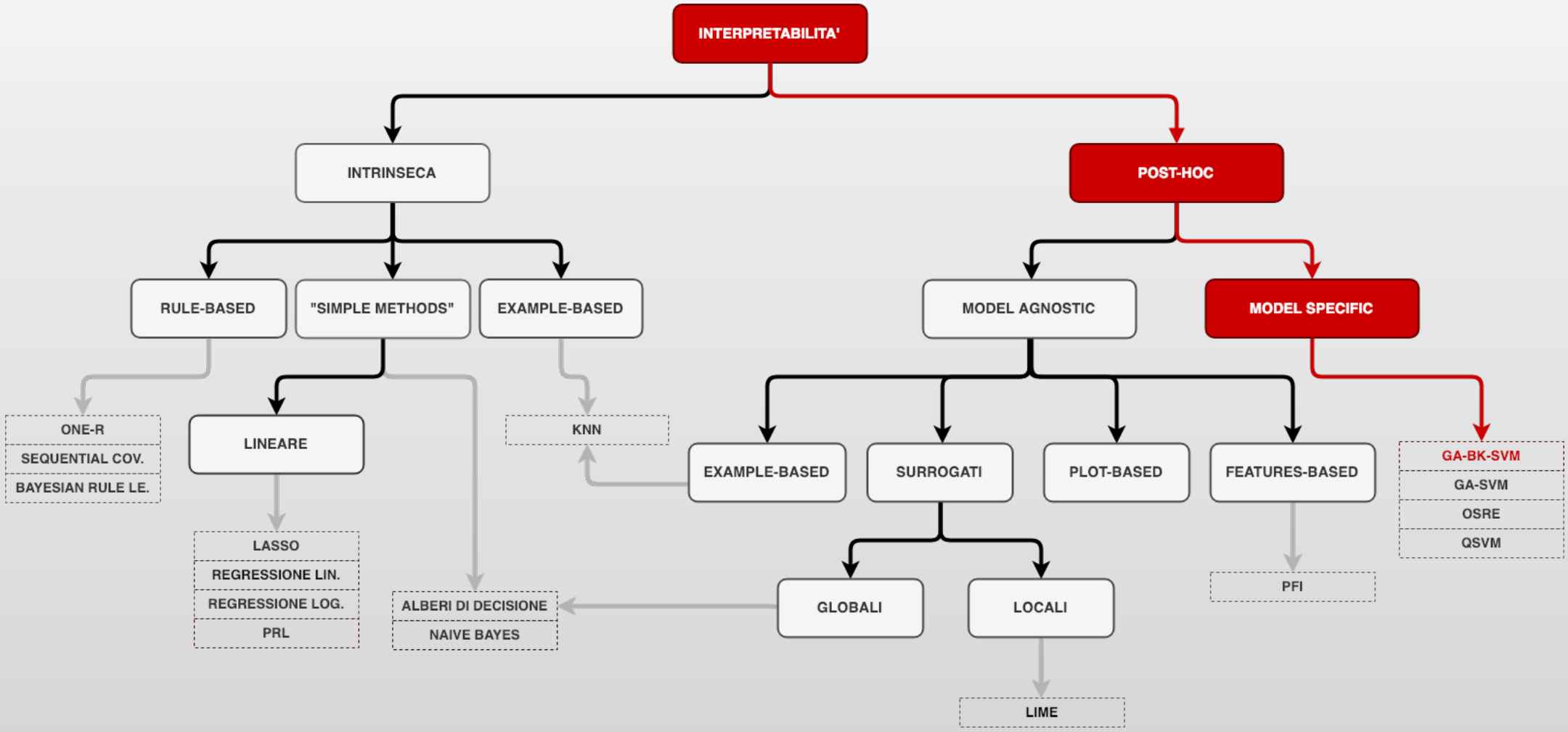
- **Flessibile**
- Approccio **intuitivo** e facile da applicare
- Alta (massima) **portabilità**

CONS

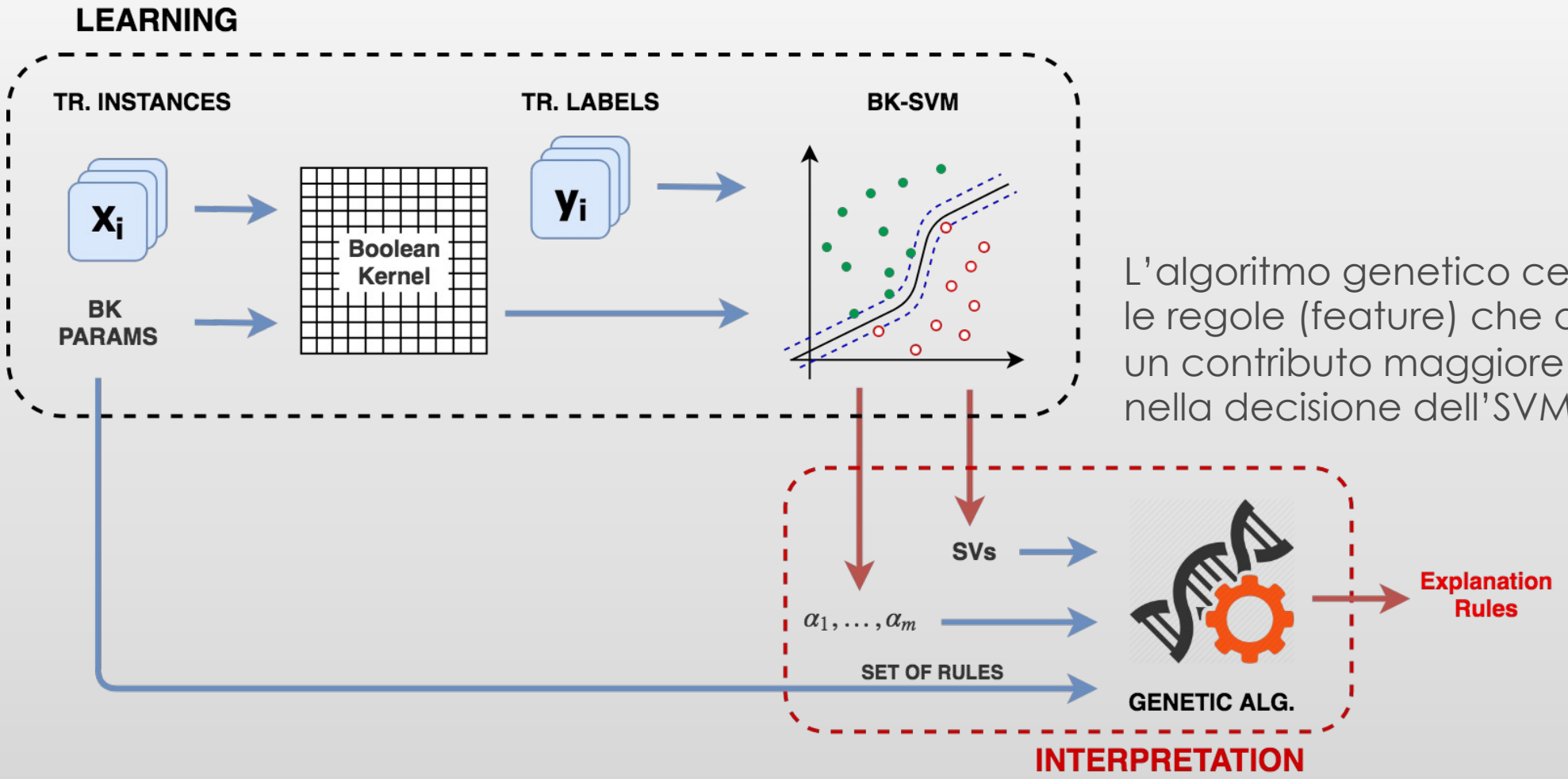
- Difficile stimare la bontà del surrogato
- Il surrogate stesso ha i suoi vantaggi/svantaggi
- Si possono trarre **conclusioni solo sul modello** e non sui dati
- La spiegazione del surrogato potrebbe non essere in linea con quella del modello di partenza

GA-BK-SVM

GA-BK-SVM


INTERPRETABILITA'
INTRINSECA
POST-HOC
RULE-BASED
"SIMPLE METHODS"
EXAMPLE-BASED
MODEL AGNOSTIC
MODEL SPECIFIC
ONE-R
SEQUENTIAL COV.
BAYESIAN RULE LE.
LINEARE
KNN
EXAMPLE-BASED
SURROGATI
PLOT-BASED
FEATURES-BASED
GA-BK-SVM
GA-SVM
OSRE
QSVM
LASSO
REGRESSIONE LIN.
REGRESSIONE LOG.
PRL
ALBERI DI DECISIONE
NAIVE BAYES
GLOBALI
LOCALI
PFI
LIME

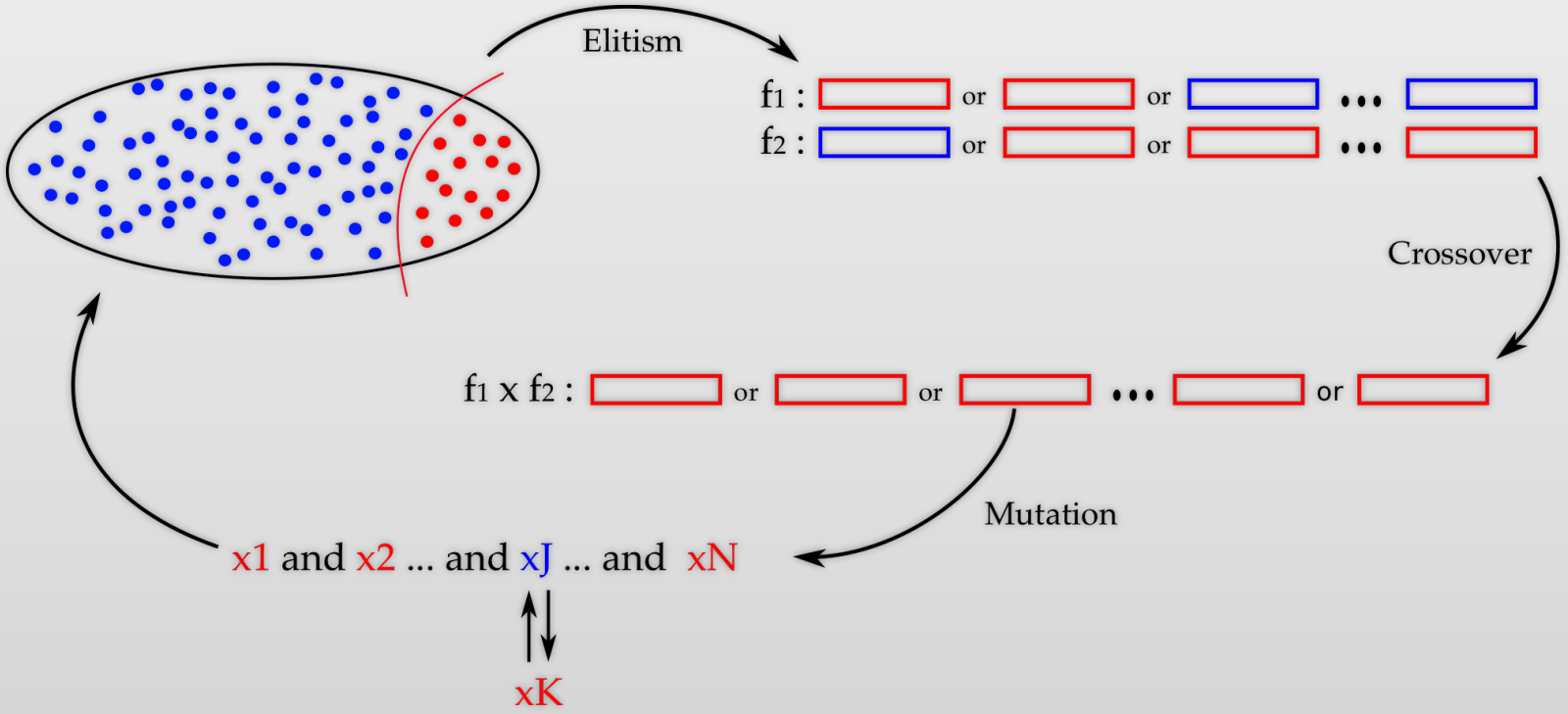
GA-BK-SVM



L'algoritmo genetico cerca le regole (feature) che danno un contributo maggiore nella decisione dell'SVM

Algoritmo genetico nel "dettaglio"

- Selezione **elitista**
- Gene ➜ mDNF formula
- **Crossover**: modifica la disgiunzione
- **Mutazione**: modifica una congiunzione



GA-BK-SVM: pro e contro

PROS

- Sfrutta un metodo allo stato dell'arte (SVM)
- **Teoricamente ben fondato**
- Empiricamente mostra un **ottima fedeltà** all'SVM

CONS

- **L'algoritmo genetico** non dà garanzie
- "By design" l'insieme di formule migliori è un intorno della *best rule*
- La *best rule* **potrebbe non essere facilmente interpretabile**
- In caso di feature continue, dipende molto da come si discretizza

Take away messages

Take away messages

L'interpretabilità è una caratteristica spesso desiderabile

Ottenerla può significare dover rinunciare ad un po' di accuracy

Tema caldo nel ML perché legato a molte tematiche importanti

Creare metodi allo stato dell'arte "interpretabili" è difficile ma possible!

Bibliografia

Molnar (2018) Christoph Molnar, *"Interpretable Machine Learning – A guide for making black box models explainable"*. Leanpub book.

Miller (2017) Tim Miller, *"Explanation in Artifical Intelligence: Insight from the Social Sciences"*. Arxiv Preprint arXiv:1706.07269

Polato (2018a) Mirko Polato e Fabio Aiolli, *"Boolean kernels for rule based interpretation of Support Vector Machines"*. Neurocomputing (Elsevier) in press.

Polato (2018b) Mirko Polato e Fabio Aiolli. *"Boolean kernels for interpretable kernel machines"*. ESANN 2018.

Polato (2019) Mirko Polato e Fabio Aiolli, *"Interpretable preference learning: a game theoretic framework for large margin on-line feature and rule learning"*. AAAI 2019.

Articoli correlati

F. Poursabzi-Sangdesh et al. *"Manipulating and Measuring Model Interpretability"*. Arxiv Preprint arXiv: 1802.07810 (2018)

F. Doshi.Velezet al. *"Towards A Rigorous Science of Interpretable Machine Learning"*. Arxiv Preprint arXiv: 1702.08608 (2017)

Z. C. Lipton. *"The Mythos of Model Interpretability"*. Arxiv Preprint arXiv: 1606.03490 (2016)

Mirko Polato, PhD

E-mail: mpolato@math.unipd.it

Sito: http://www.math.unipd.ith/~mpolato