What is explainable AI?

Interpretable
Machine Learning

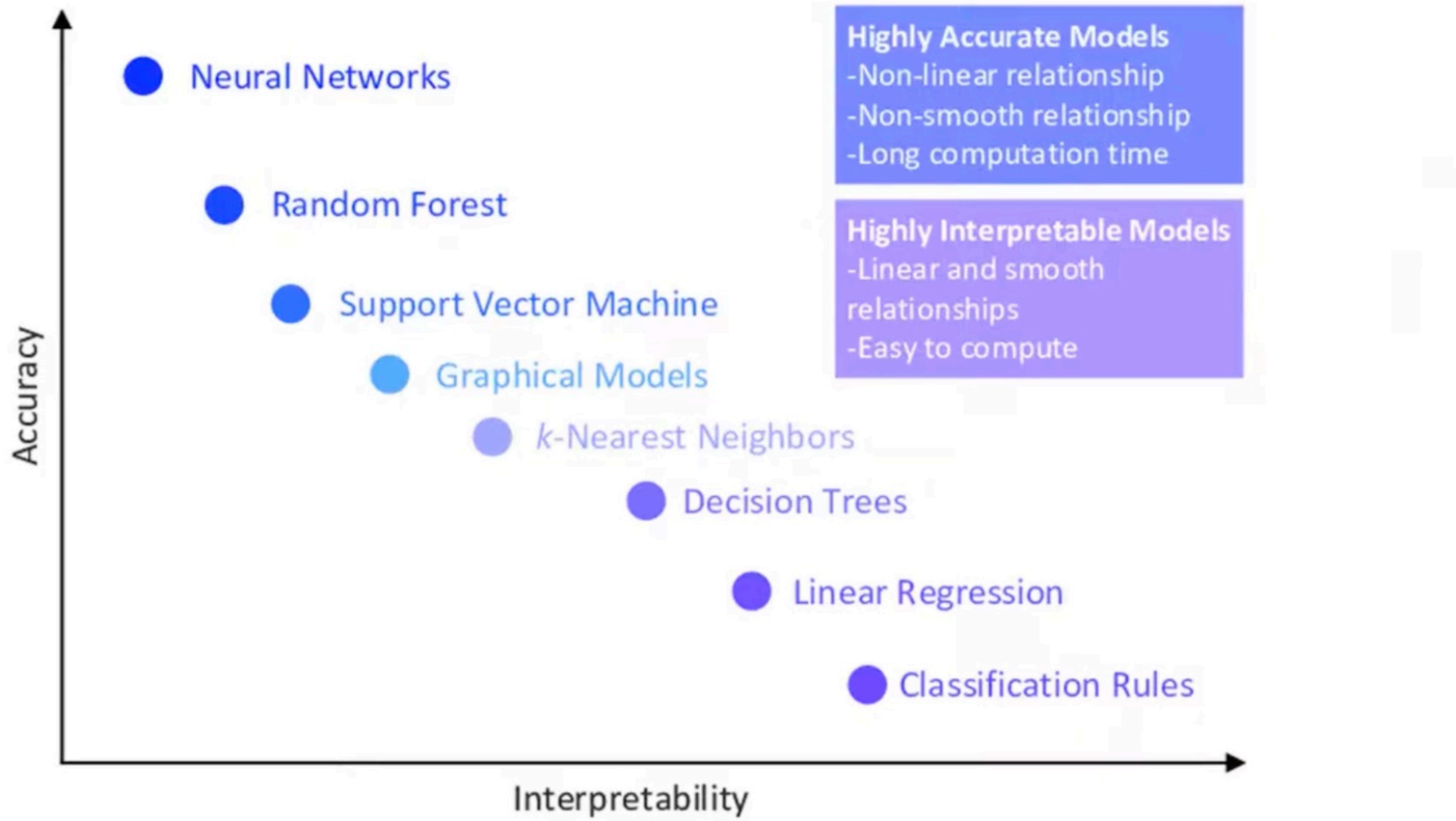
What is explainable AI?

Source: Google Trends for “Explainable AI”

Data Scientists &
ML Engineers

Users of the
algorithms

What is explainable AI?



Source: Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions

Understanding Machine Learning models

„Model based“

Build **interpretable** ML models

„Post-hoc“

Derive explanations for **complex** ML models

Black-box approach

White-box approach

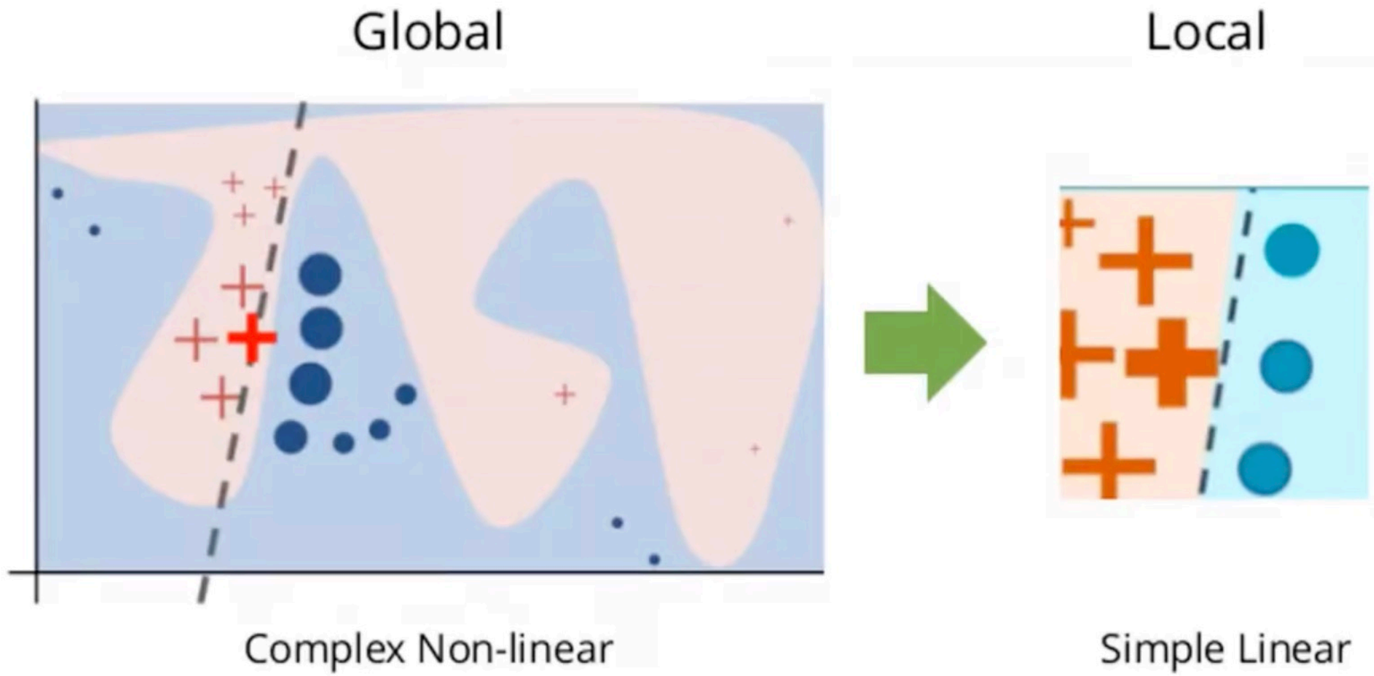
Categorization in XAI

Model-**agnostic**	Applicable to all model types
Model-**specific**	Only applicable to a specific model type

Agnosticity

Global explanation	Explaining the whole model
Local explanation	Explaining individual predictions

Scope



Source: "Why Should I Trust You?": Explaining the Predictions of Any Classifier, Ribeiro et al.

Categorization in XAI

Model-**agnostic** — Applicable to all model types

Model-**specific** — Only applicable to a specific model type

Agnosticity

Global explanation — Explaining the whole model

Local explanation — Explaining individual predictions

Scope

Graph

Image

Text / Speech

Tabular

Data type

Visual

Feature importance

Data points

Surrogate models

Explaination type

Explainable AI

Cheat sheet ex.pegg.io v.0.2

Start

Explaining a model's decisions?

No

Ah, so we're exploring the model and its internals

Yes

Is the model interpretable by design?

Yes

Good choice! *Especially* for high-stakes decisions

Interpretable Models

Logistic regression

KNN

Linear models

No

So less-interpretable methods perform better for your use-case? We understand!

Need a method that works on all models?

Yes

Model Agnostic

SHAP

Perturbation

LIME

No

Model Specific

Gradient saliency

Integrated gradients

Attention

Also yes

Using special examples from the dataset could uncover insights about the model

Example Based

Adversarial

Influence functions

Counterfactual

Neural Representations

SVCCA

Activation maximization

Probes

Feature visualization

TCAV

Arpeggio