O QUE VOCÊ PRECISA SABER SOBRE AUTOML

Luciana Lima

Luciana Lima

Head de Analytics na A3Data

Formação

Estatística

Pós-graduação em Business Intelligence

Interesses

MLOps

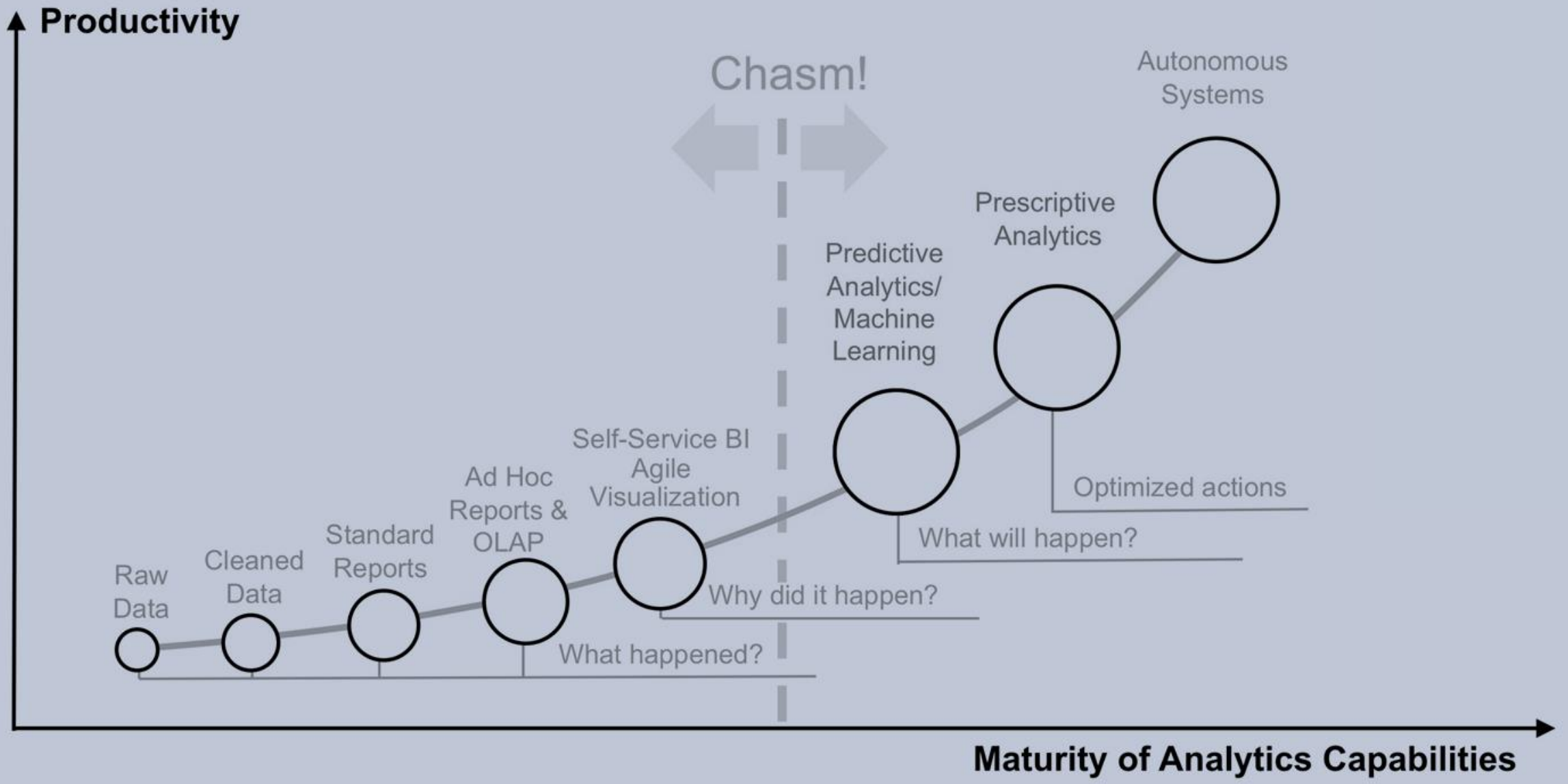
AutoML

Gatos

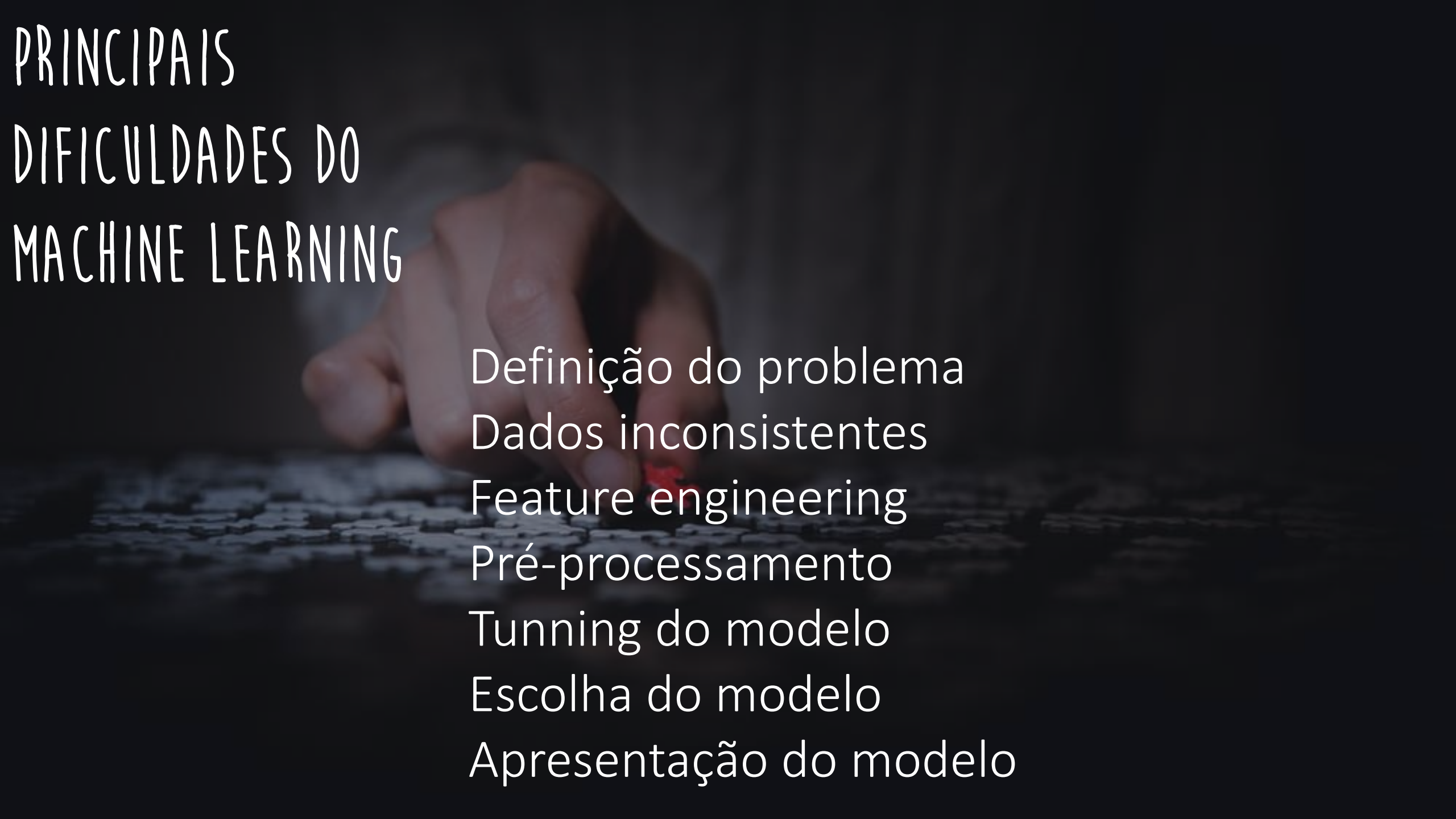
Séries

Trekking

Viagem


Productivity
Chasm!
Autonomous Systems
Prescriptive Analytics
Predictive Analytics/ Machine Learning
Self-Service BI Agile Visualization
Ad Hoc Reports & OLAP
Standard Reports
Cleaned Data
Raw Data
Optimized actions
What will happen?
Why did it happen?
What happened?
Maturity of Analytics Capabilities


Entendimento do problema
Coleta e preparação das bases de dados
Análise exploratória dos dados
Modelagem
Validação
Implantação
Avaliação periódica



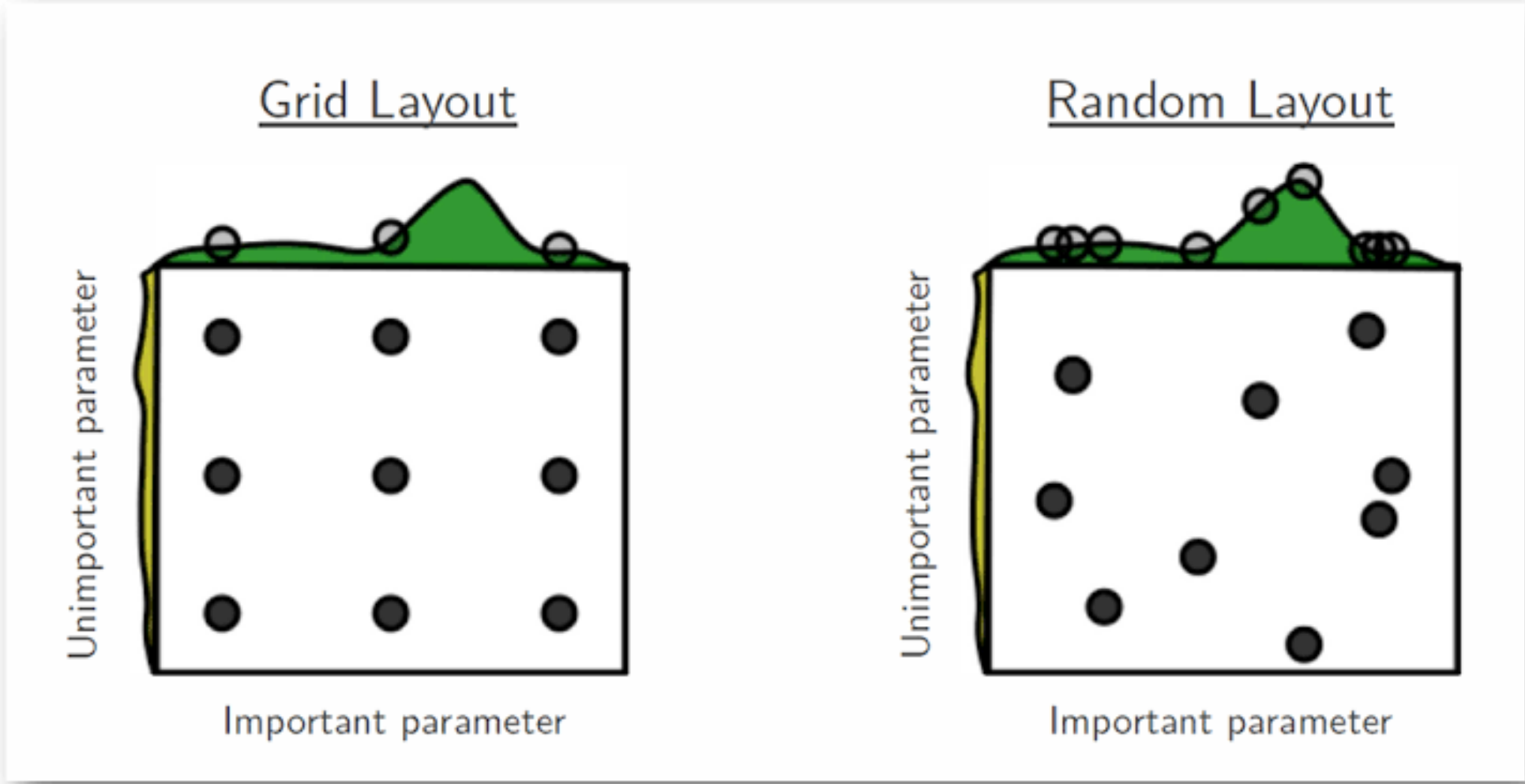
PRINCIPAIS DIFICULDADES DO MACHINE LEARNING

Definição do problema
Dados inconsistentes
Feature engineering
Pré-processamento
Tunning do modelo
Escolha do modelo
Apresentação do modelo

O que é Tunning do Modelo?

class sklearn.linear_model.**LogisticRegression**(*penalty='l2', dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None*)

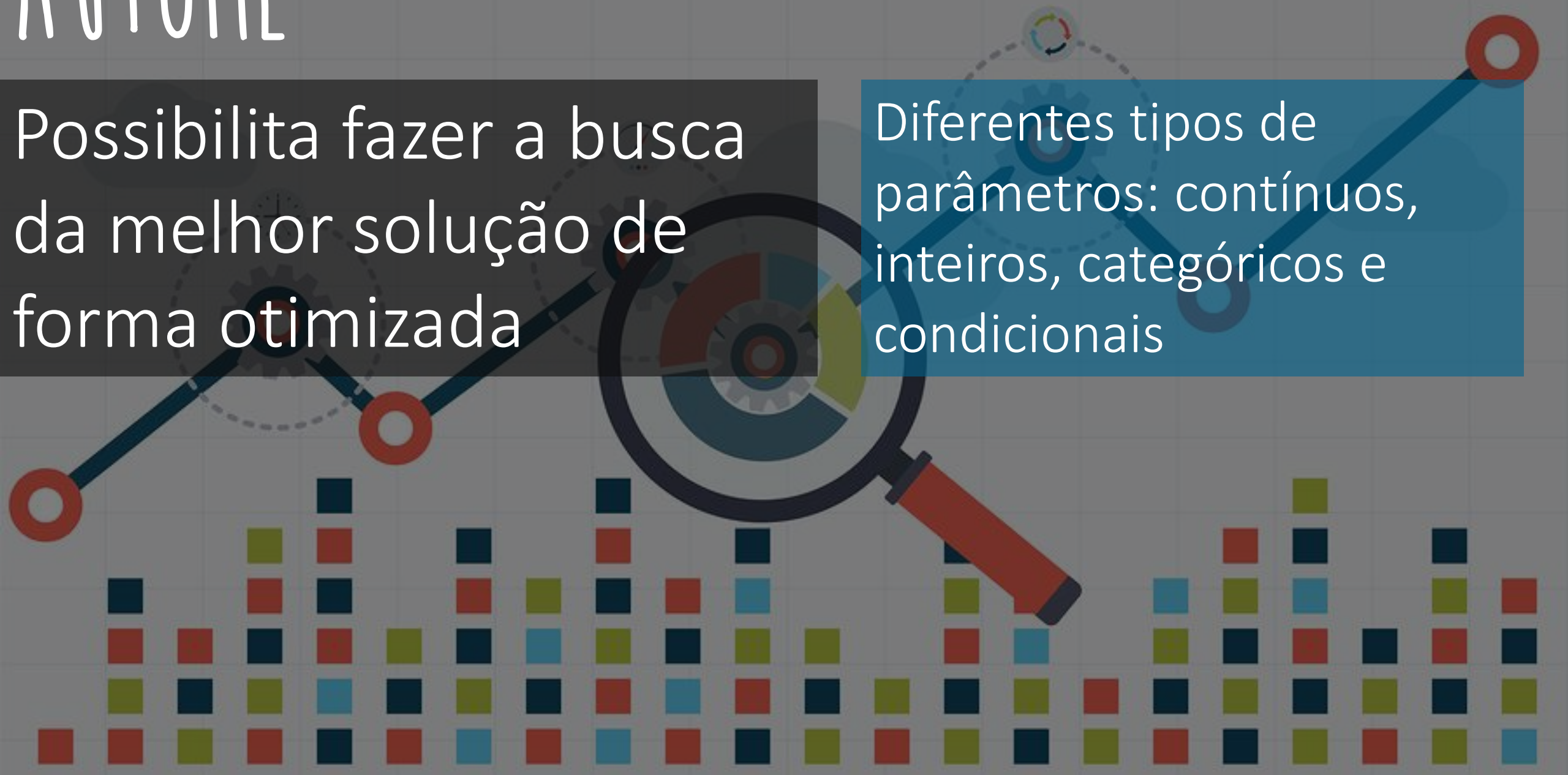
*class sklearn.ensemble.**RandomForestClassifier**(n_estimators=100, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None)*


Grid Layout
Unimportant parameter
Important parameter
Random Layout
Unimportant parameter
Important parameter

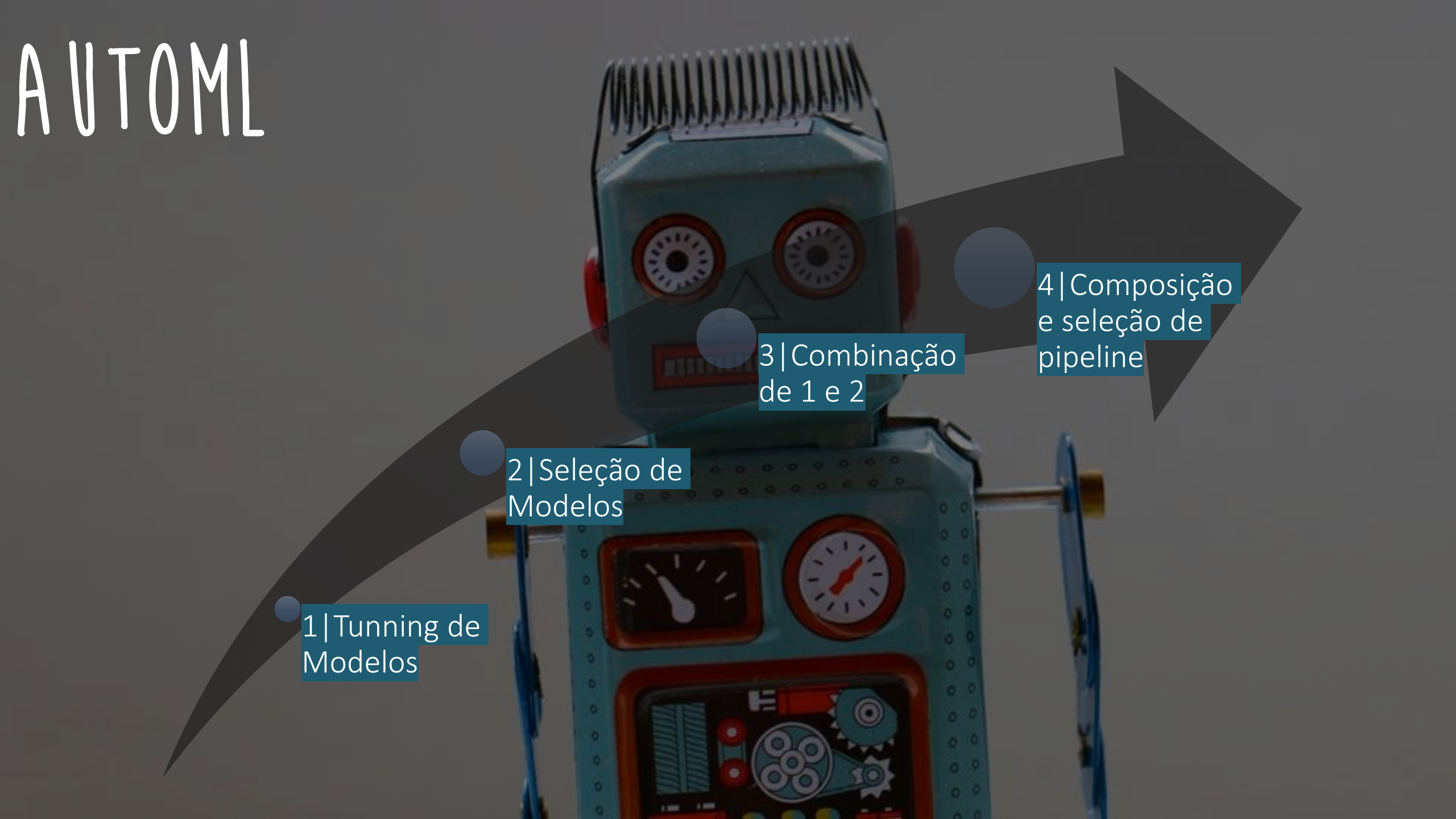
AUTOML

Possibilita fazer a busca da melhor solução de forma otimizada

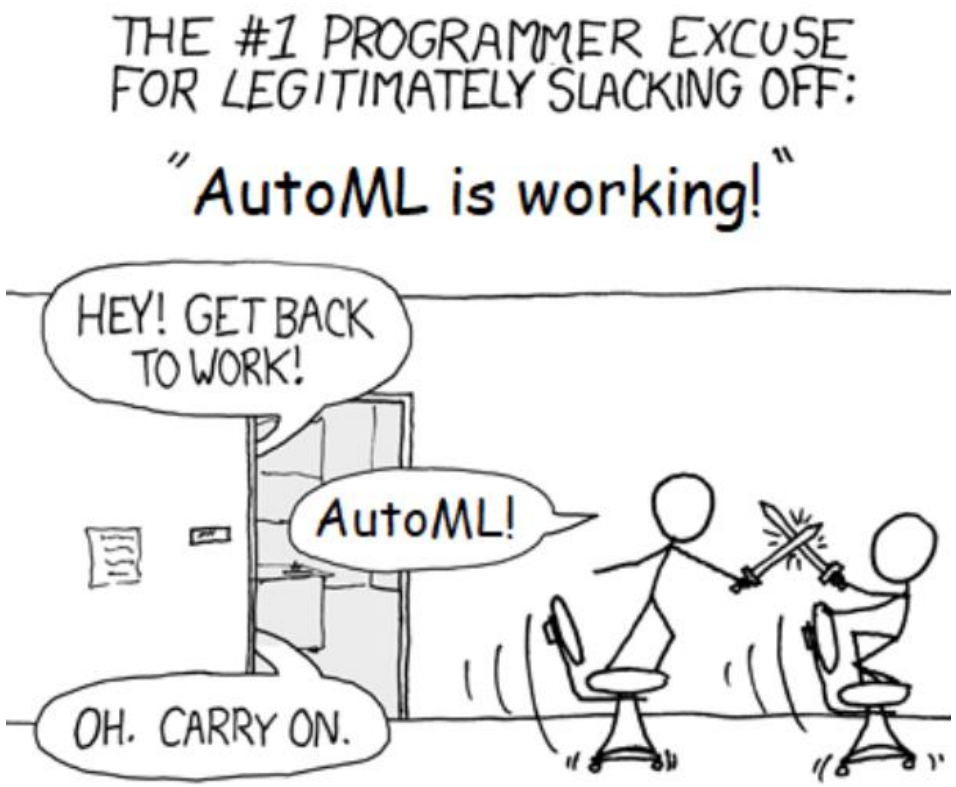
Diferentes tipos de parâmetros: contínuos, inteiros, categóricos e condicionais



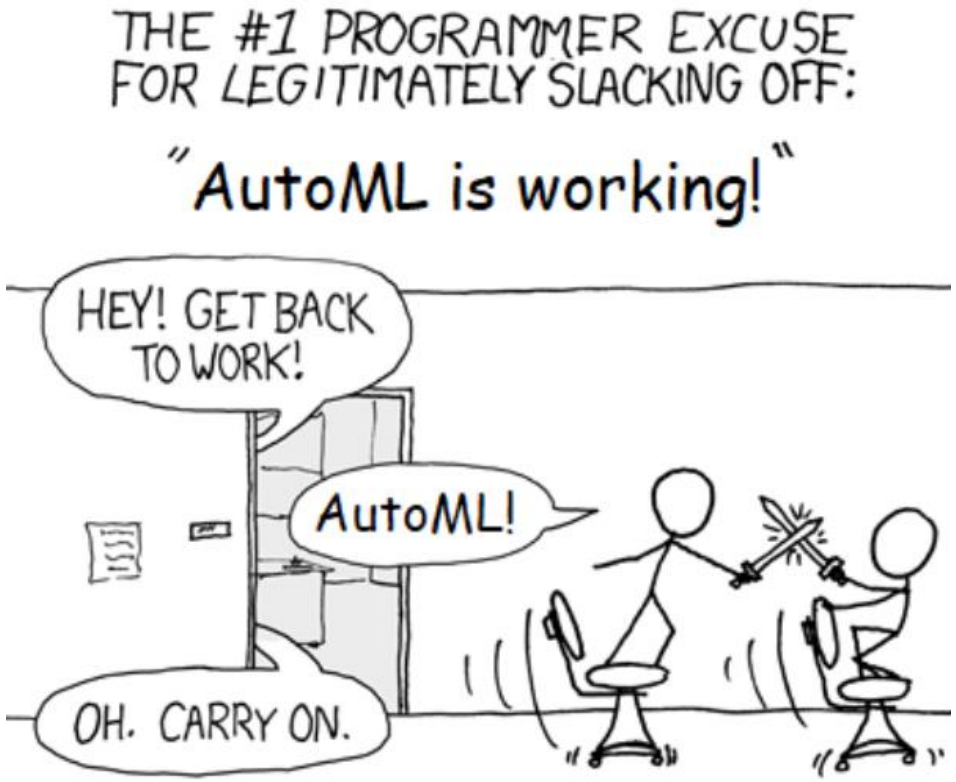
AUTOML


1|Tunning de Modelos
2|Seleção de Modelos
3|Combinação de 1 e 2
4|Composição e seleção de pipeline

ENTÃO NÃO PRECISA MAIS DE CIENTISTA DE DADOS?



ENTÃO NÃO PRECISA MAIS DE CIENTISTA DE DADOS?



Definição do problema

Dados inconsistentes

Feature engineering

Pré-processamento

Tunning do modelo

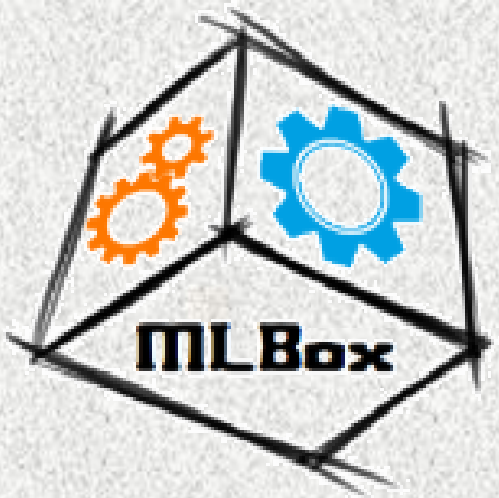
Escolha do modelo

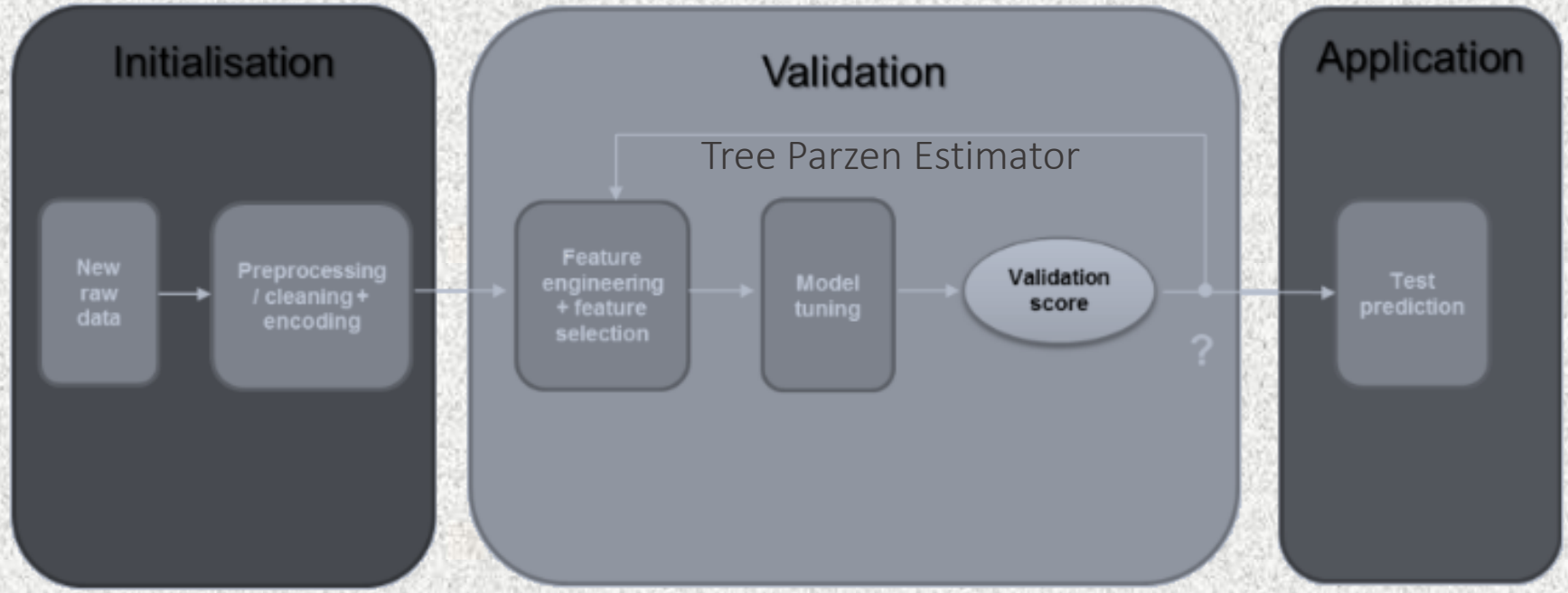
Apresentação do modelo

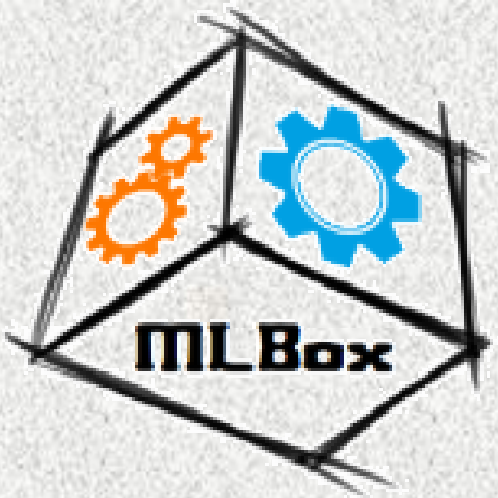
PRINCIPAIS FRAMEWORKS

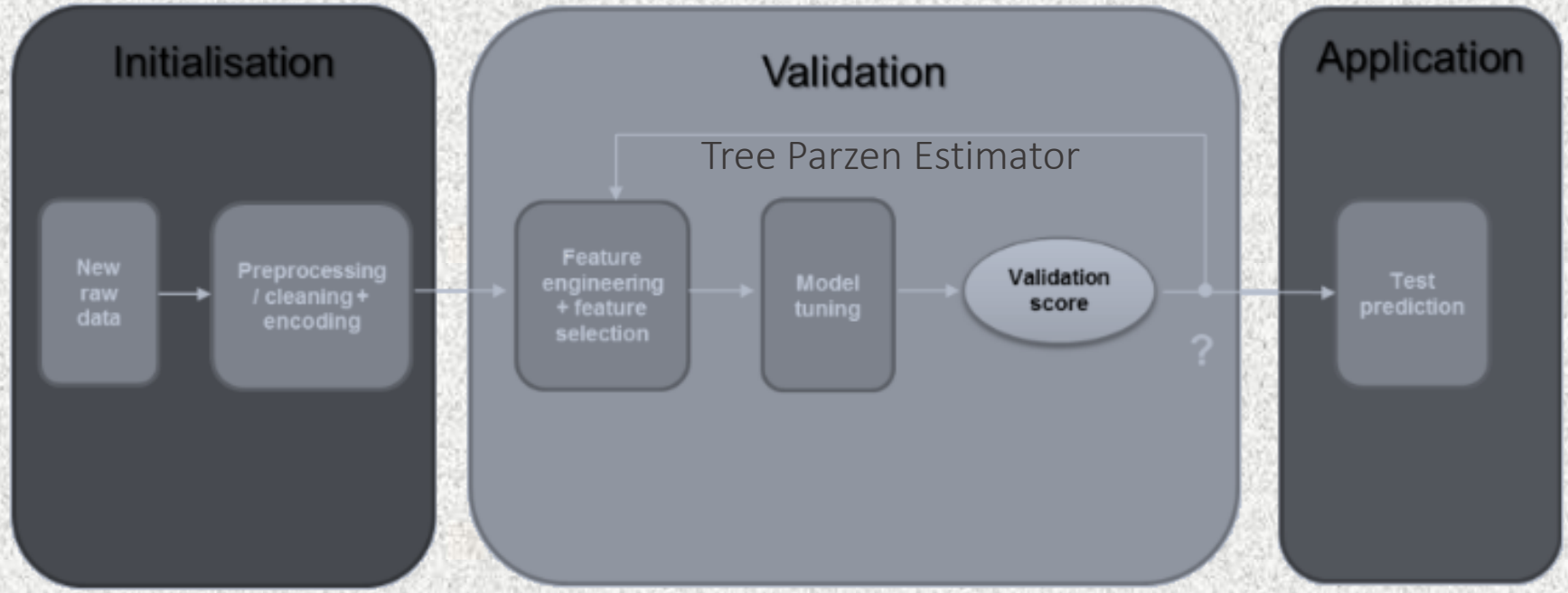
Auto-Sklearn

H_2O.ai


MLBox

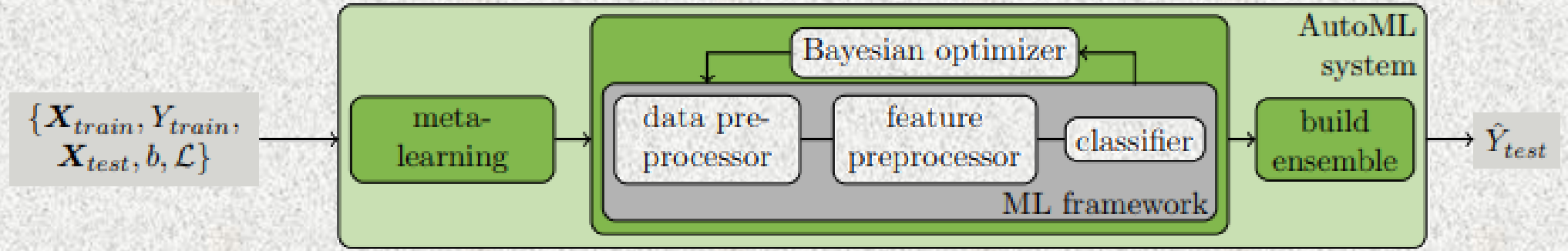

Initialisation
New raw data
Preprocessing / cleaning + encoding
Validation
Tree Parzen Estimator
Feature engineering + feature selection
Model tuning
Validation score
?
Application
Test prediction



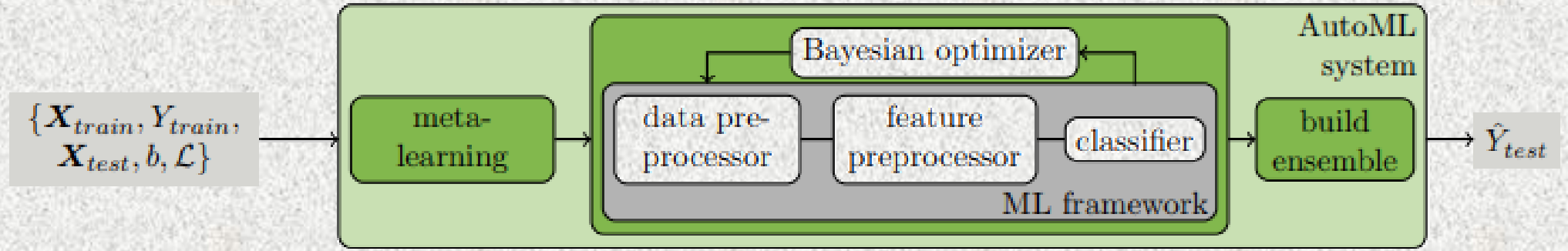


Processamento paralelizado	Restrição do espaço de busca	Dados categóricos	Stacking de modelos
Modelo com interpretabilidade	Restrição de tempo	Dados faltantes	Métricas parametrizadas

Auto-Sklearn

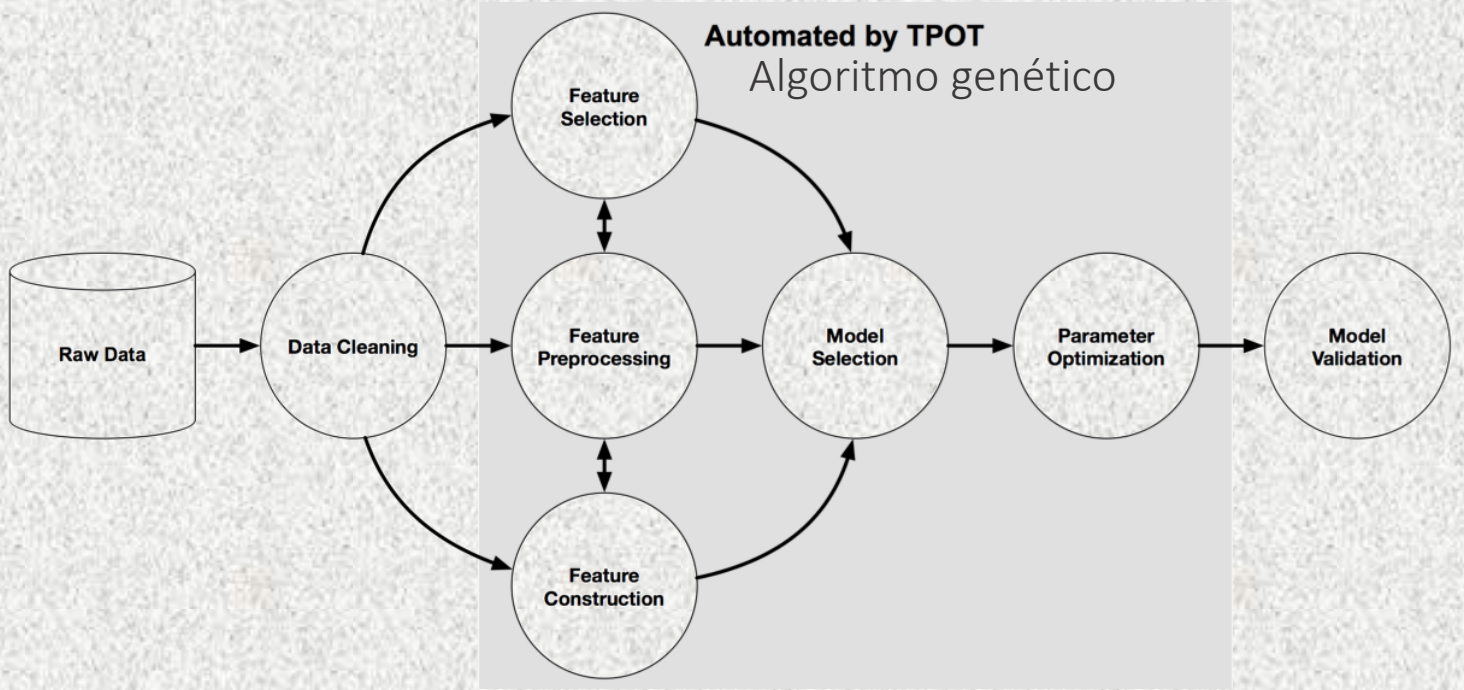


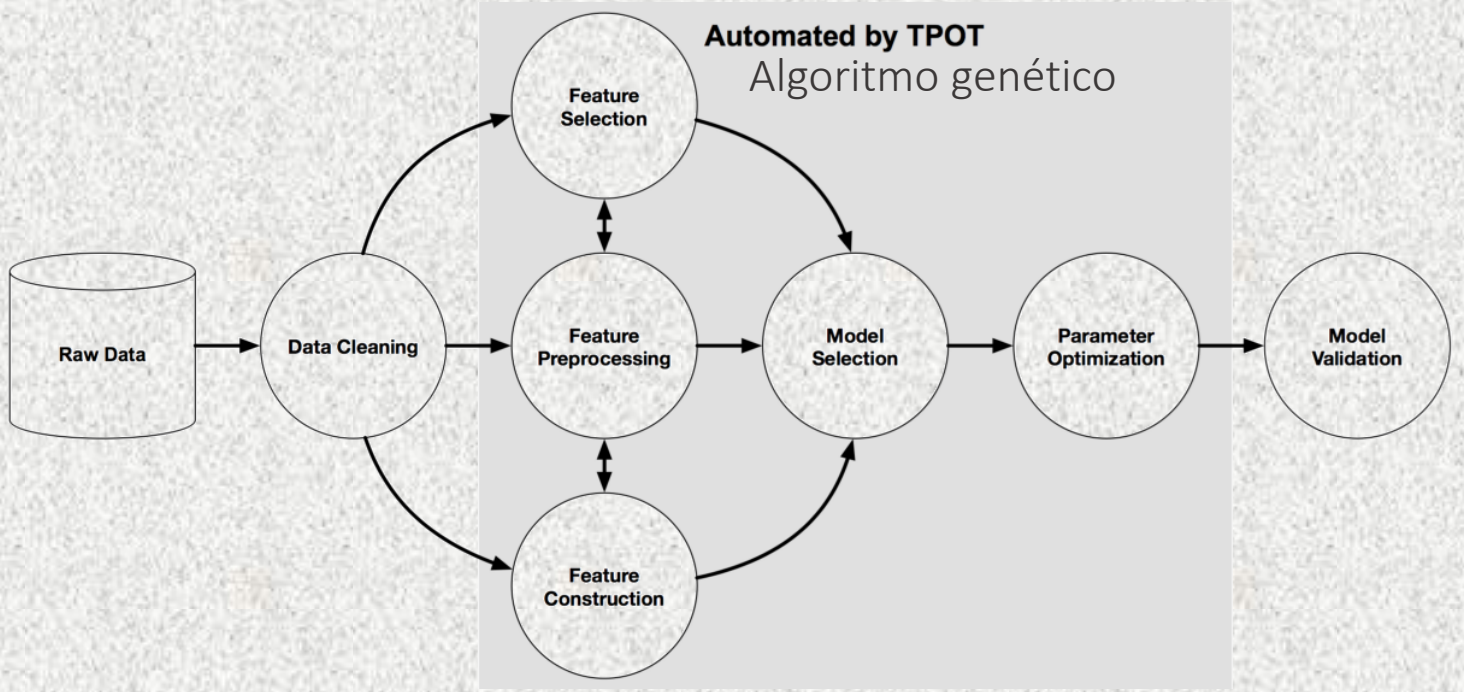
Auto-Sklearn



Processamento paralelizado	Restrição do espaço de busca	Dados categóricos	Stacking de modelos
Modelo com interpretabilidade	Restrição de tempo	Dados faltantes	Métricas parametrizadas

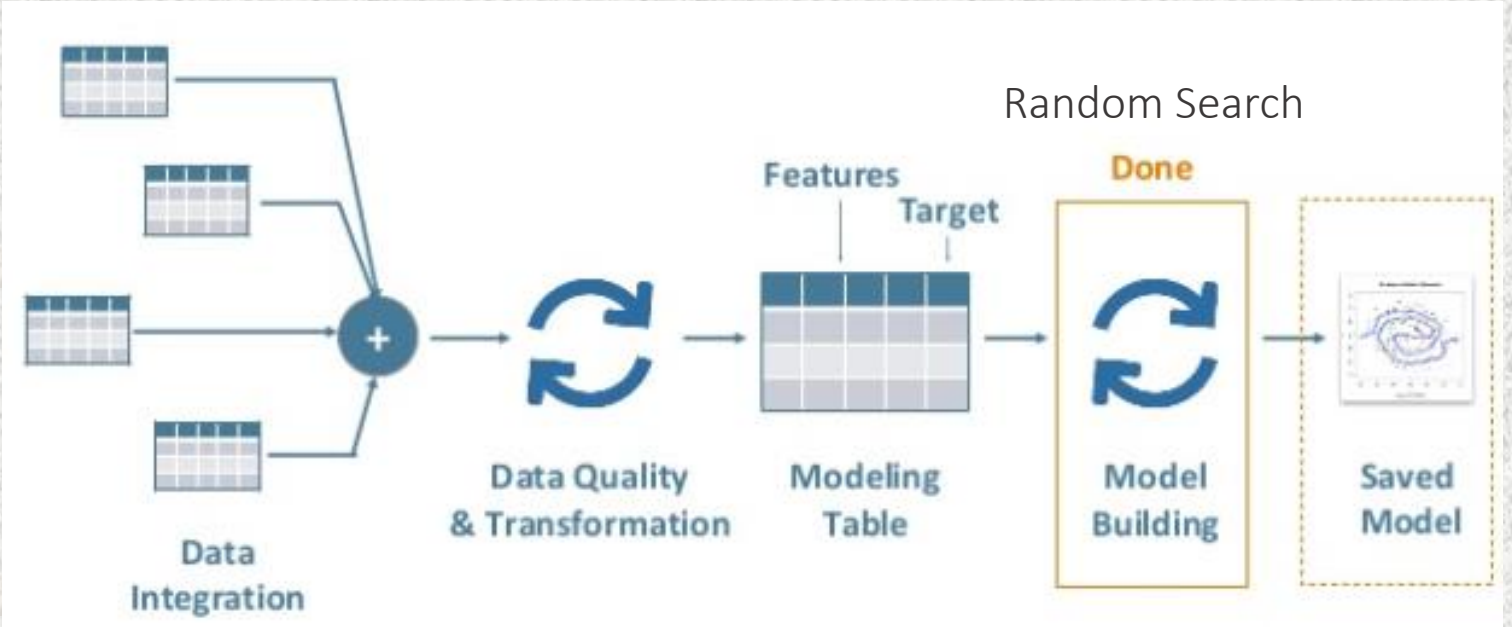

y = a + bX + e
TPOT


Automated by TPOT
Algoritmo genético
Feature Selection
Raw Data
Data Cleaning
Feature Preprocessing
Model Selection
Parameter Optimization
Model Validation
Feature Construction

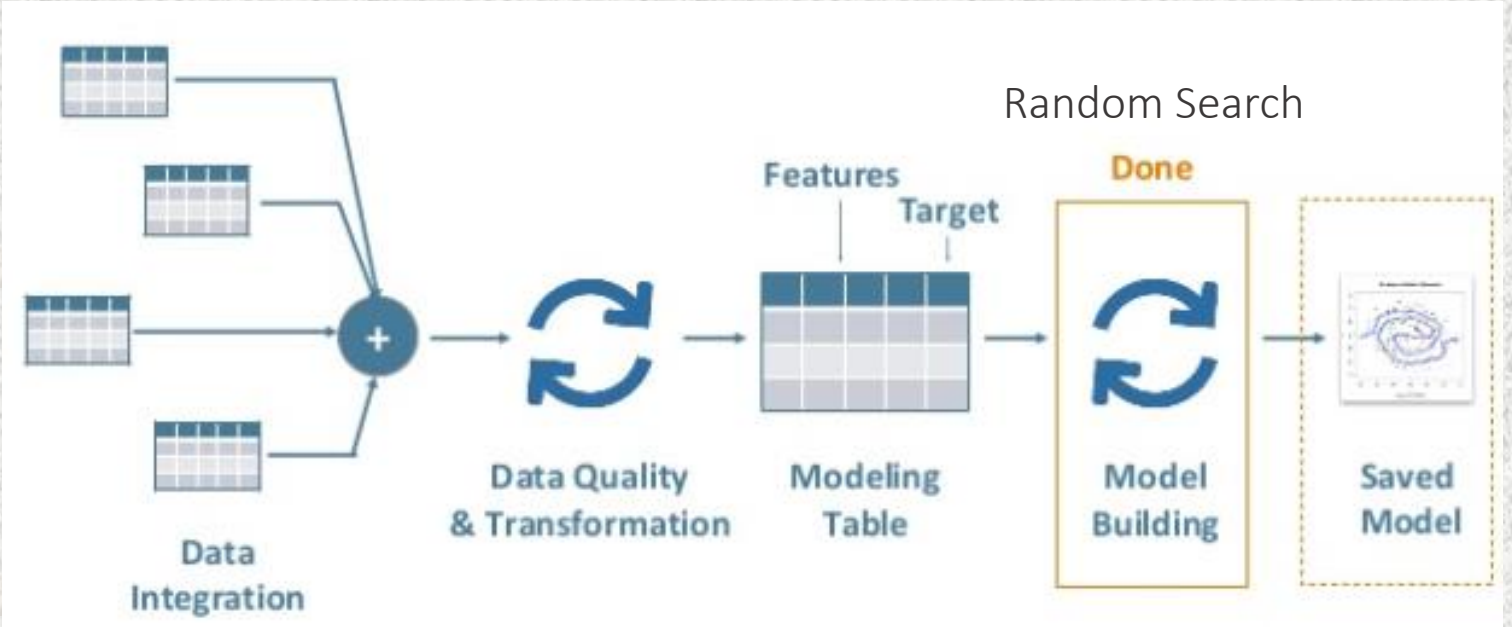


Processamento paralelizado	Restrição do espaço de busca	Dados categóricos	Stacking de modelos
Modelo com interpretabilidade	Restrição de tempo	Dados faltantes	Métricas parametrizadas

H2O.ai

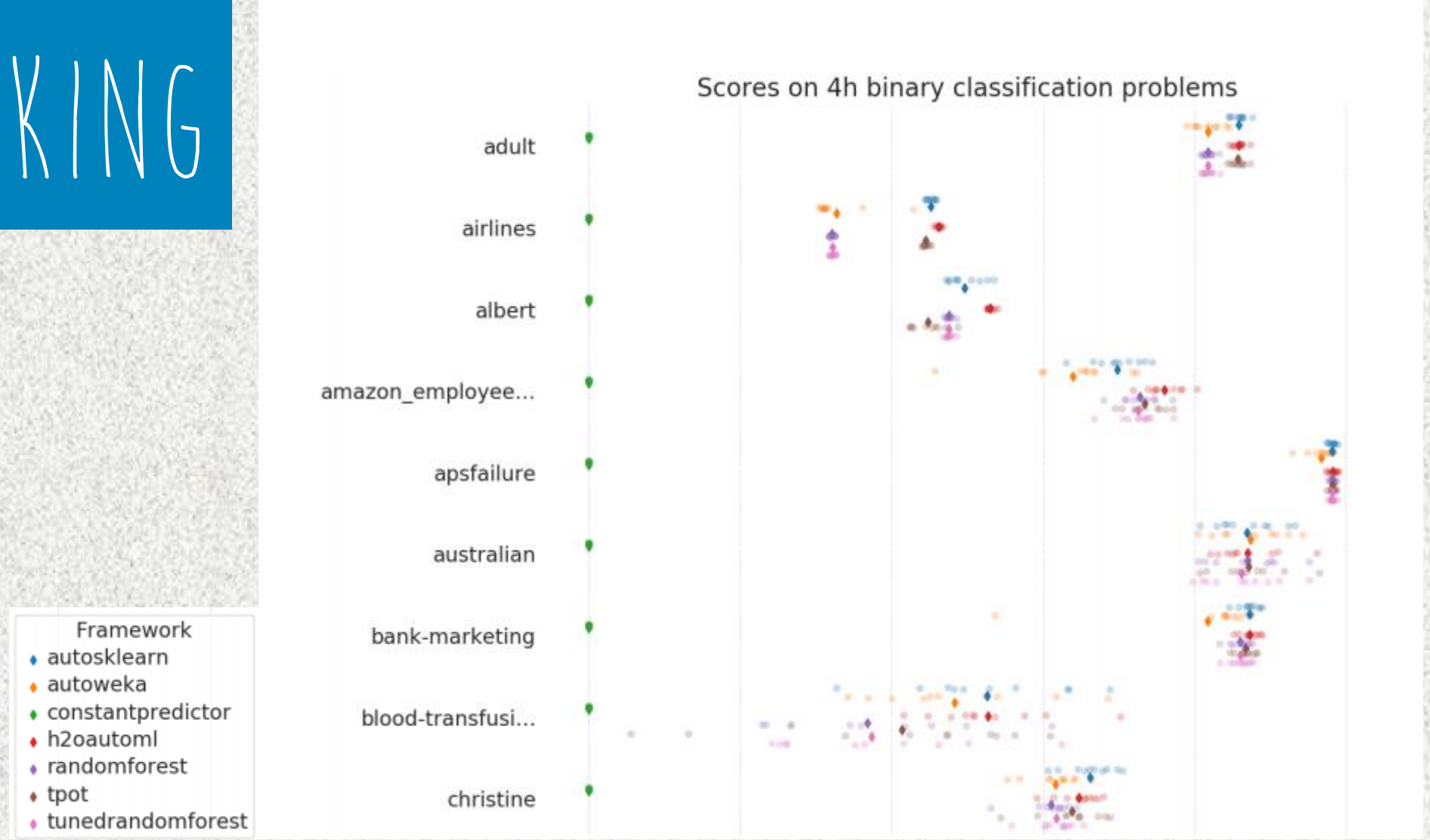

Random Search
Features
Target
Done
Data Integration
Data Quality & Transformation
Modeling Table
Model Building
Saved Model

H_2O.ai



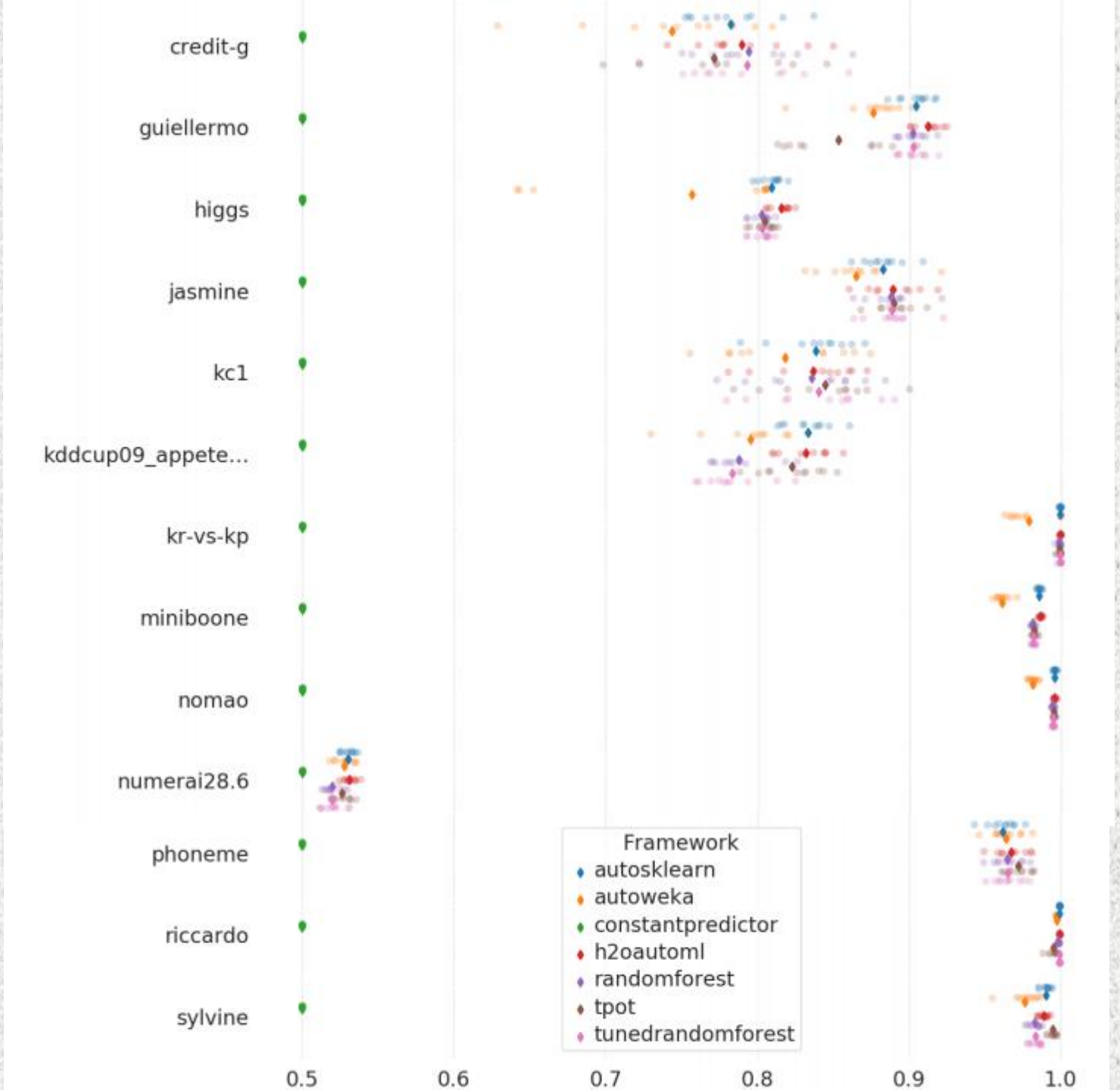
Processamento paralelizado	Restrição do espaço de busca	Dados categóricos	Stacking de modelos
Modelo com interpretabilidade	Restrição de tempo	Dados faltantes	Métricas parametrizadas

BENCHMARKING



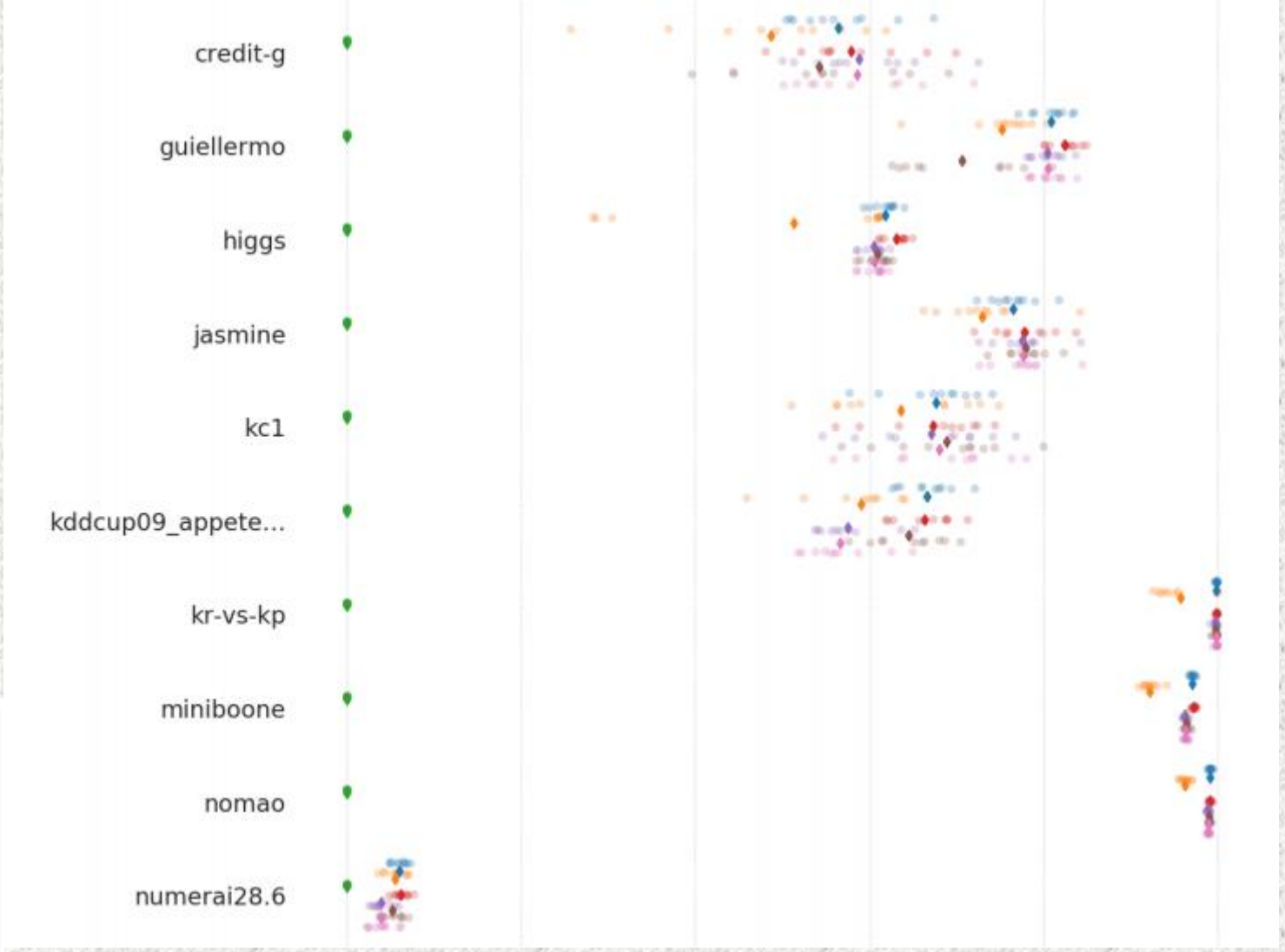
https://www.automl.org/wp-content/uploads/2019/06/automlws2019_Paper45.pdf

BENCHMARKING



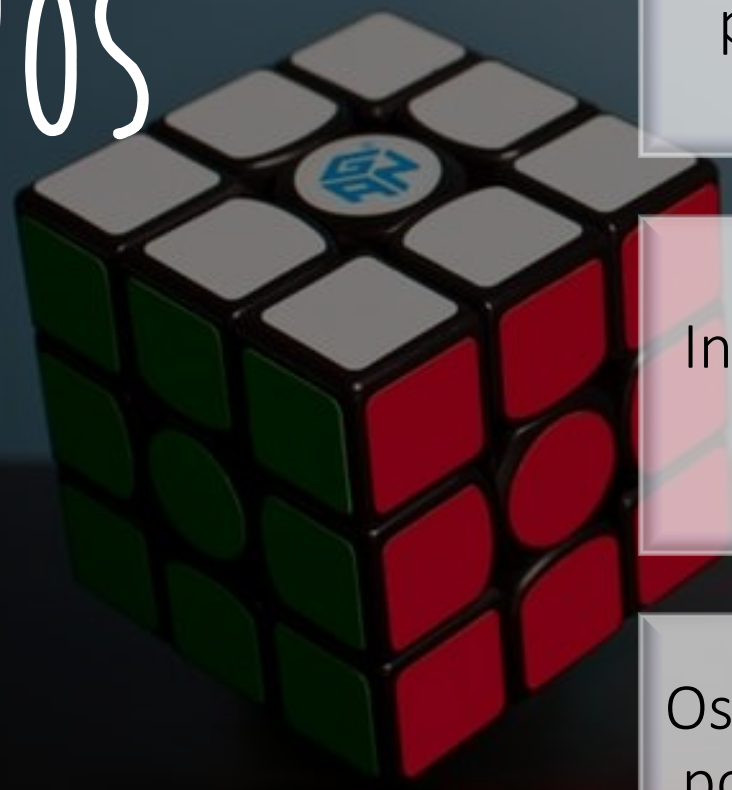
https://www.automl.org/wp-content/uploads/2019/06/automlws2019_Paper45.pdf

BENCHMARKING



https://www.automl.org/wp-content/uploads/2019/06/automlws2019_Paper45.pdf

AUTOML: DESAFIOS



- Tempo de processamento
- Convergência dos resultados
- Interpretabilidade dos modelos
- Não é multi-objetivo
- Os frameworks não possuem todos os algoritmos
- Os frameworks não lidam com texto e imagens

"Some people call this artificial intelligence, but the reality is this technology will enhance us. So instead of artificial intelligence, I think we'll augment our intelligence." Ginni Rometty – CEO IBM

OBRIGADA!

luciana-lima-data-science