Understanding the Machine Learning Workflow with scikit-learn

Janani Ravi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

scikit-learn in the typical ML workflow

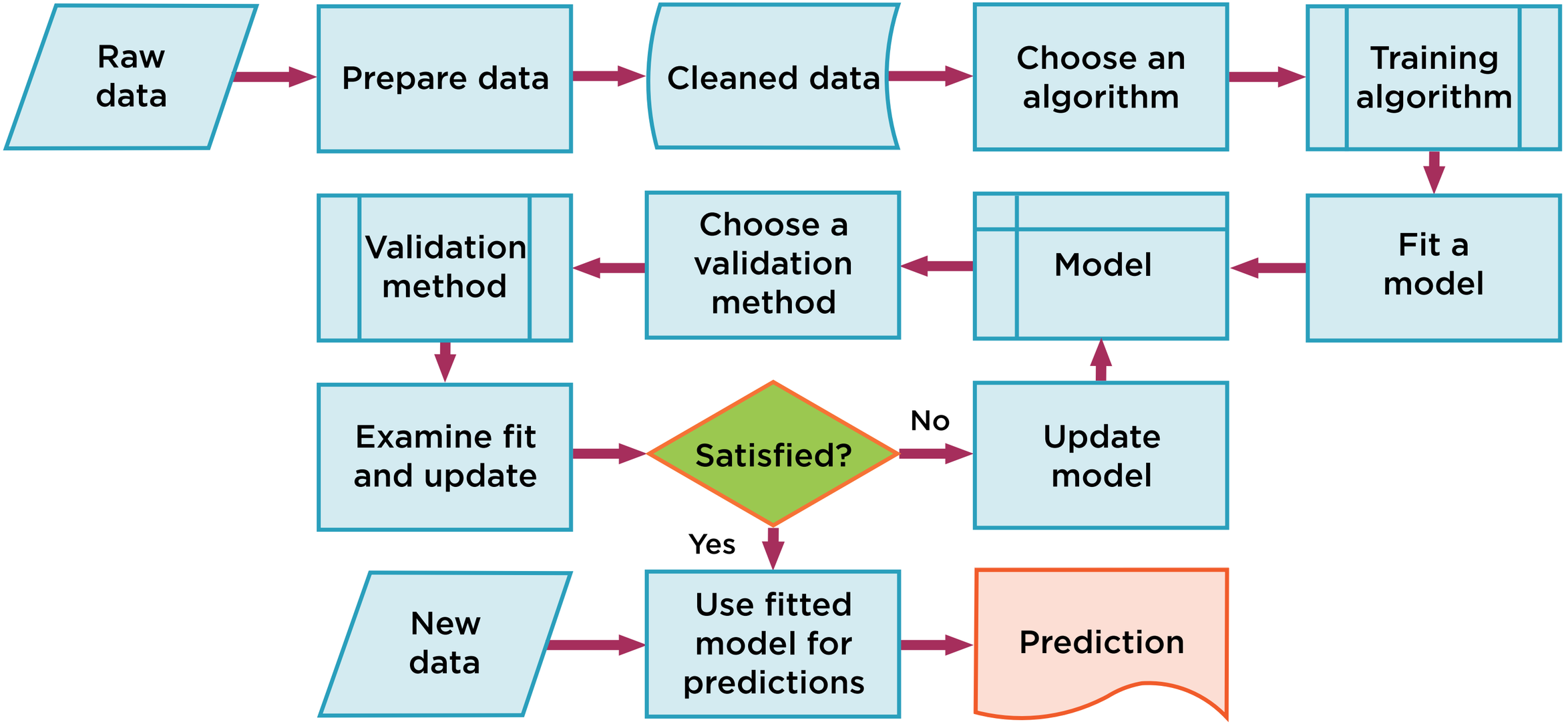
Estimators and pipelines

Model evaluation and transformation

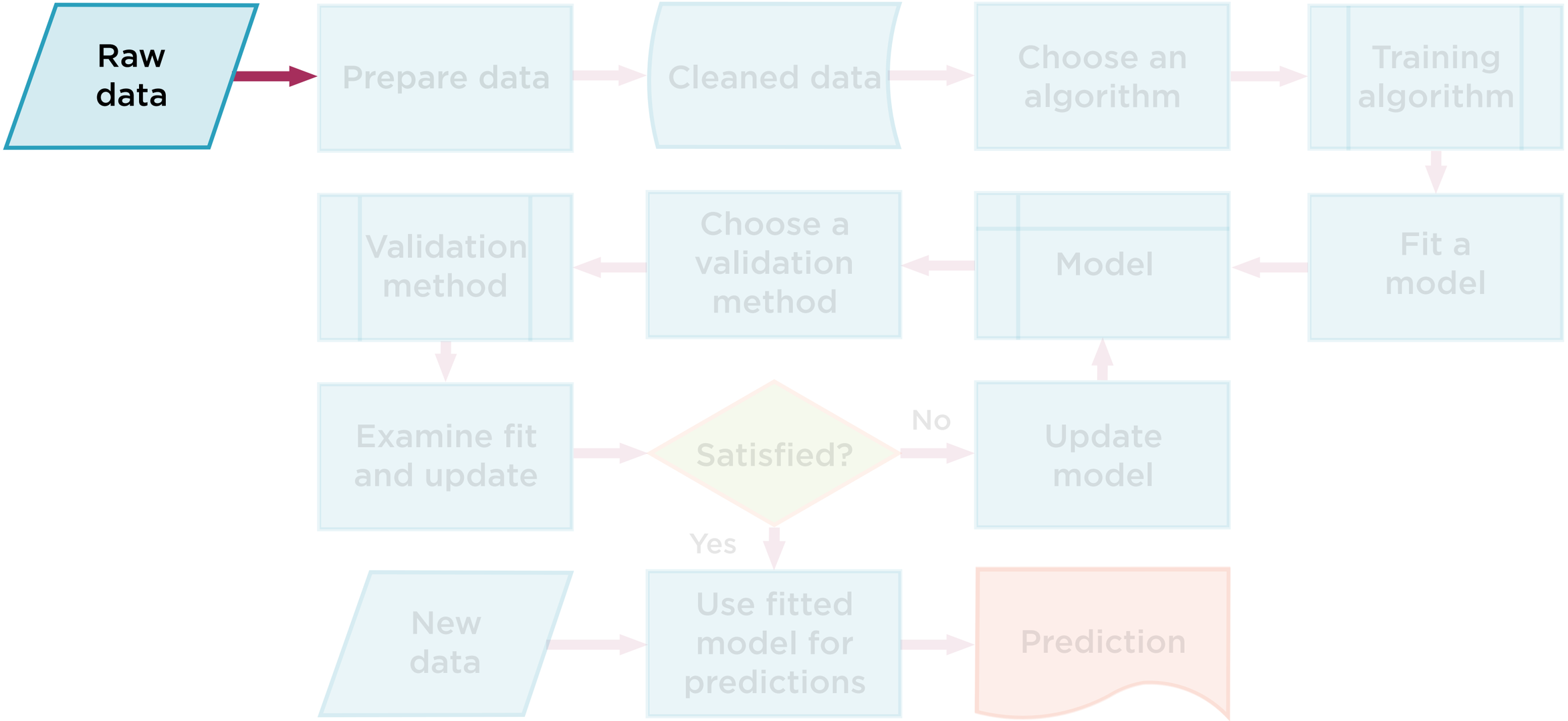
Loading, cleaning, transforming, and visualizing datasets

Machine Learning Workflow

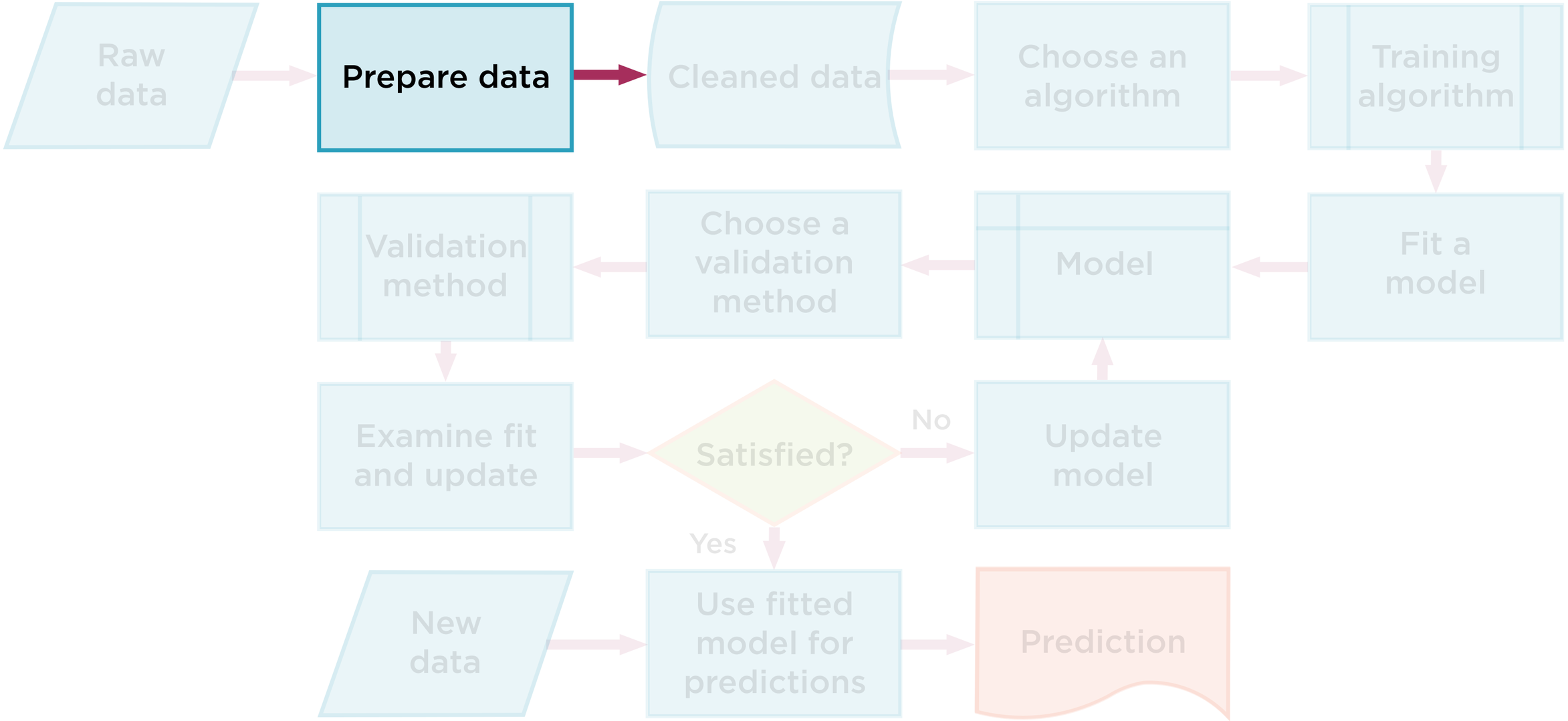
Basic Machine Learning Workflow



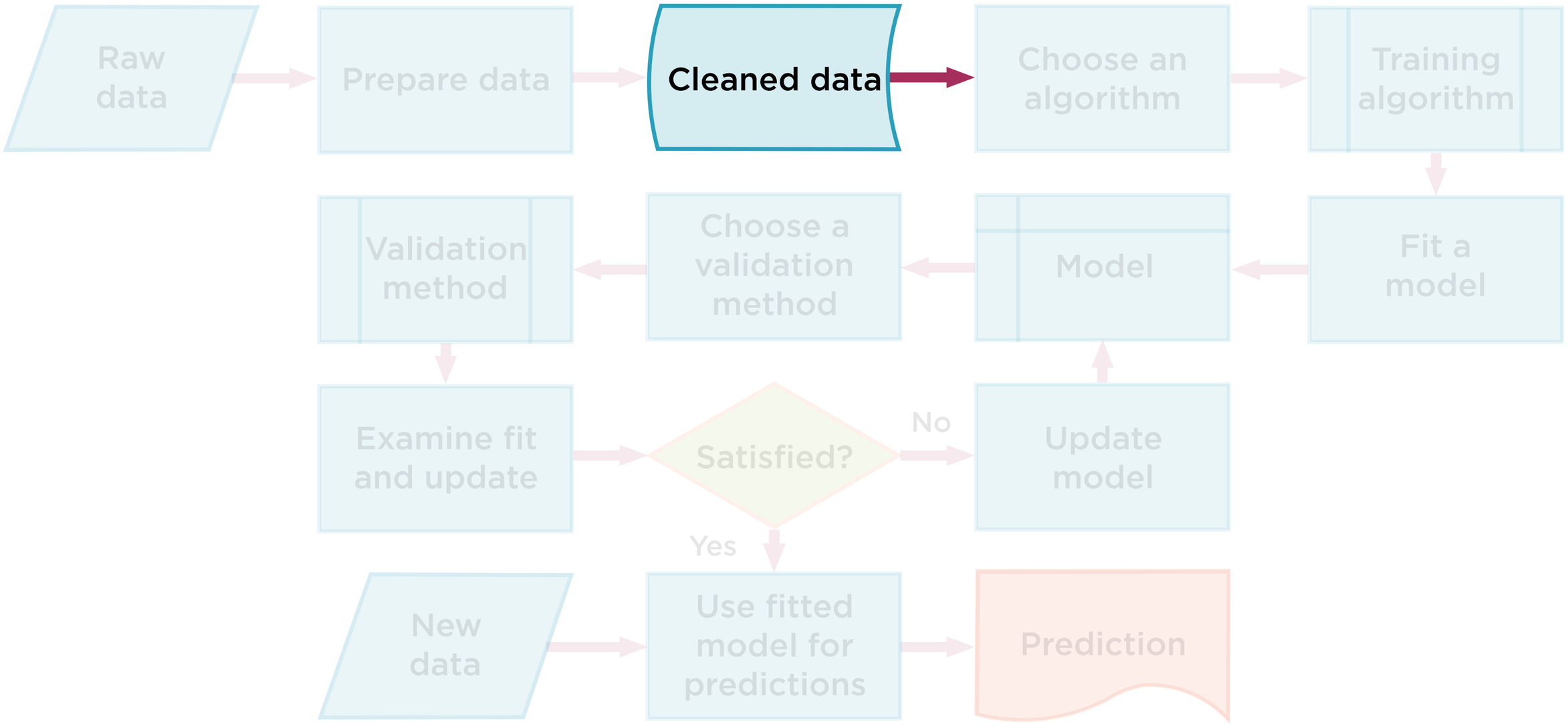
What Data Do You Have to Work With?



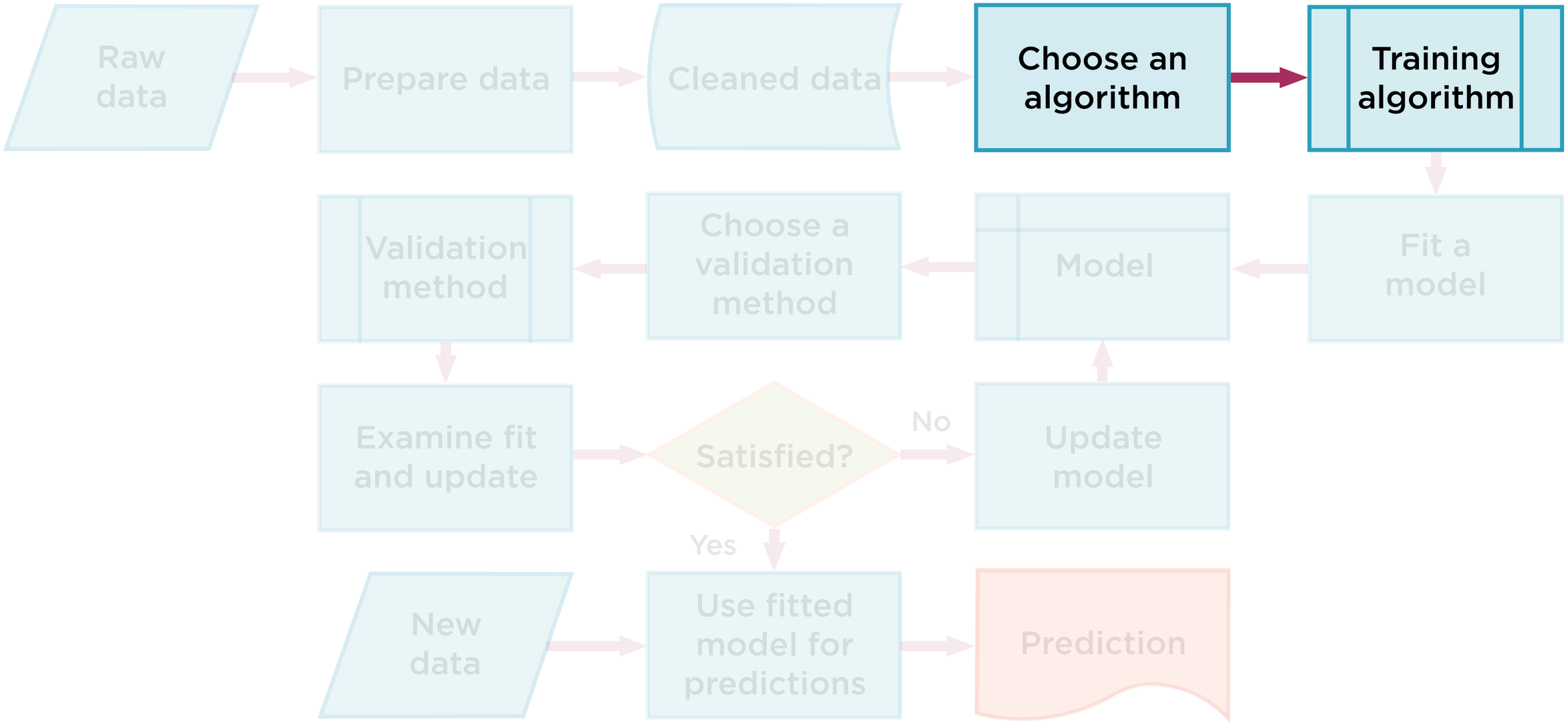
Load and Store Data



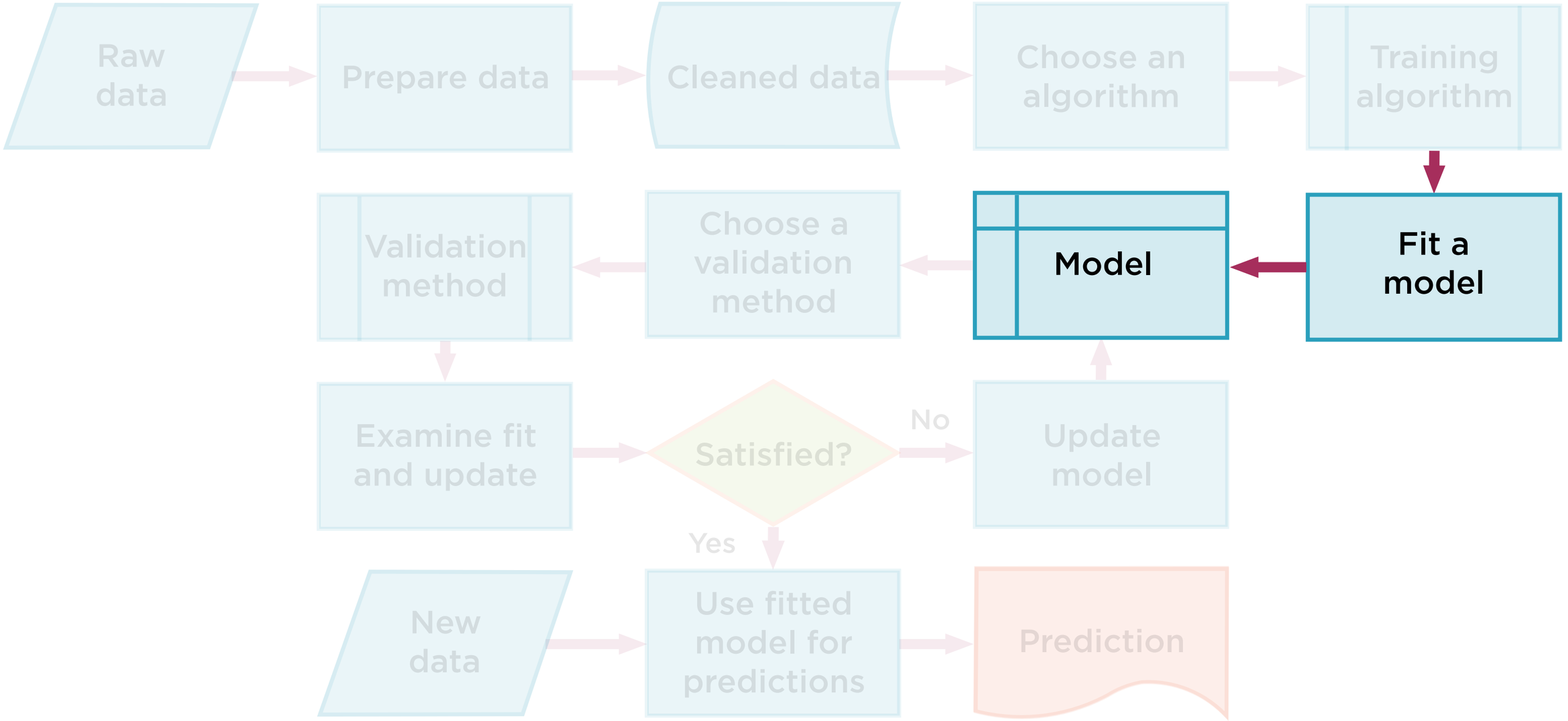
Data Preprocessing



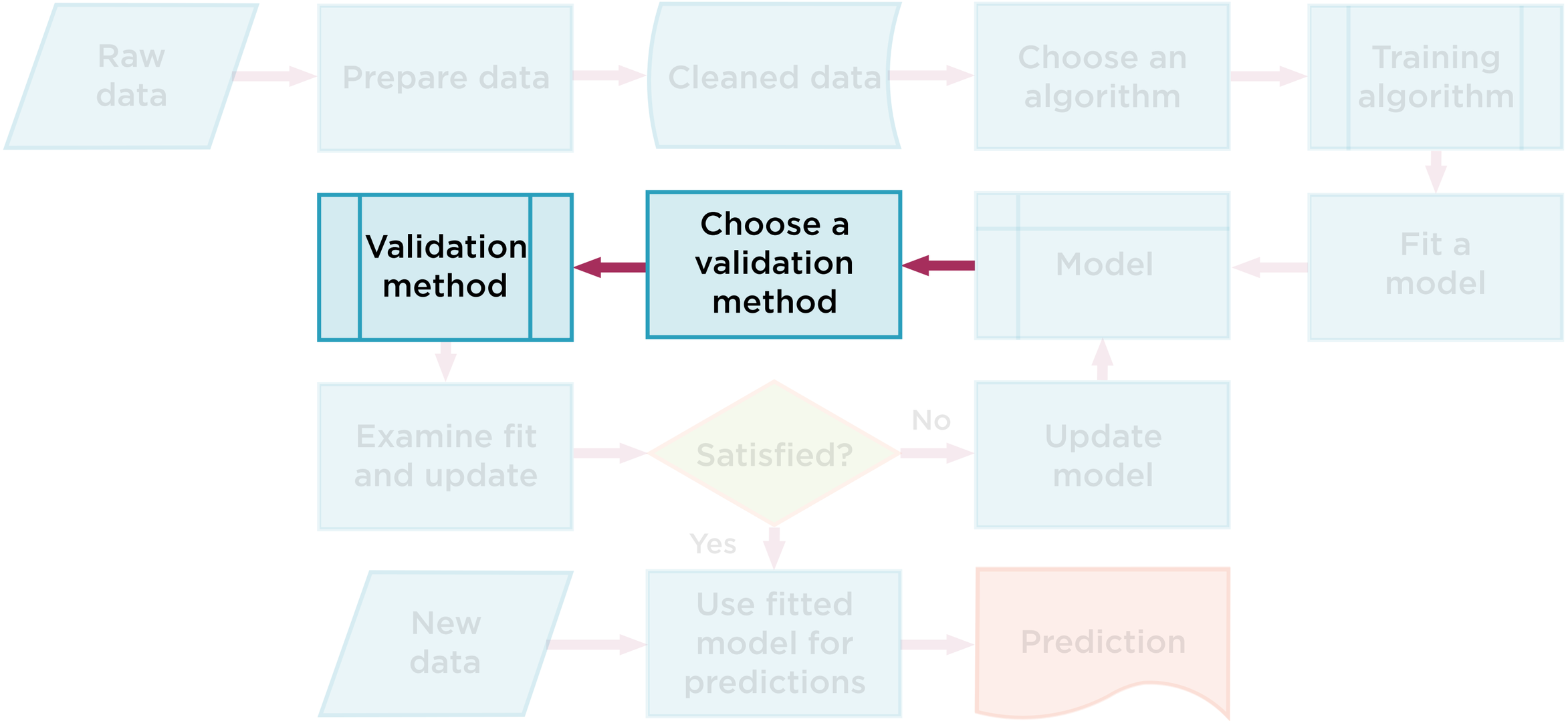
Decision Trees, Support Vector Machines?



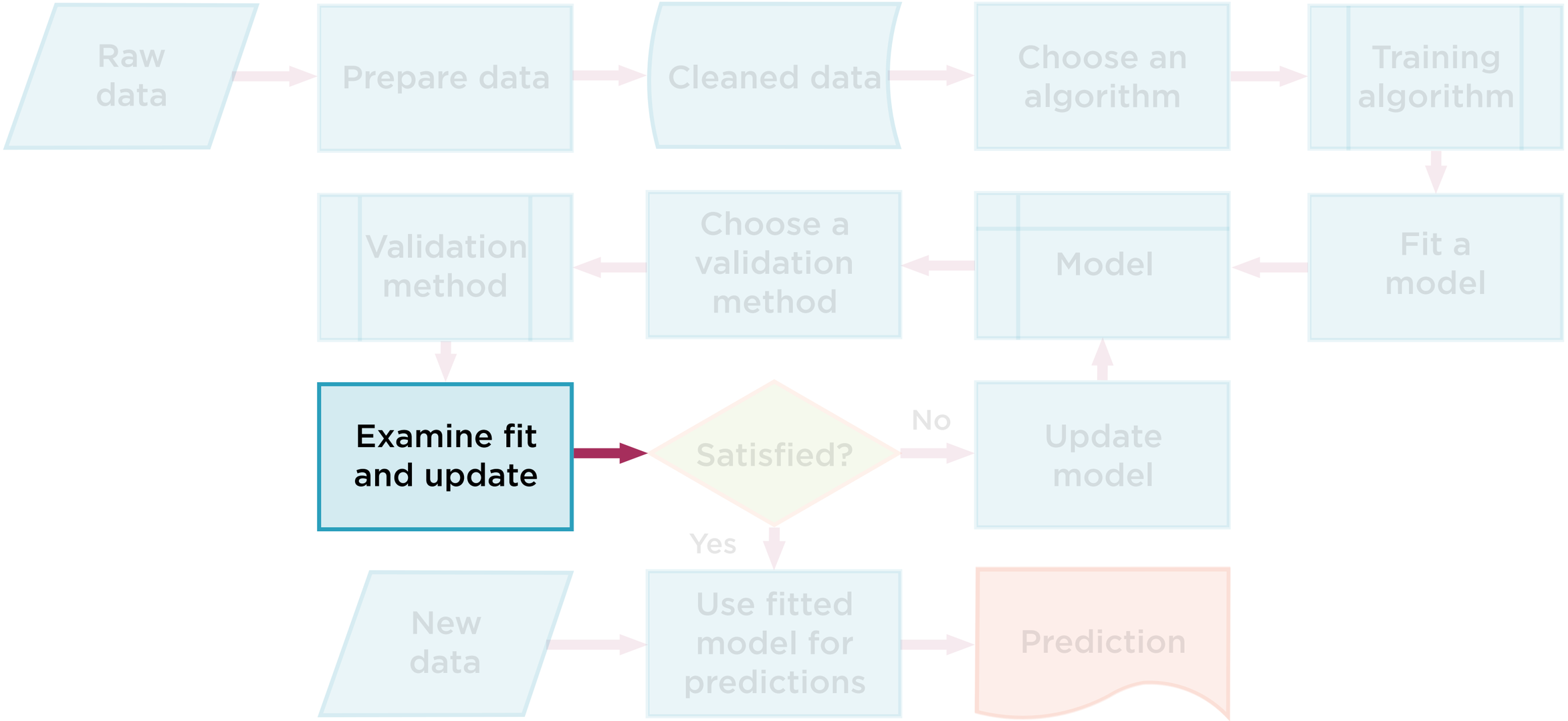
Training to Find Model Parameters



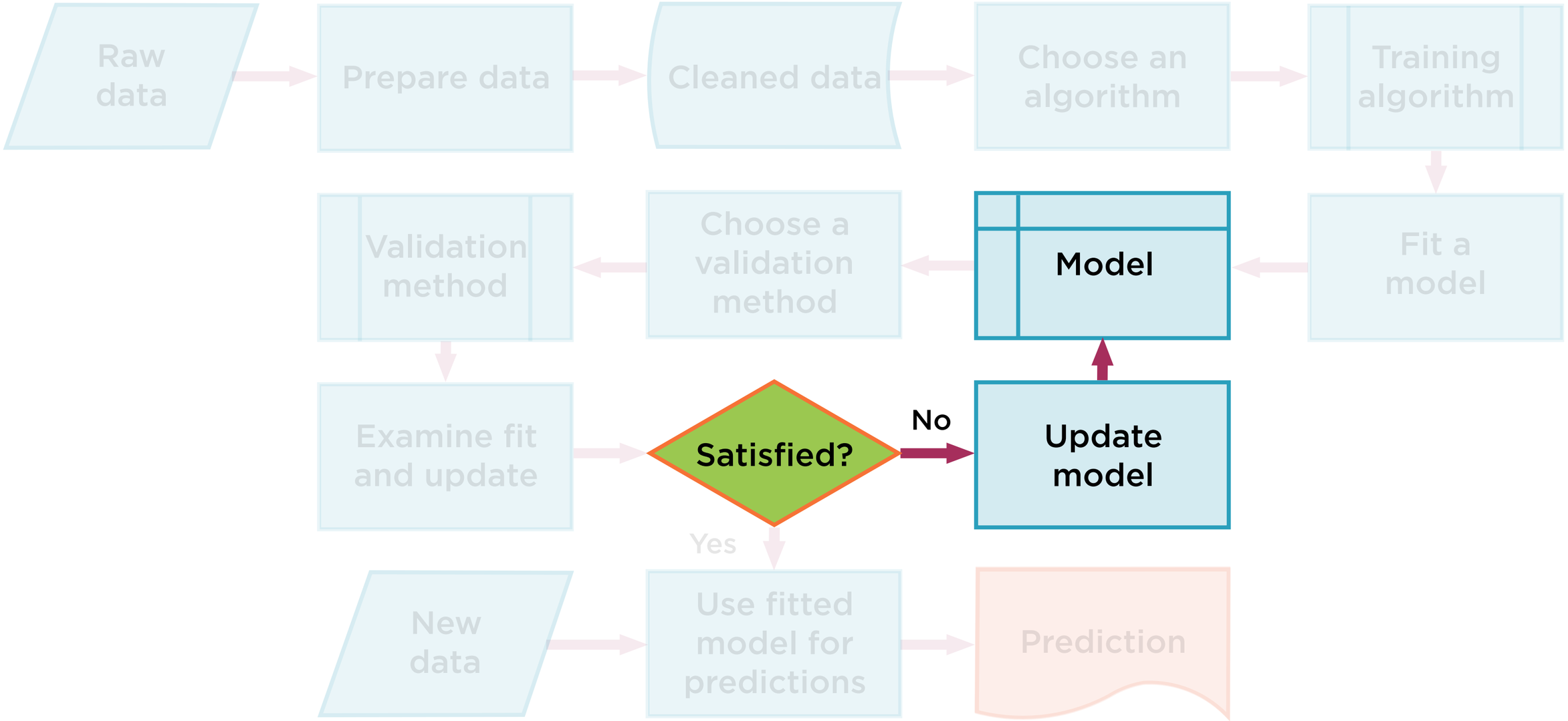
Evaluate the Model



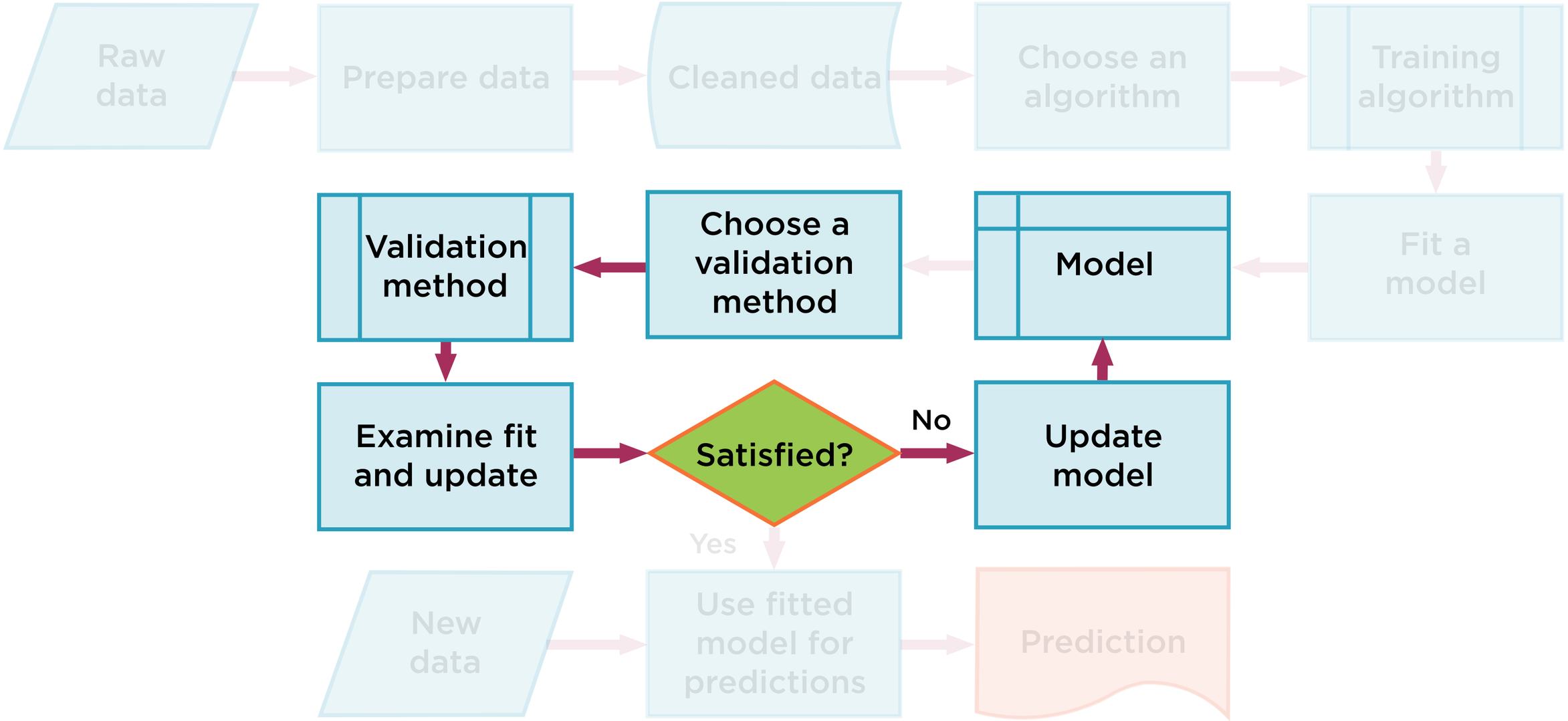
Score the Model



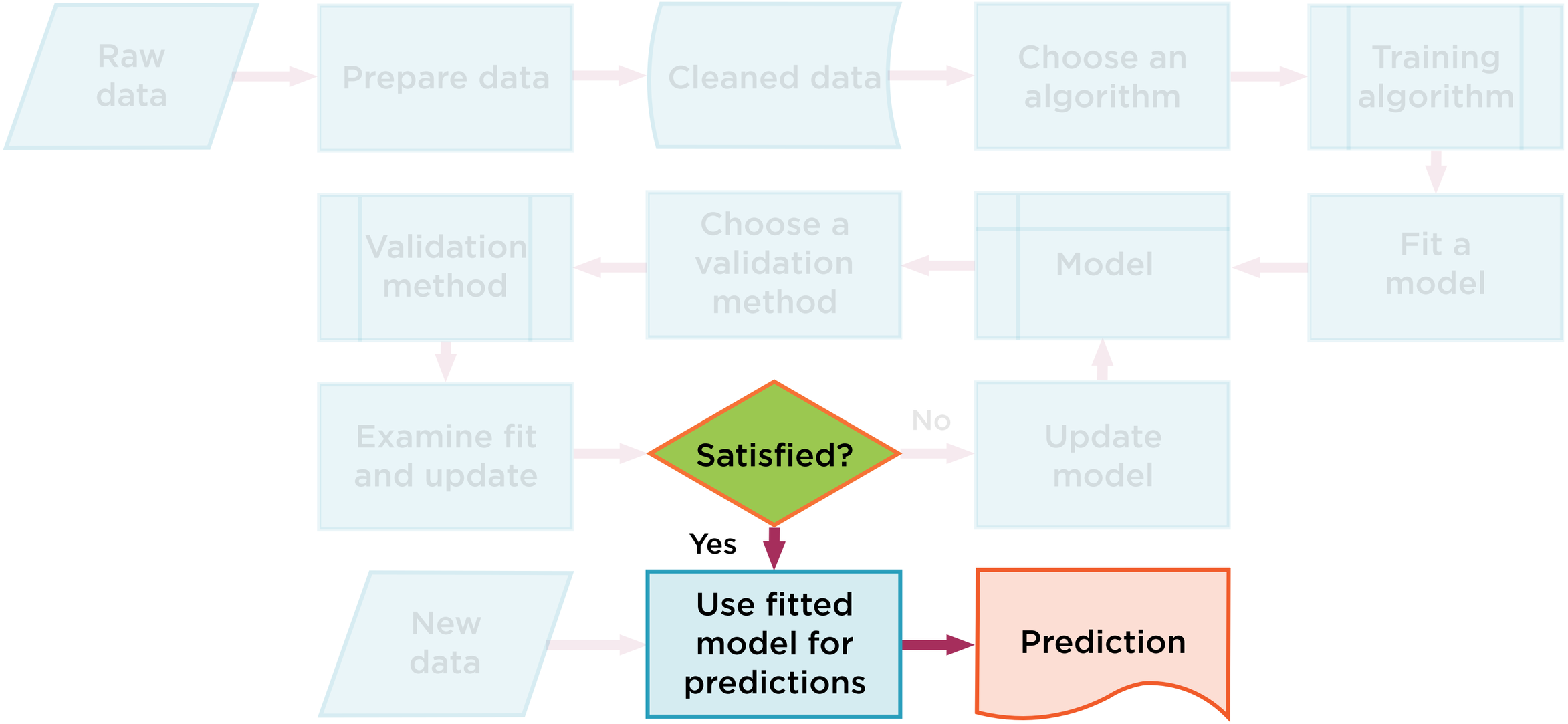
Different Algorithm, More Data, More Training?



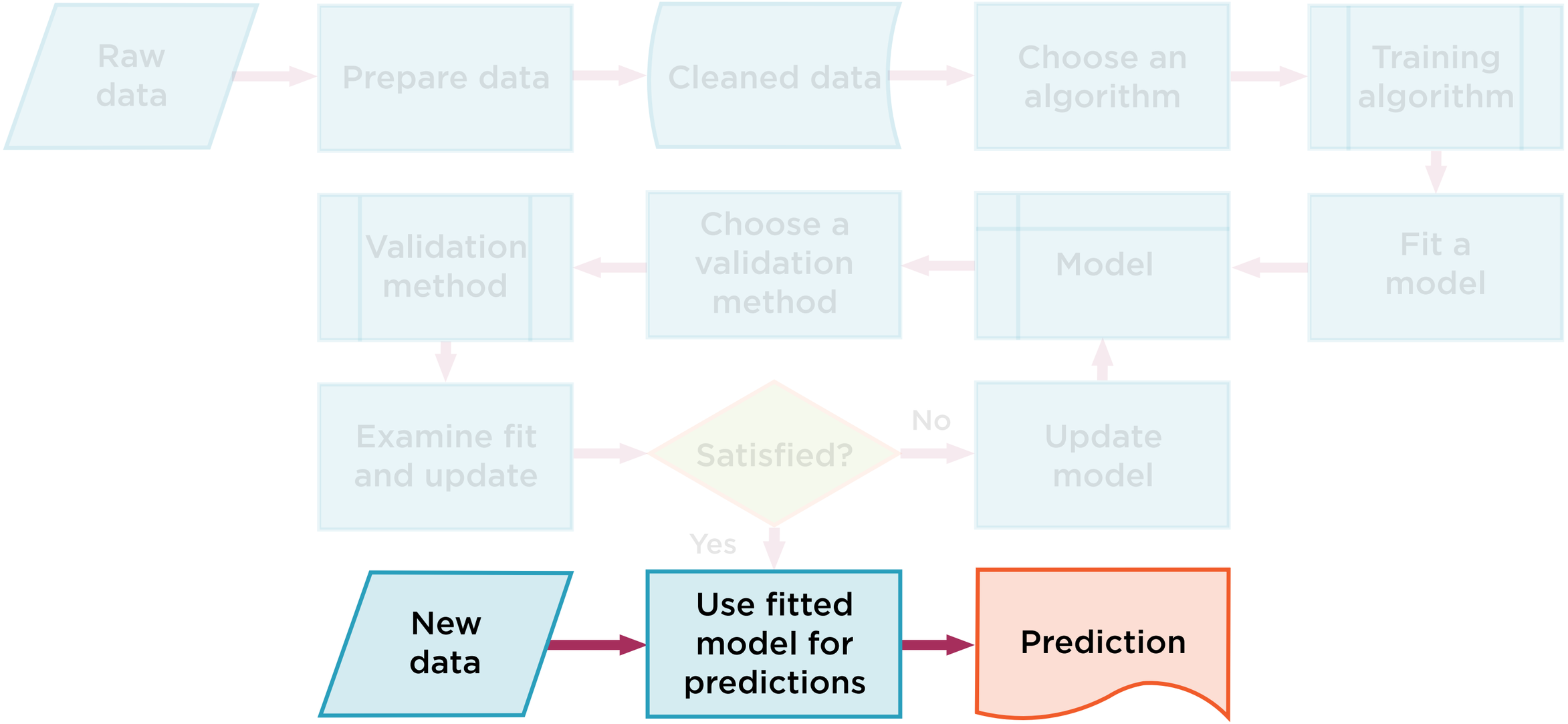
Iterate Till Model Finalized



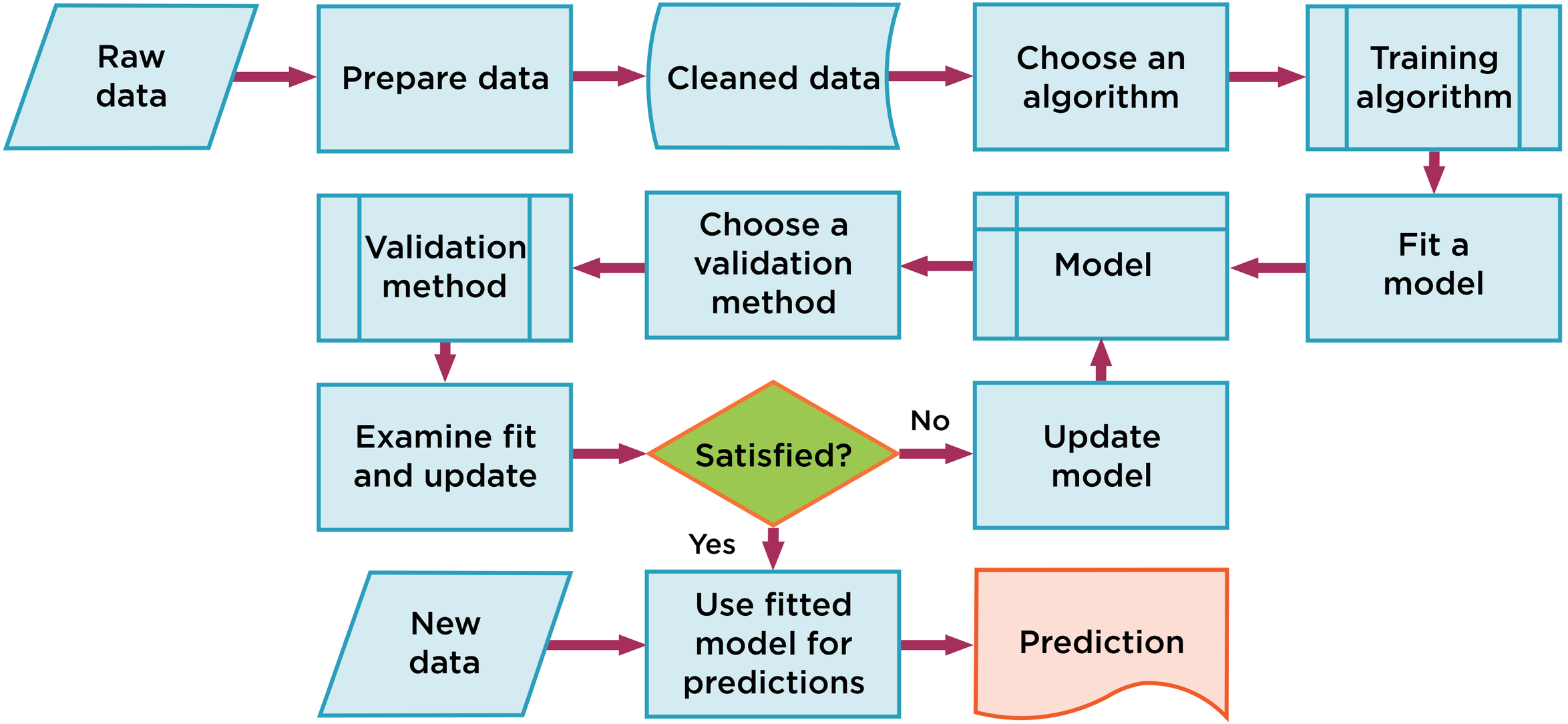
Model Used for Predictions



Retrained Using New Data



Basic Machine Learning Workflow

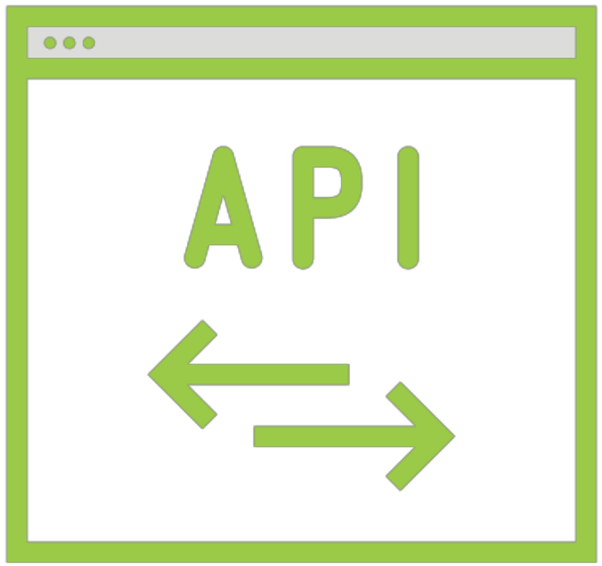


scikit-learn has objects and functions for every step in the ML workflow

The Estimator API

scikit-learn - easy-to-use, very comprehensive and efficient Python library for traditional ML models

The Estimator API



Estimator API for consistent interface

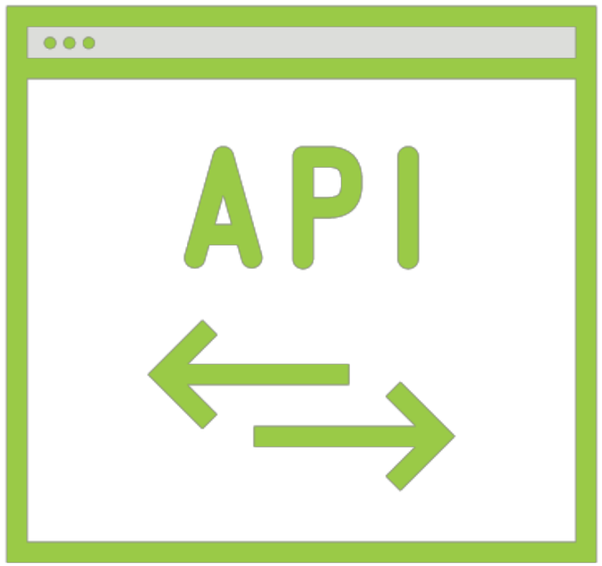
Create a model object

Fit to training data

Predict for new data

Pipelines for complex operations

Principles Underlying Estimator APIs



Consistency

Inspection

Limited object hierarchy

Composition

Sensible defaults

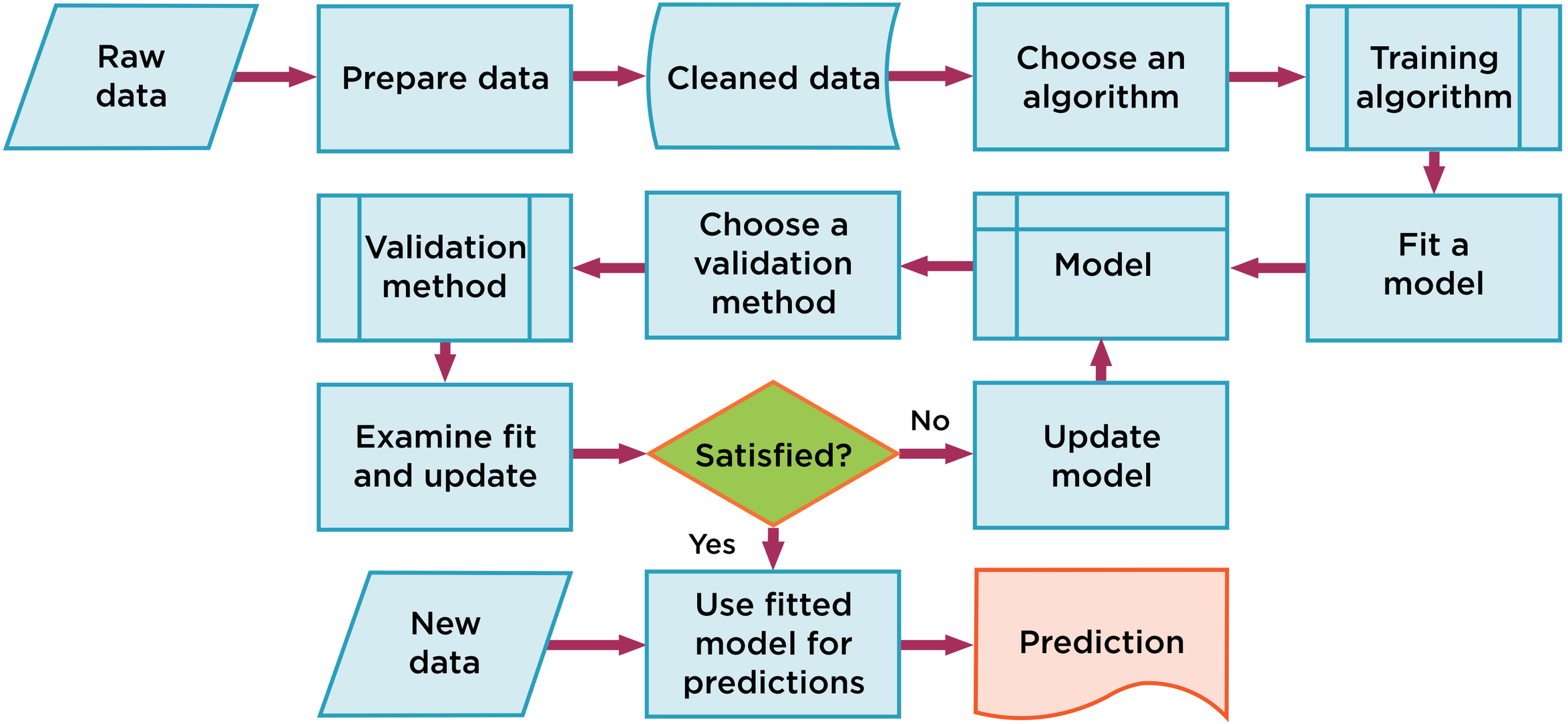
Scikit-Learn's Estimator API

The Scikit-Learn API is designed with the following guiding principles in mind, as outlined in the Scikit-Learn API paper:

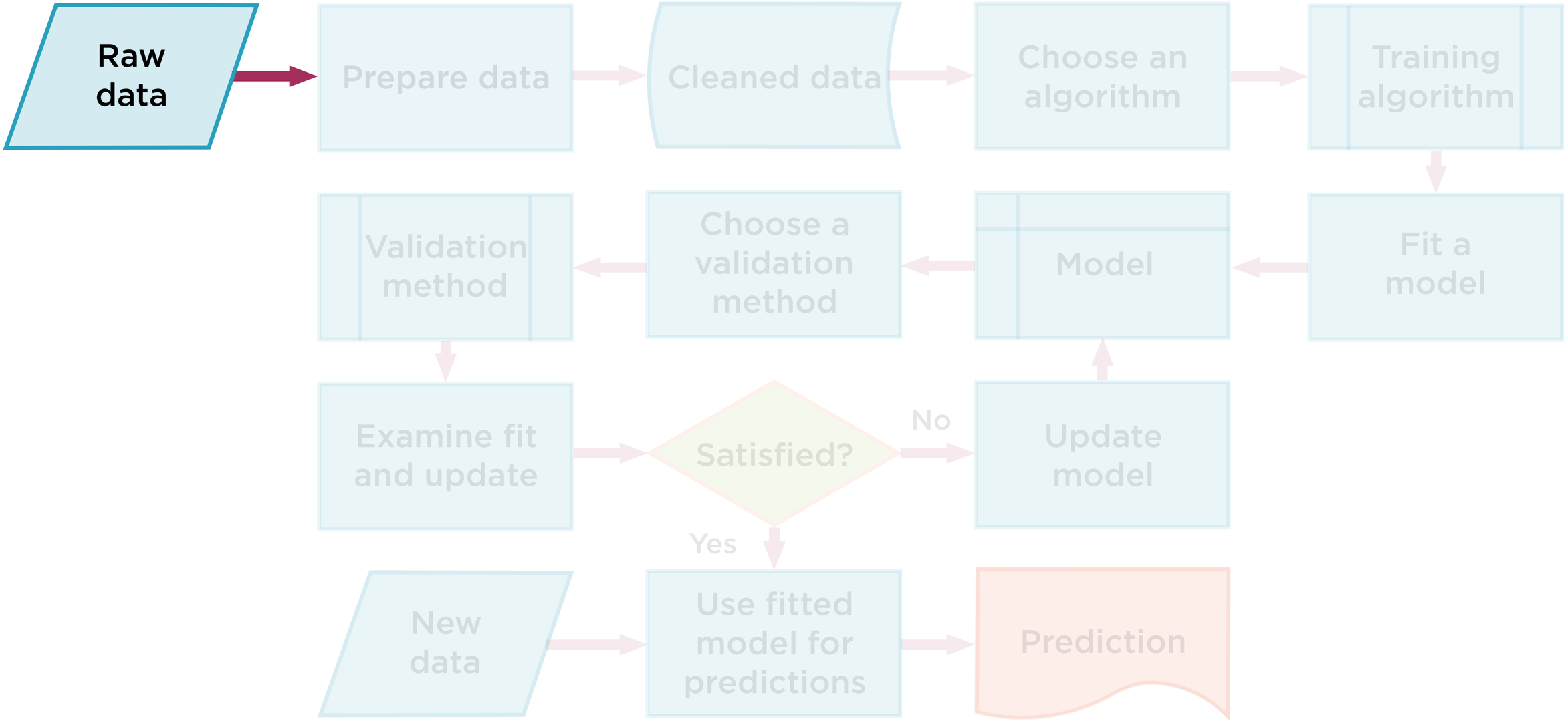
- *Consistency*: All objects share a common interface drawn from a limited set of methods, with consistent documentation.
- *Inspection*: All specified parameter values are exposed as public attributes.
- *Limited object hierarchy*: Only algorithms are represented by Python classes; datasets are represented in standard formats (NumPy arrays, Pandas `DataFrame`s, SciPy sparse matrices) and parameter names use standard Python strings.
- *Composition*: Many machine learning tasks can be expressed as sequences of more fundamental algorithms, and Scikit-Learn makes use of this wherever possible.
- *Sensible defaults*: When models require user-specified parameters, the library defines an appropriate default value.

In practice, these principles make Scikit-Learn very easy to use, once the basic principles are understood. Every machine learning algorithm in Scikit-Learn is implemented via the Estimator API, which provides a consistent interface for a wide range of machine learning applications.

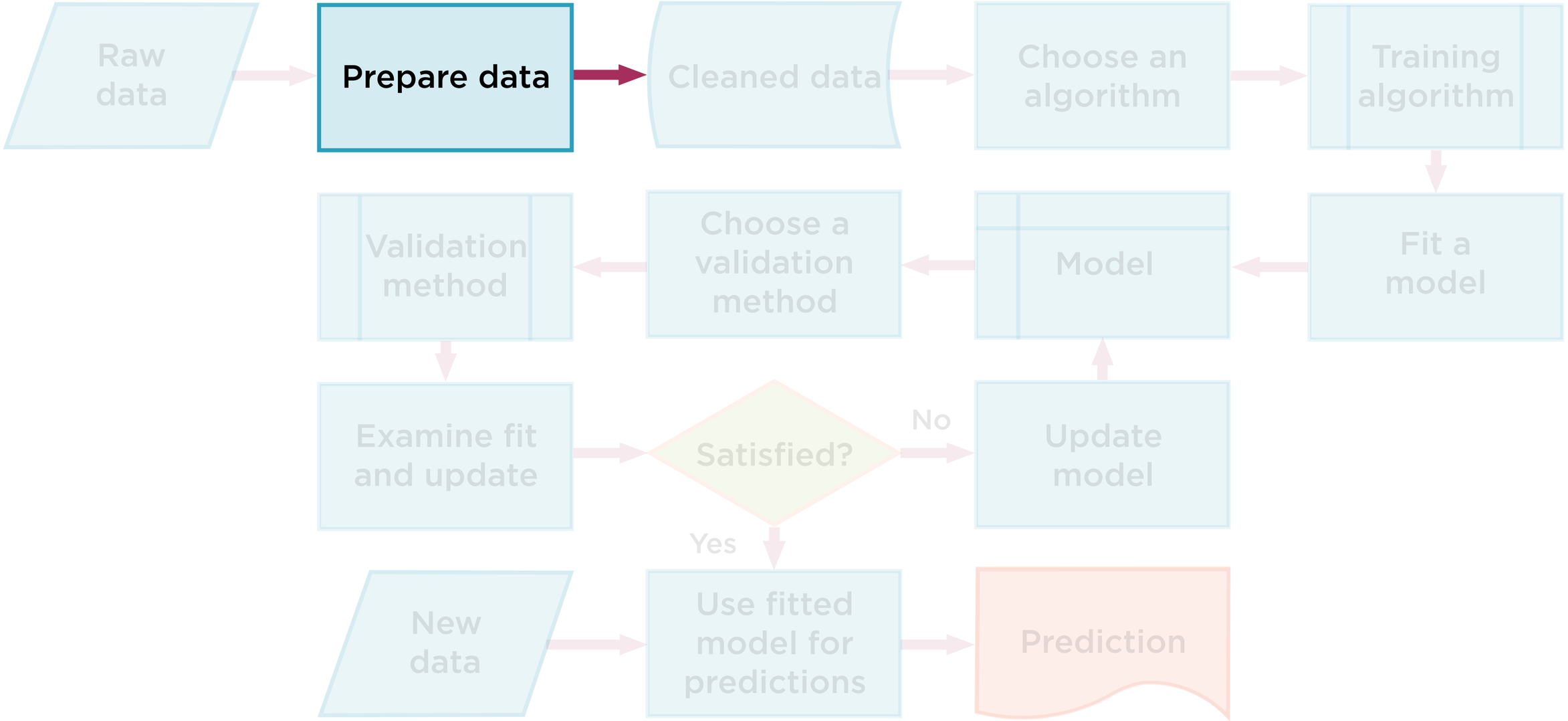
Basic Machine Learning Workflow



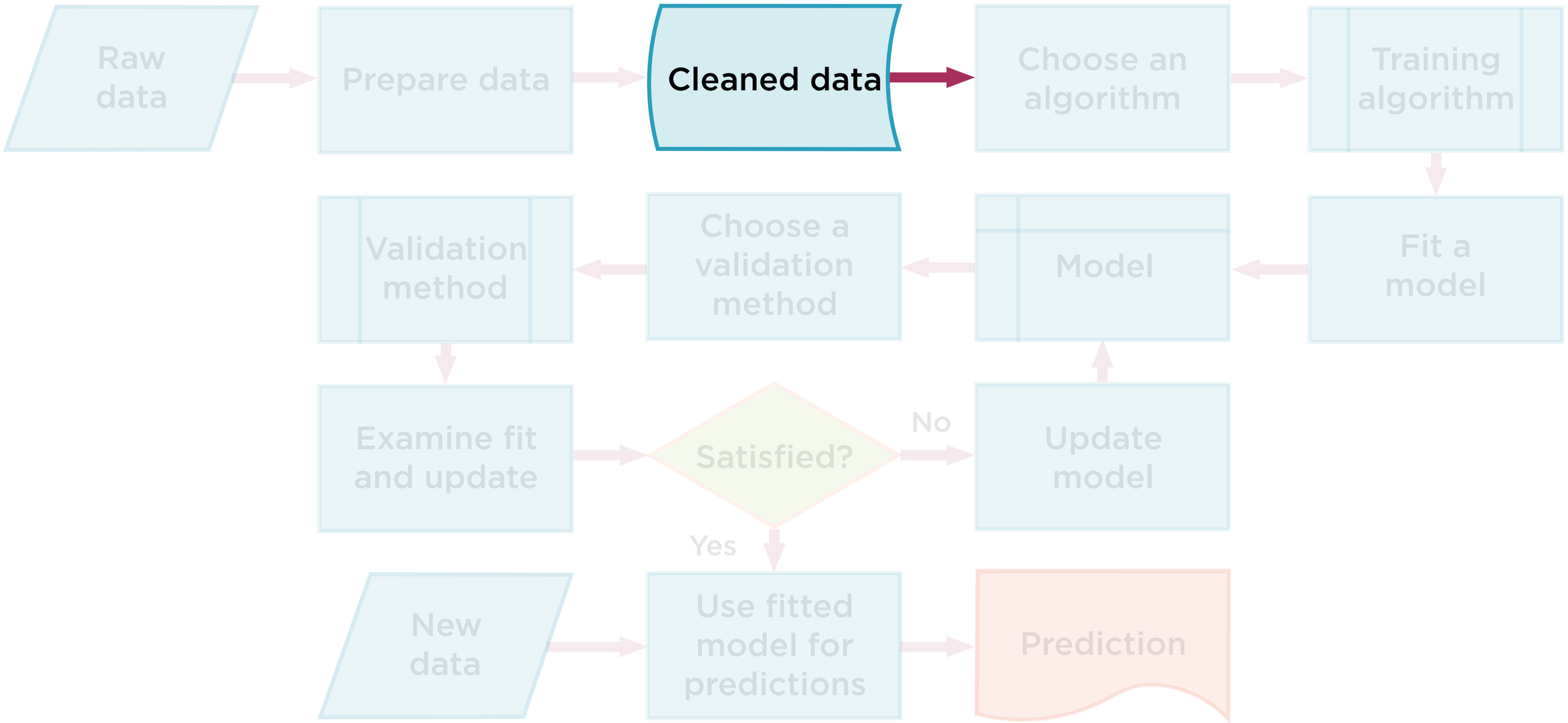
Pandas and NumPy Inter-operability



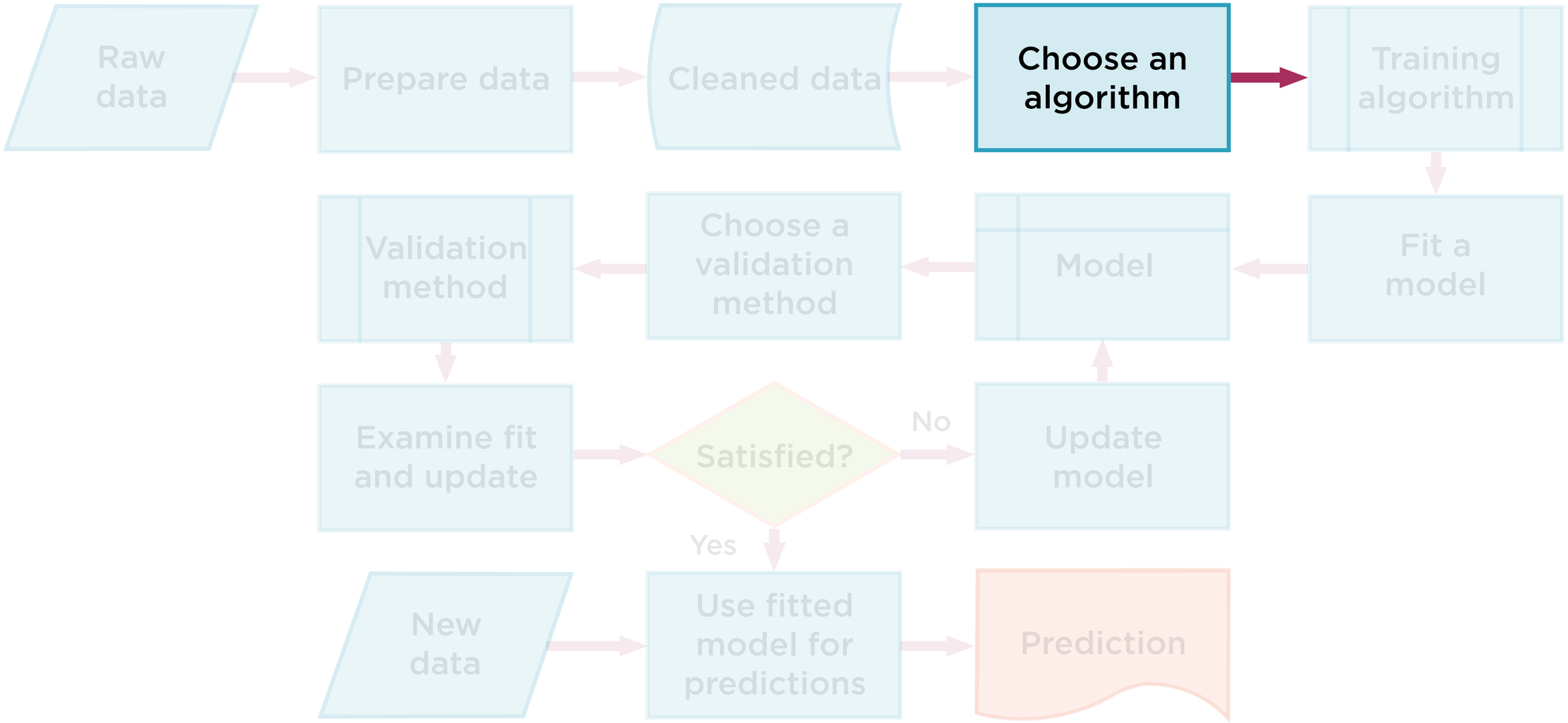
Standardization, Normalization, Scaling



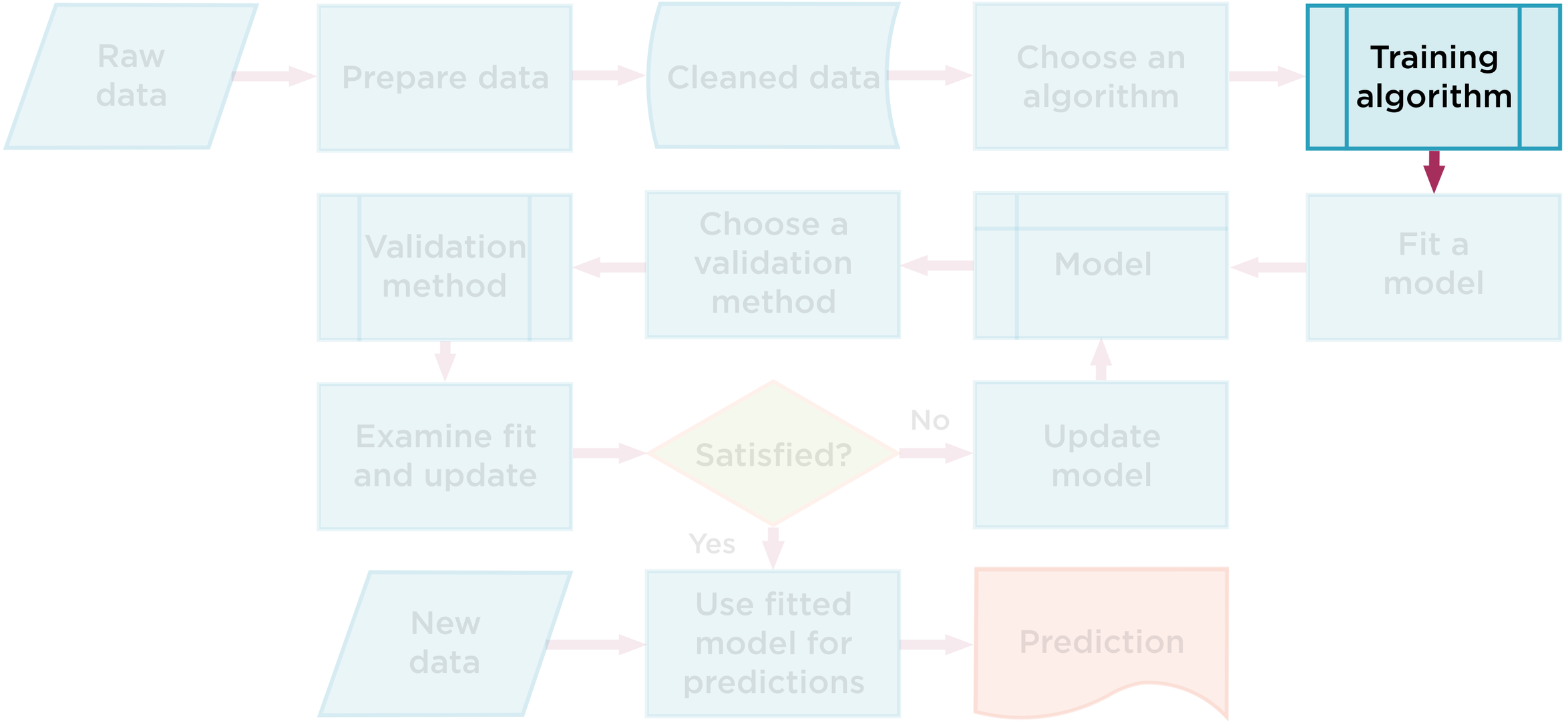
Missing Values and Outliers



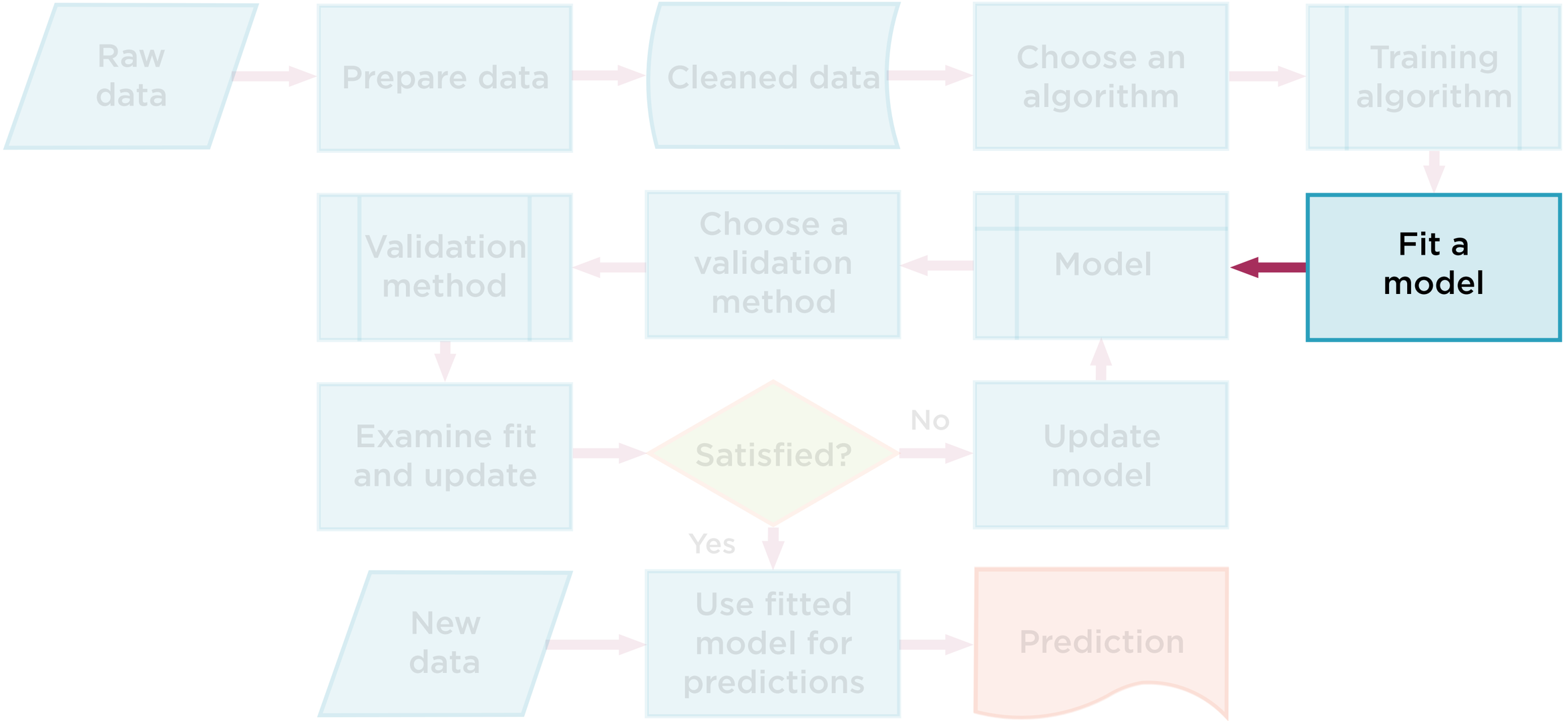
Comprehensive Suite of Algorithms



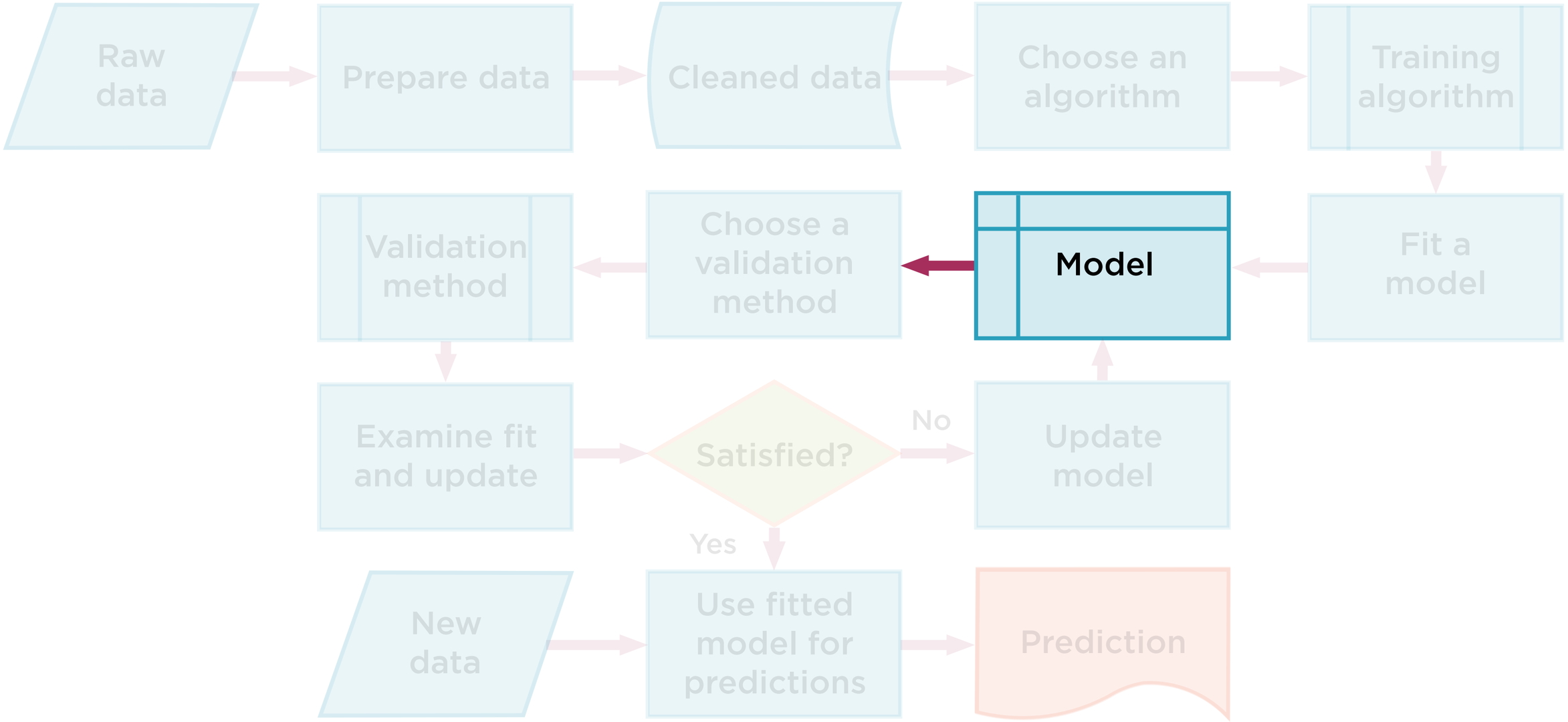
Find Best Model Parameters



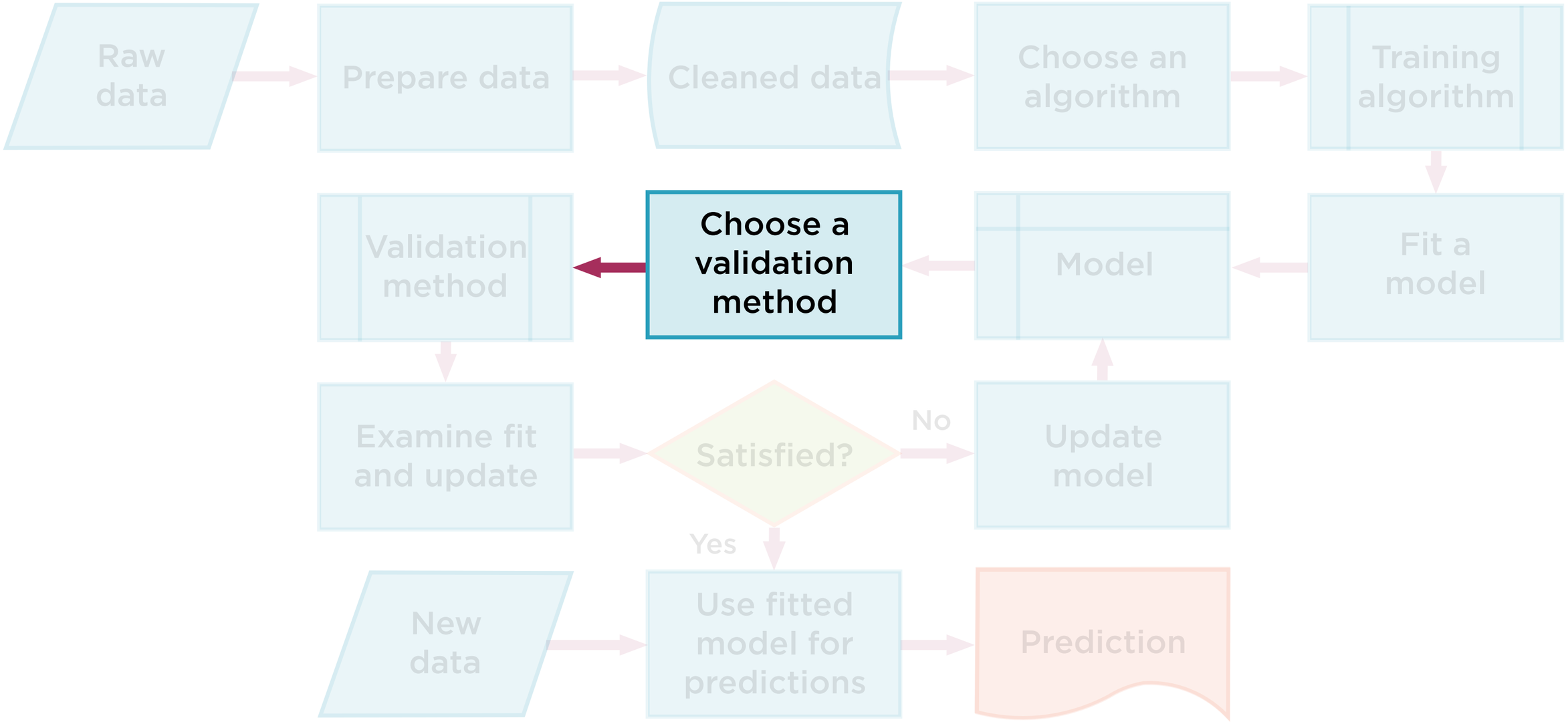
Invoke fit() Or fit_transform()



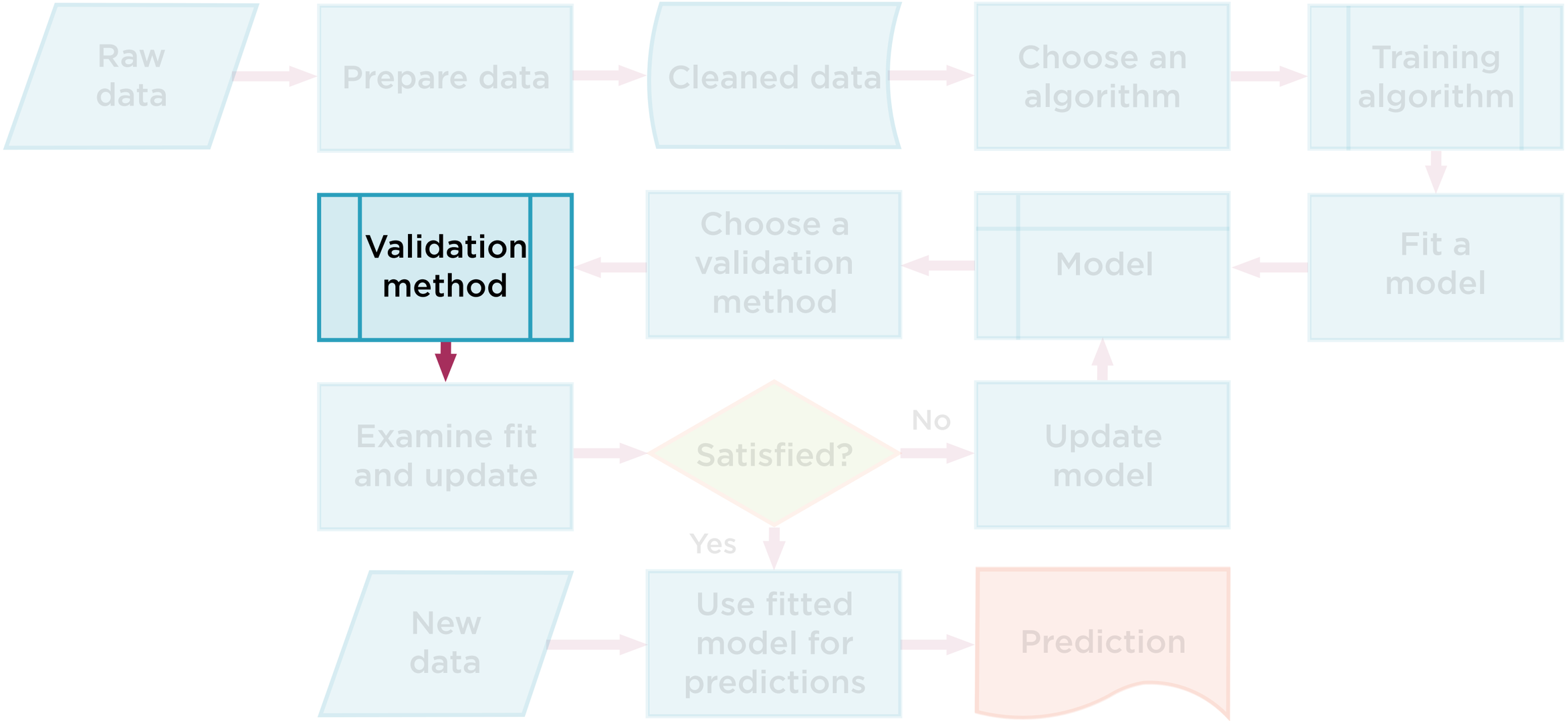
Trained Model Available



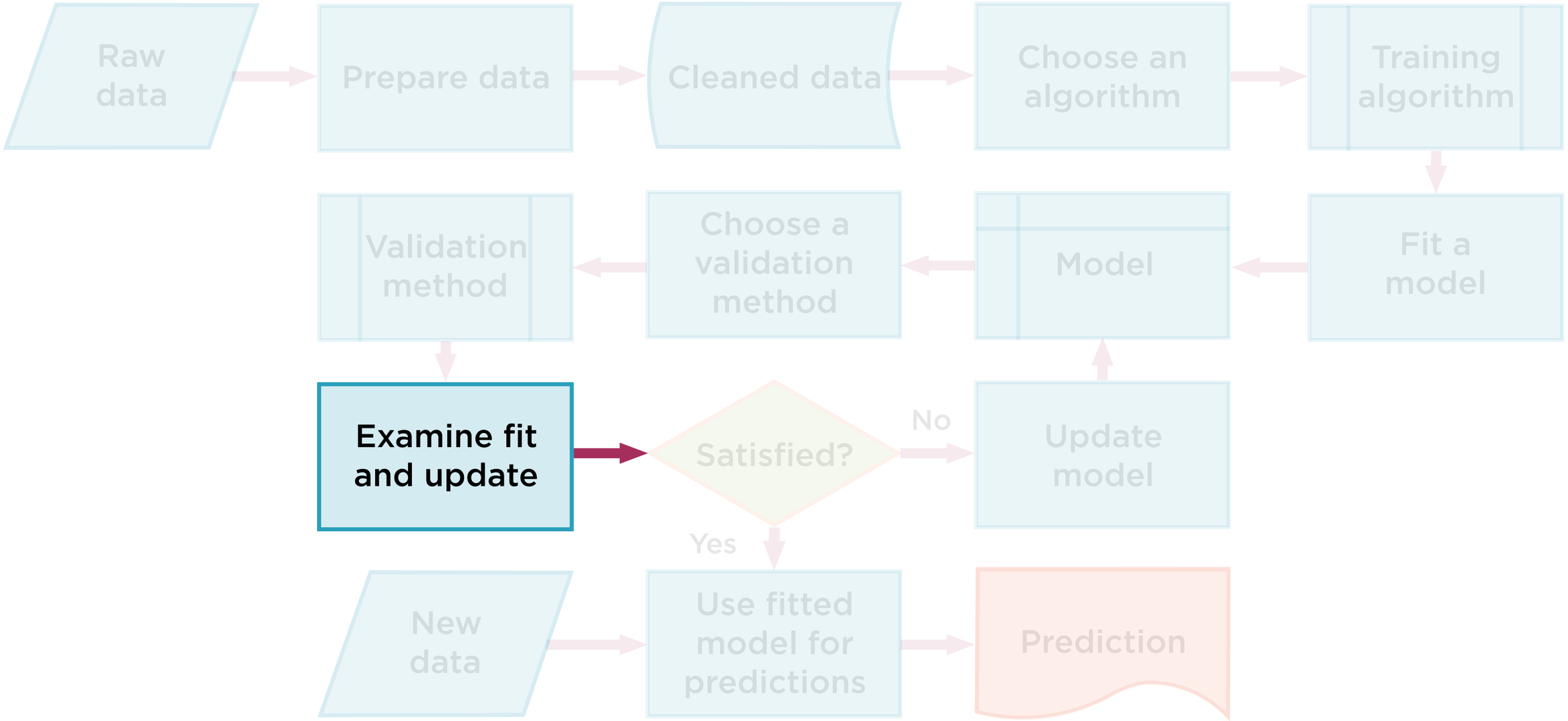
Comprehensive Suite of Cross-validation Tools



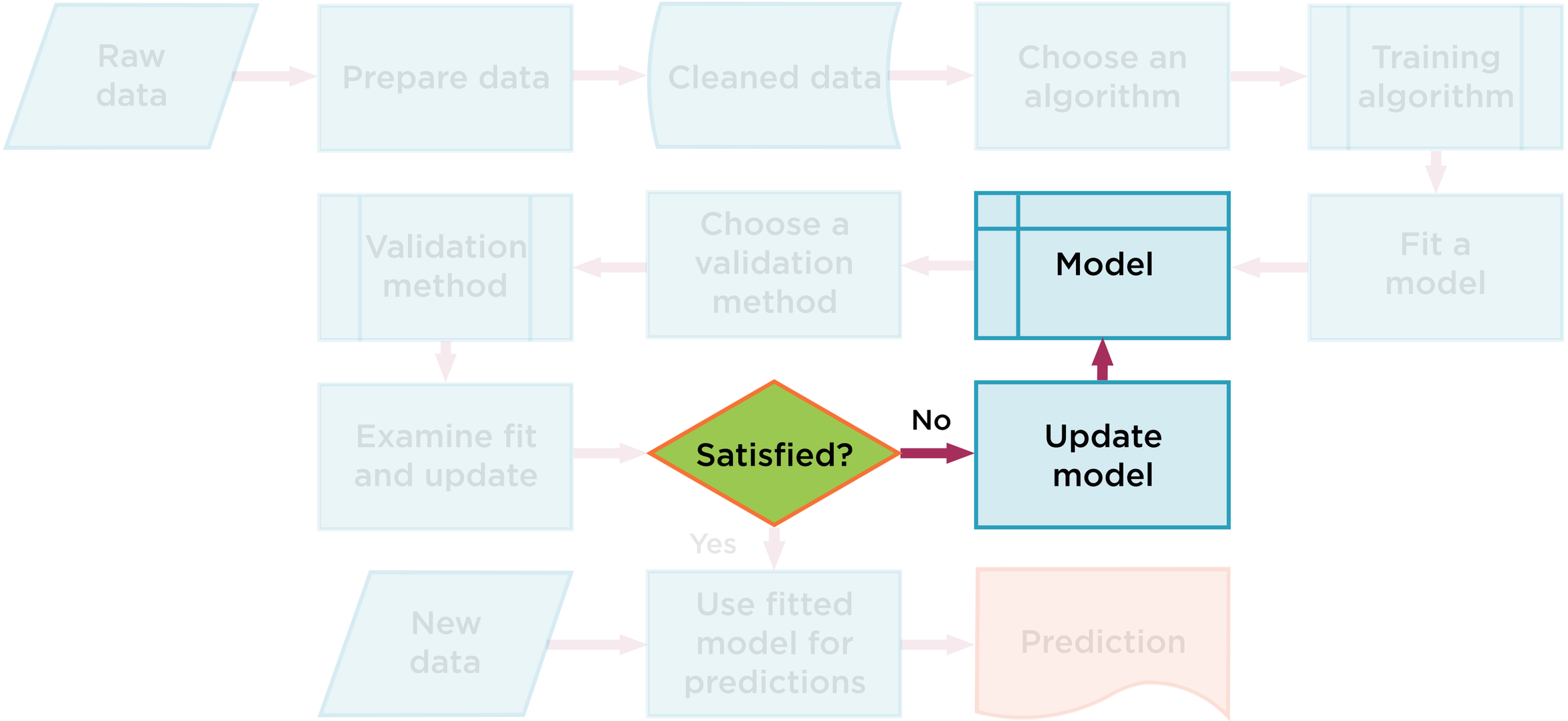
Cross-validation, K-fold, Group K-fold



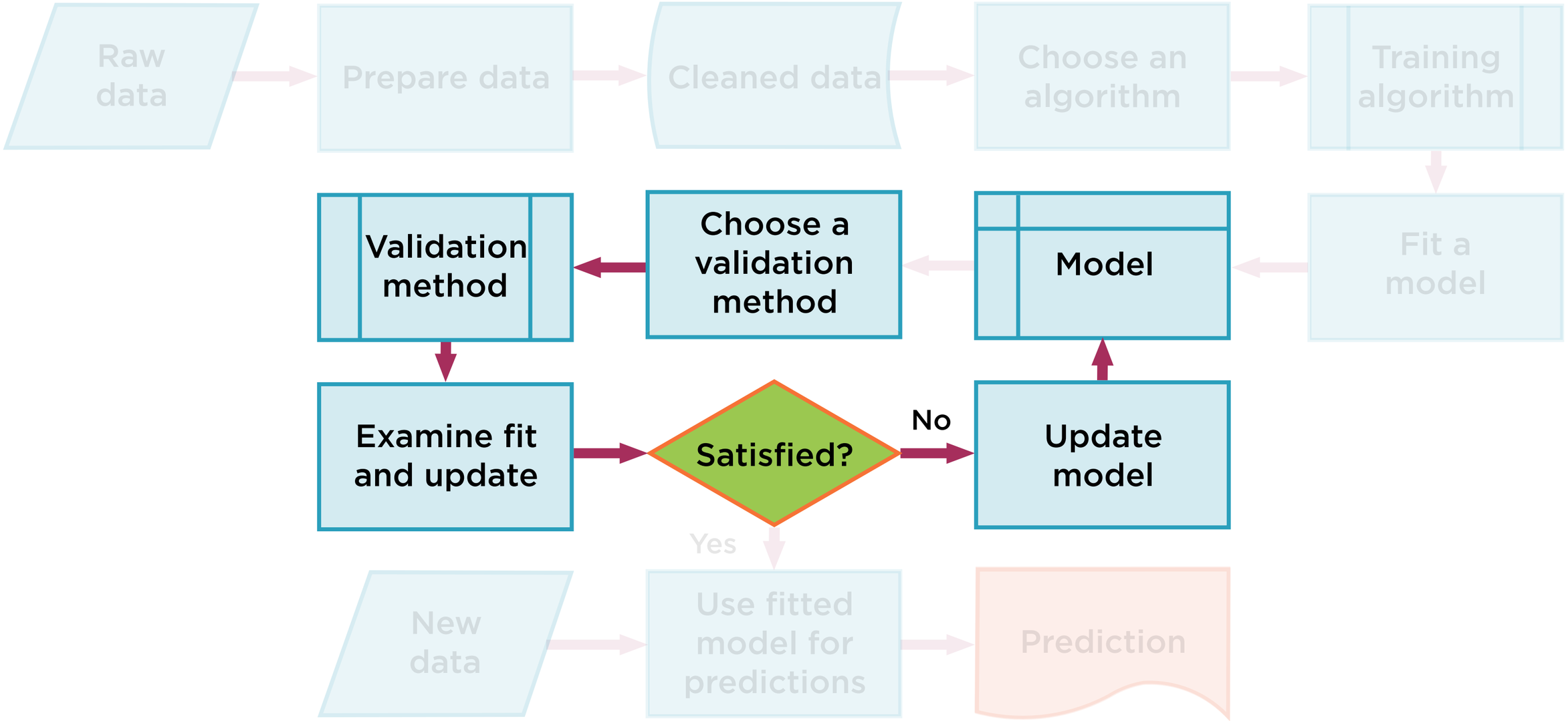
Metrics for Model Evaluation



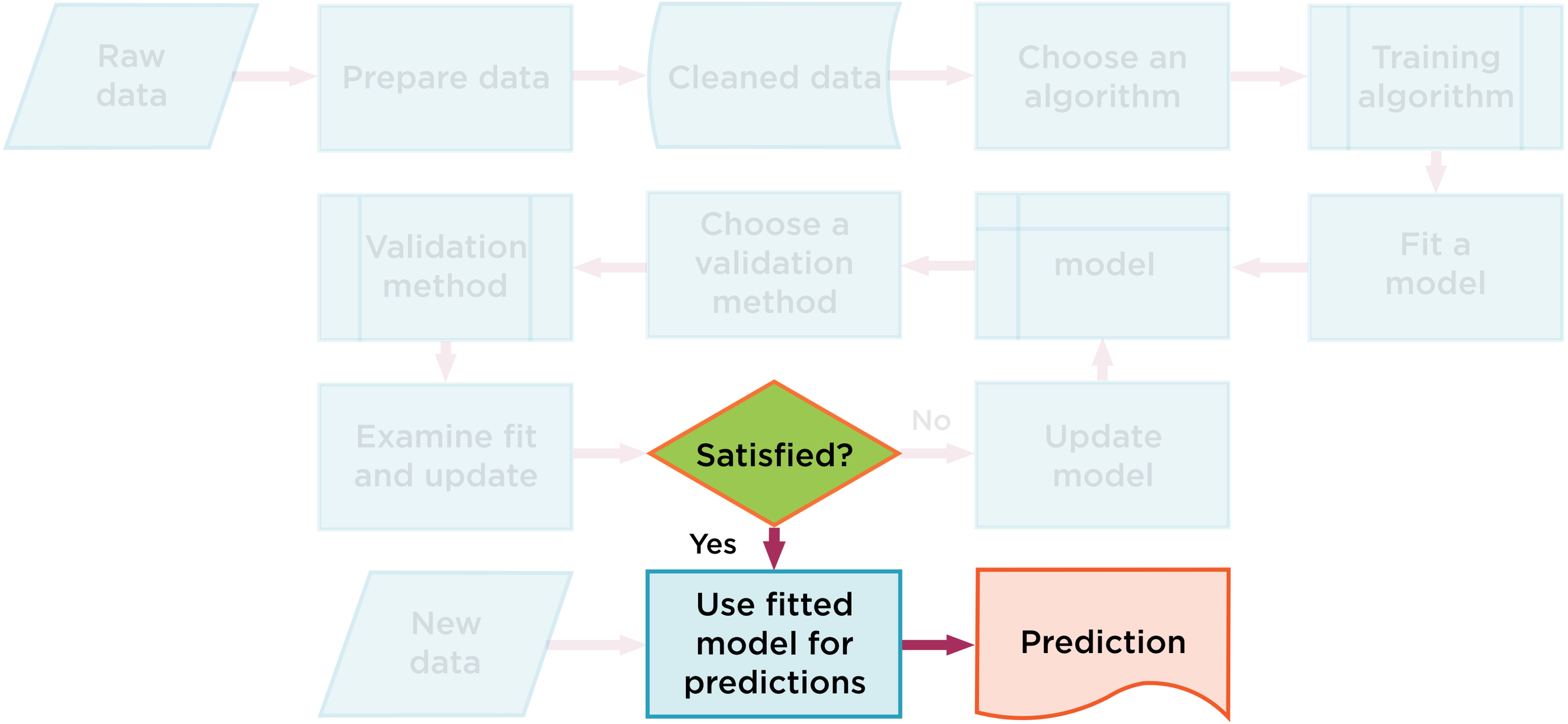
Rinse and Repeat



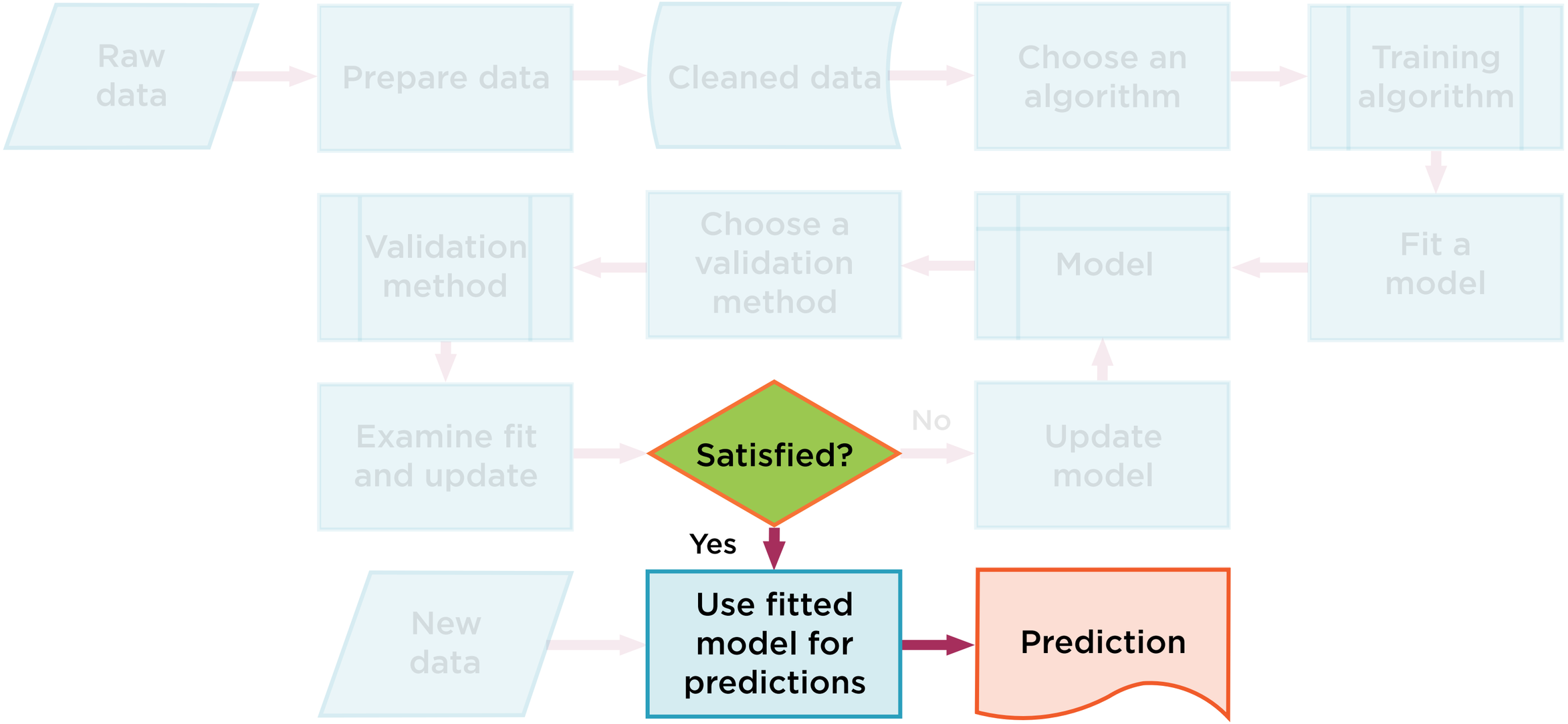
Rinse and Repeat



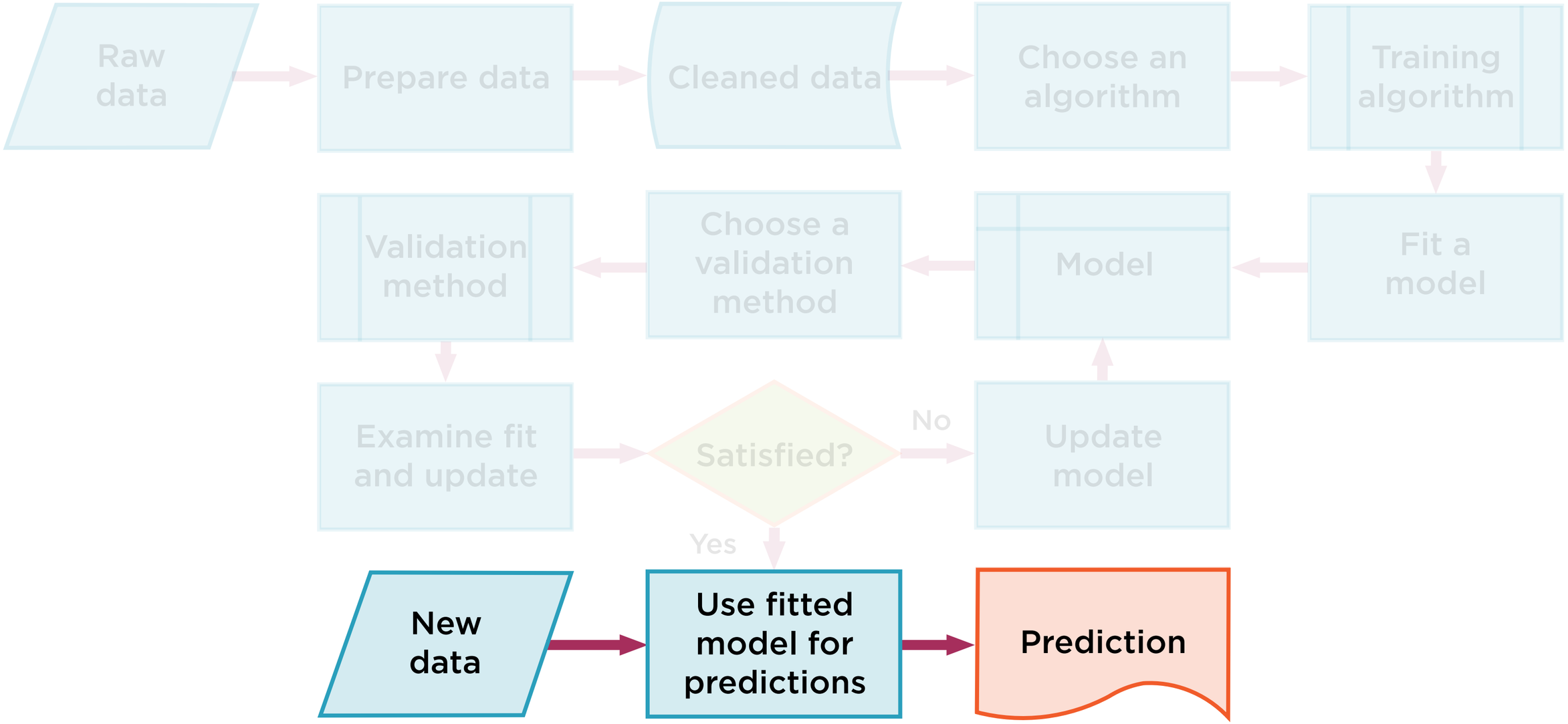
Fitted Model for Production Use



Invoke predict() Method



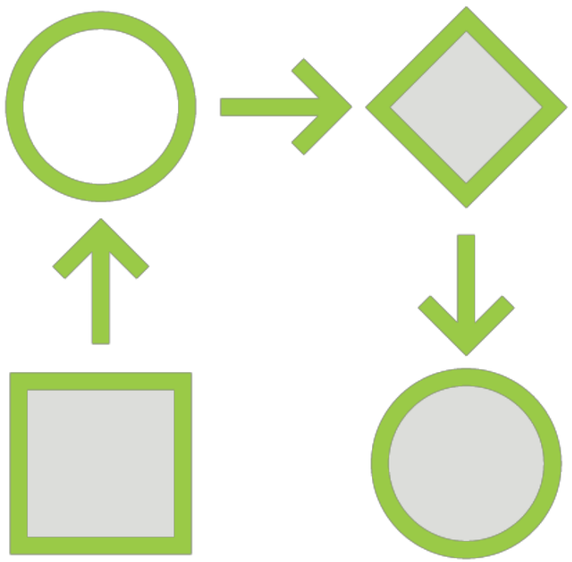
Invoke predict() Method



scikit-learn Pipeline

Estimator object that sequentially applies several transforms. Pipeline can be evaluated and tuned as a whole.

Pipelines



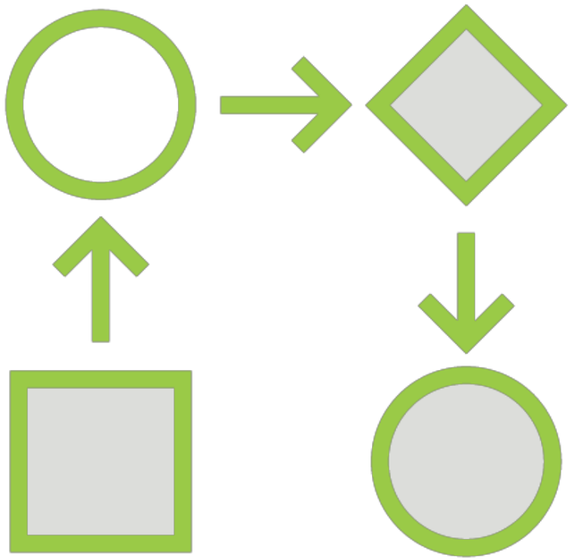
Pipeline objects are estimators too

Apply transforms sequentially

Return fitted estimator

Final output only implements fit

Pipelines



Intermediate transforms can be cached

Easily tune pipeline as a whole

Cross-validation

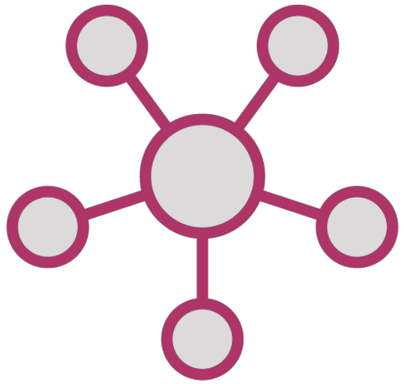
Switch in or switch out individual steps

Choosing the Right Estimator

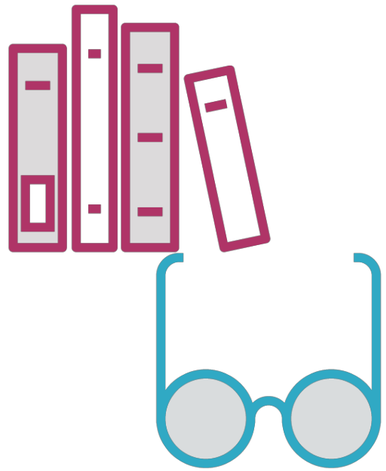
Types of Machine Learning Problems

Classification

Regression



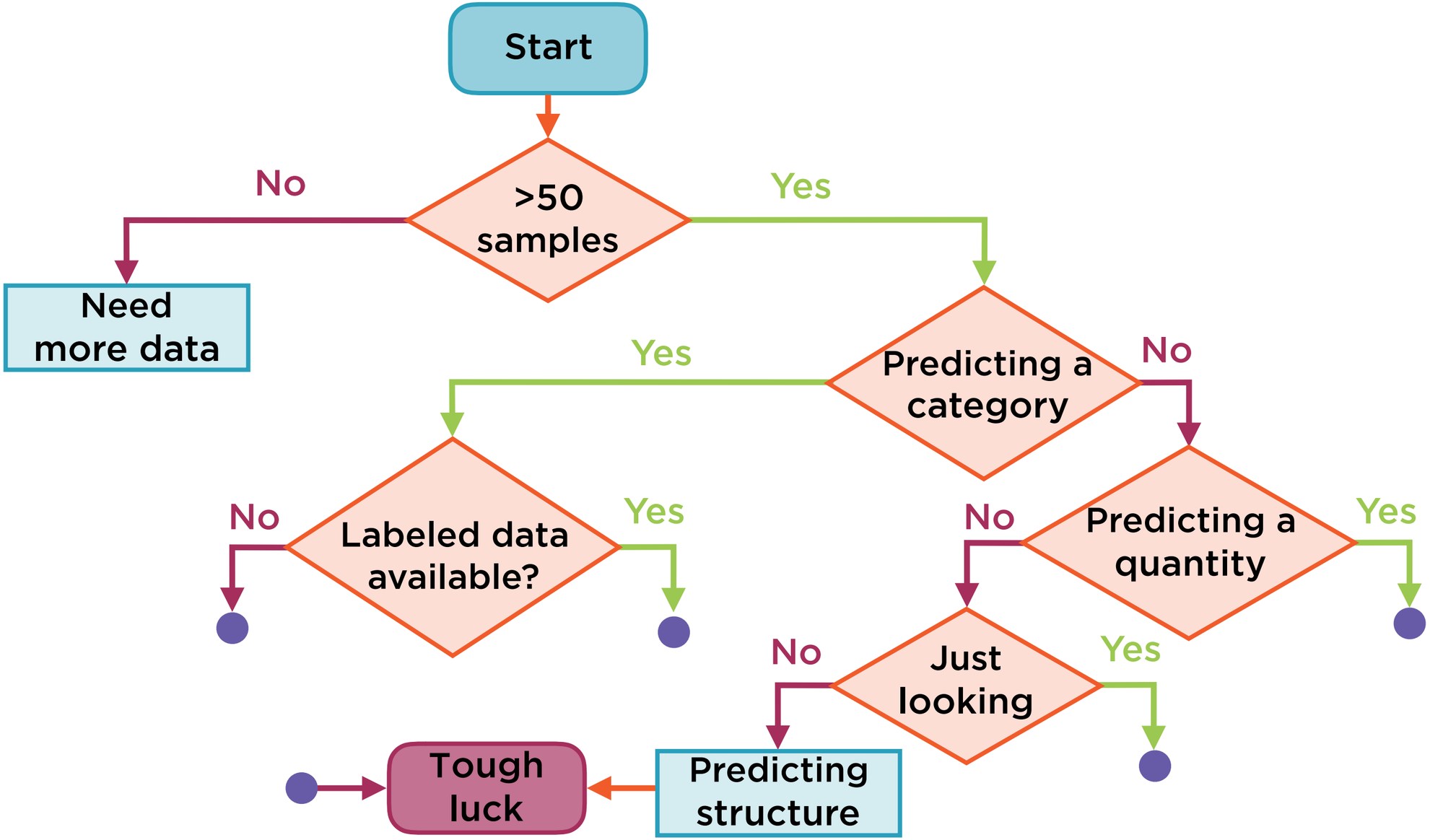
Clustering



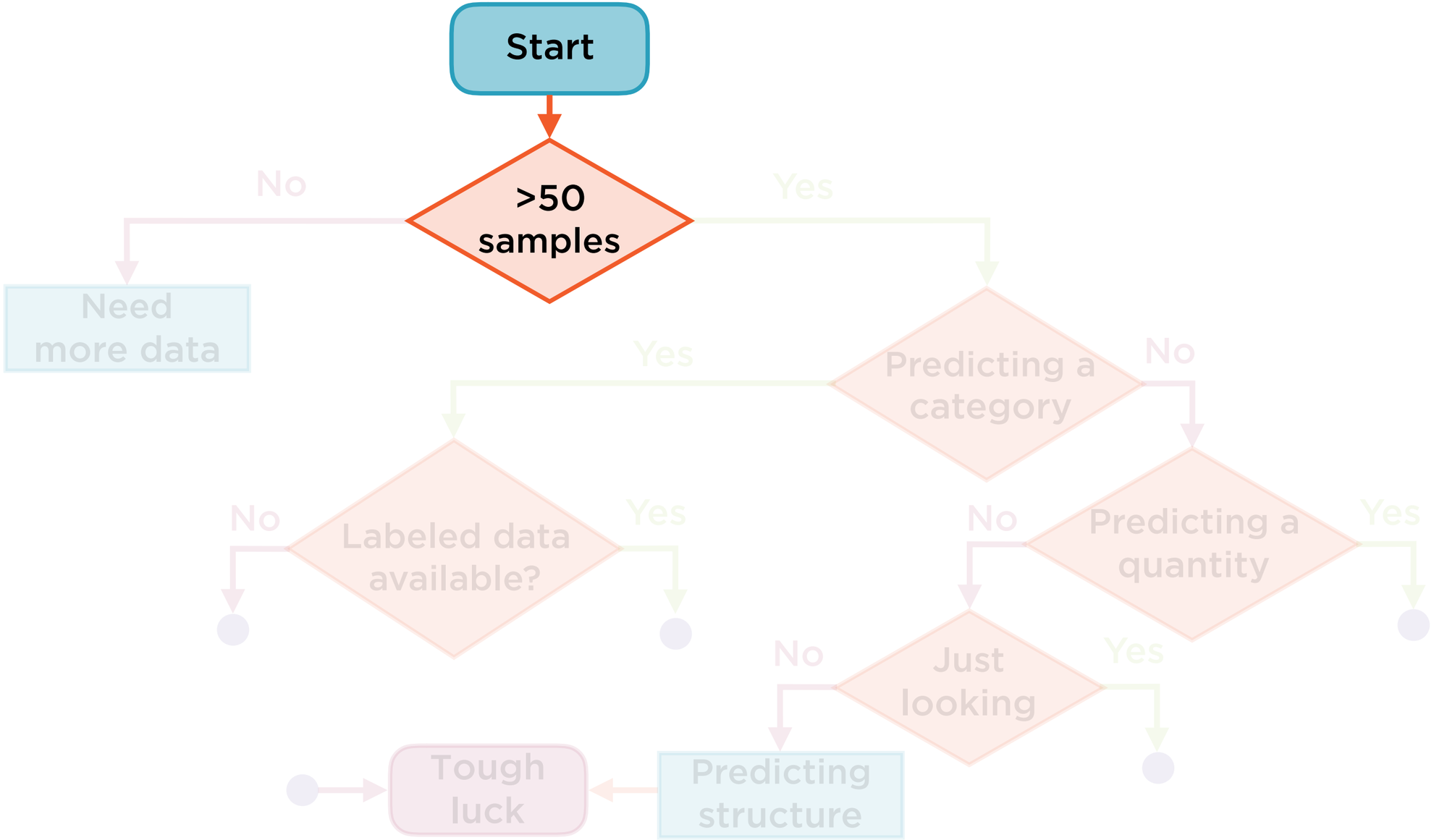
Dimensionality reduction

Focus first on defining the right problem to solve, then on choosing the right estimator to solve it

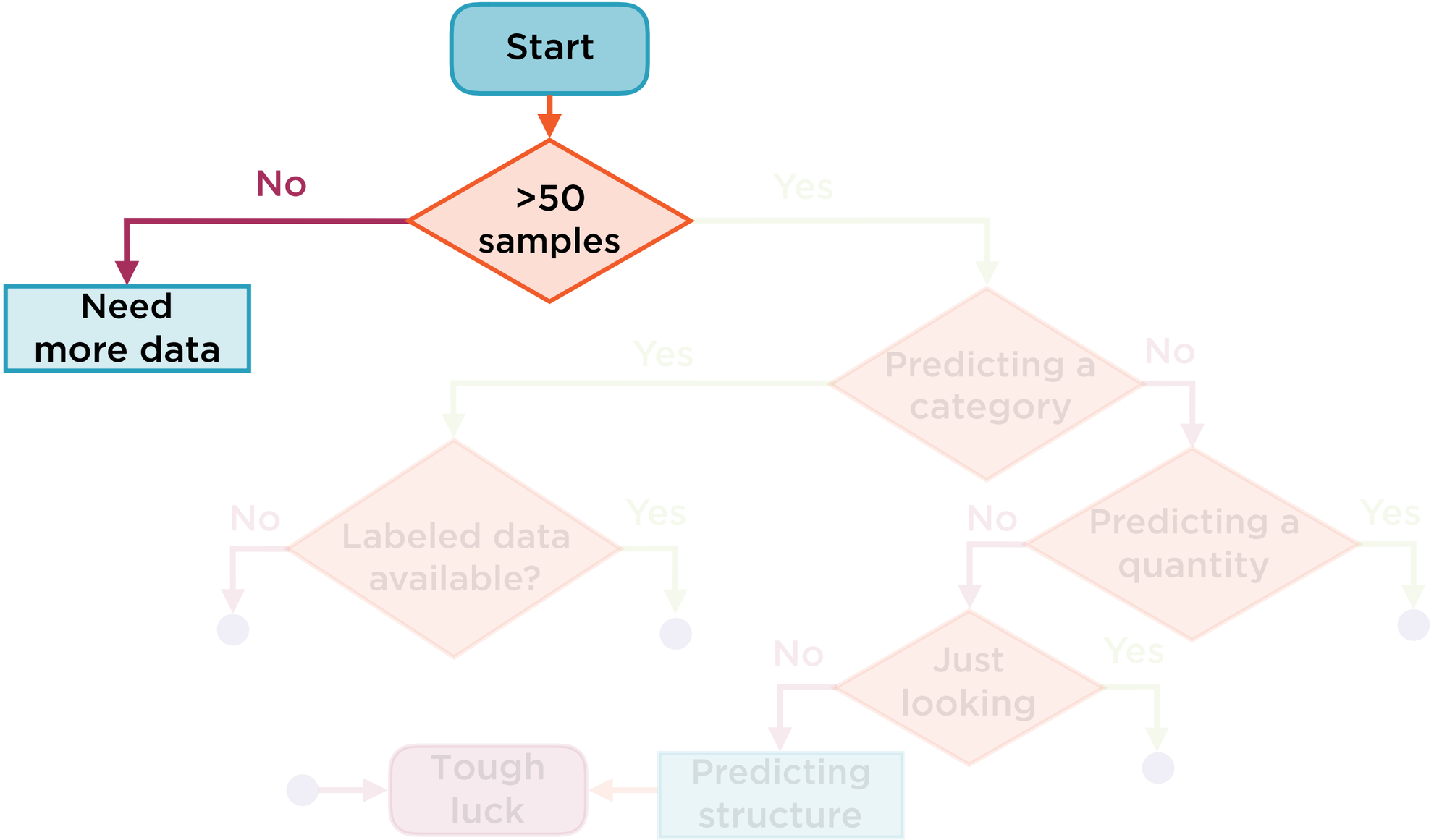
Choosing the Right Estimator



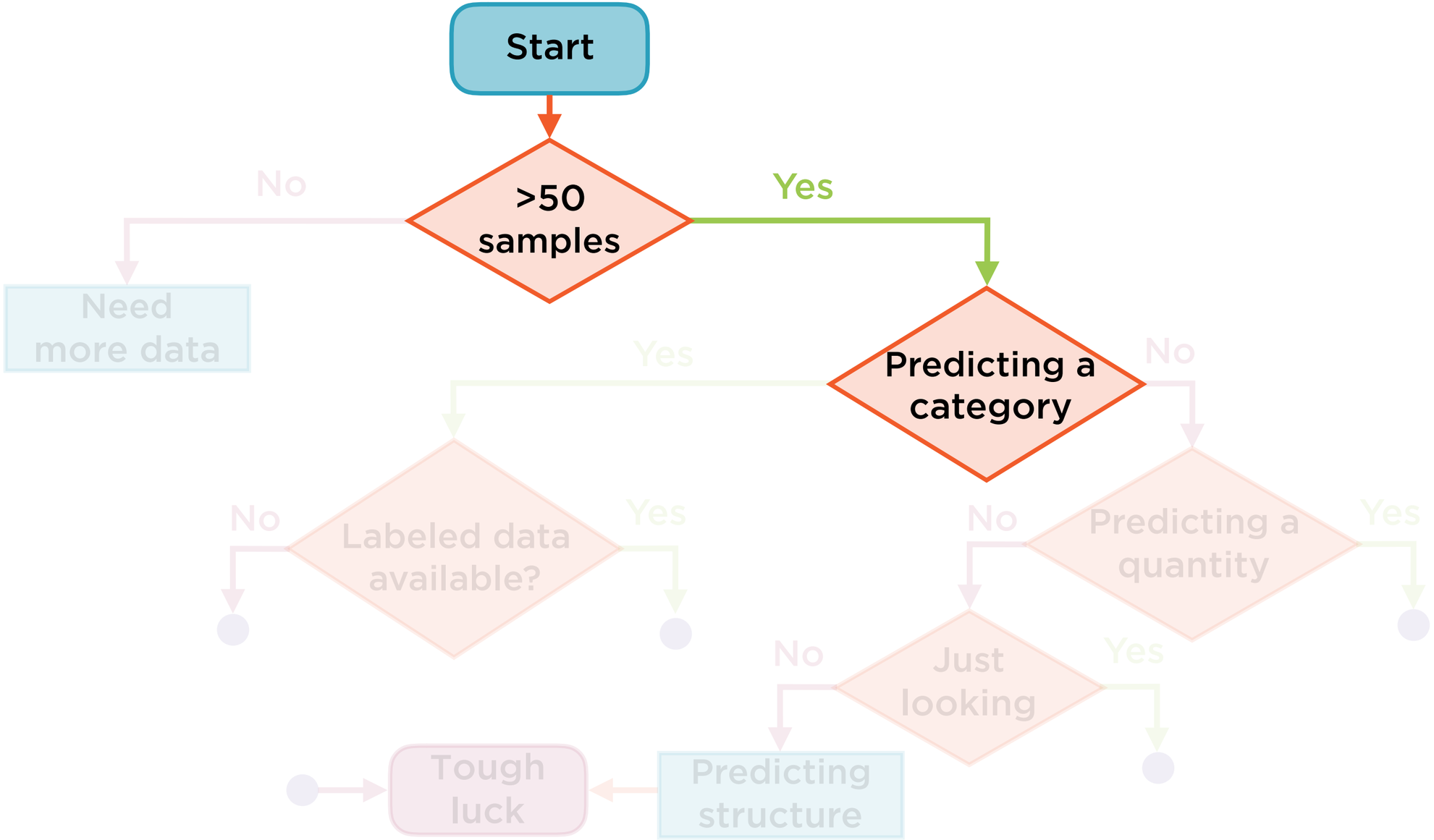
Choosing the Right Estimator



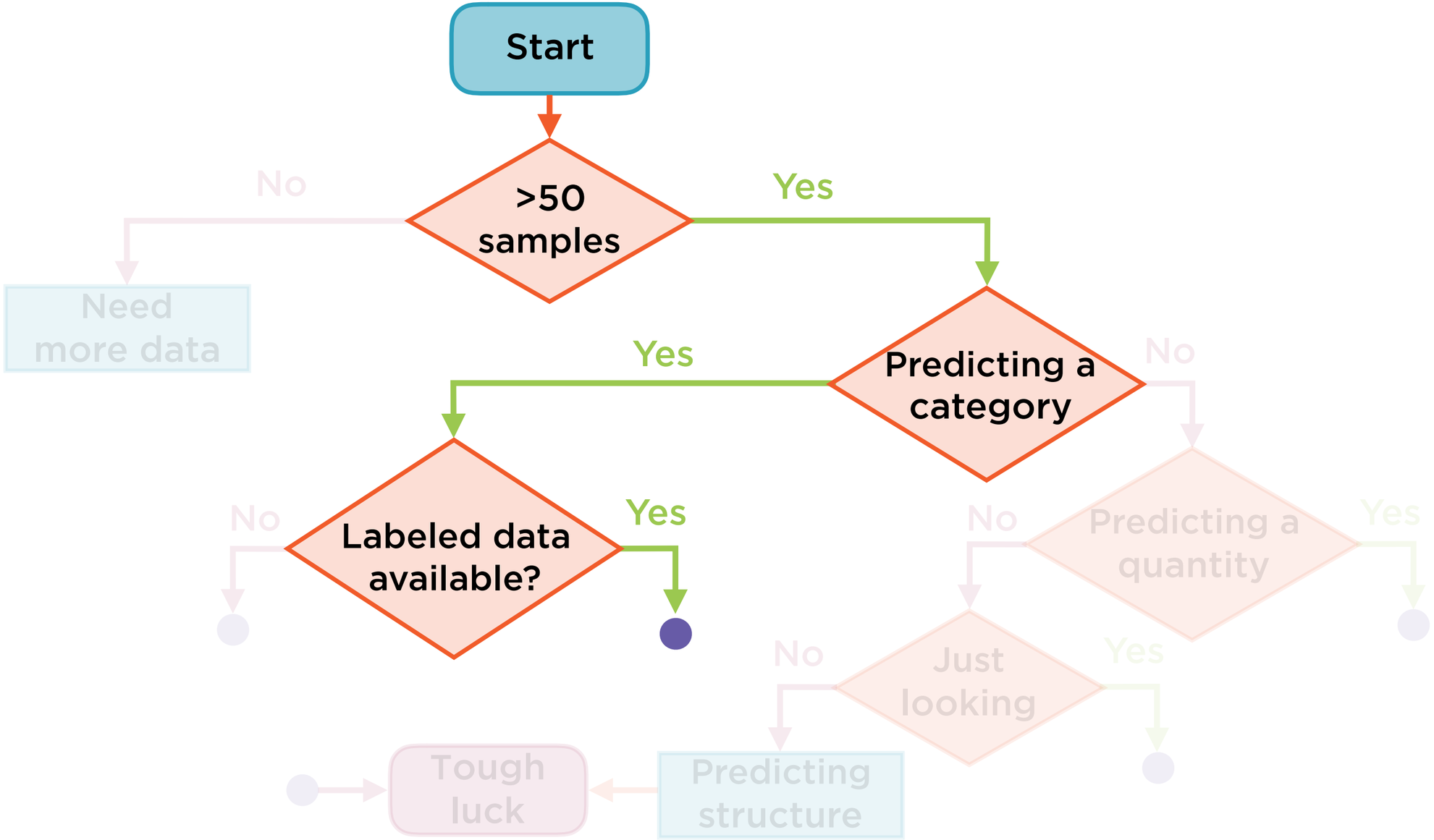
Choosing the Right Estimator



Choosing the Right Estimator



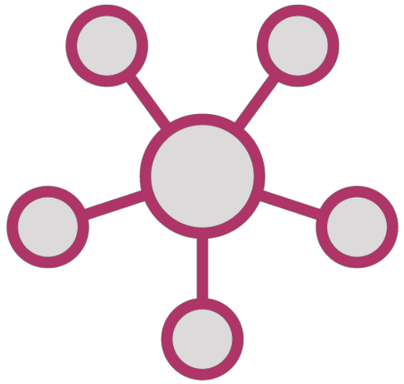
Choosing the Right Estimator



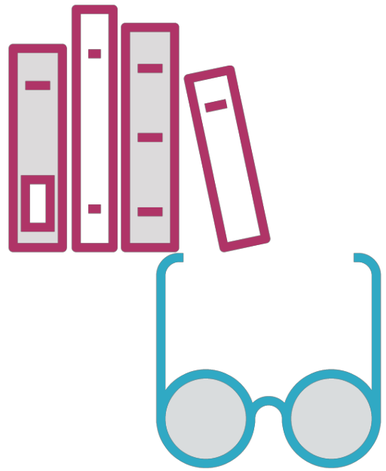
Types of Machine Learning Problems

Classification

Regression



Clustering



Dimensionality reduction

Types of Machine Learning Problems

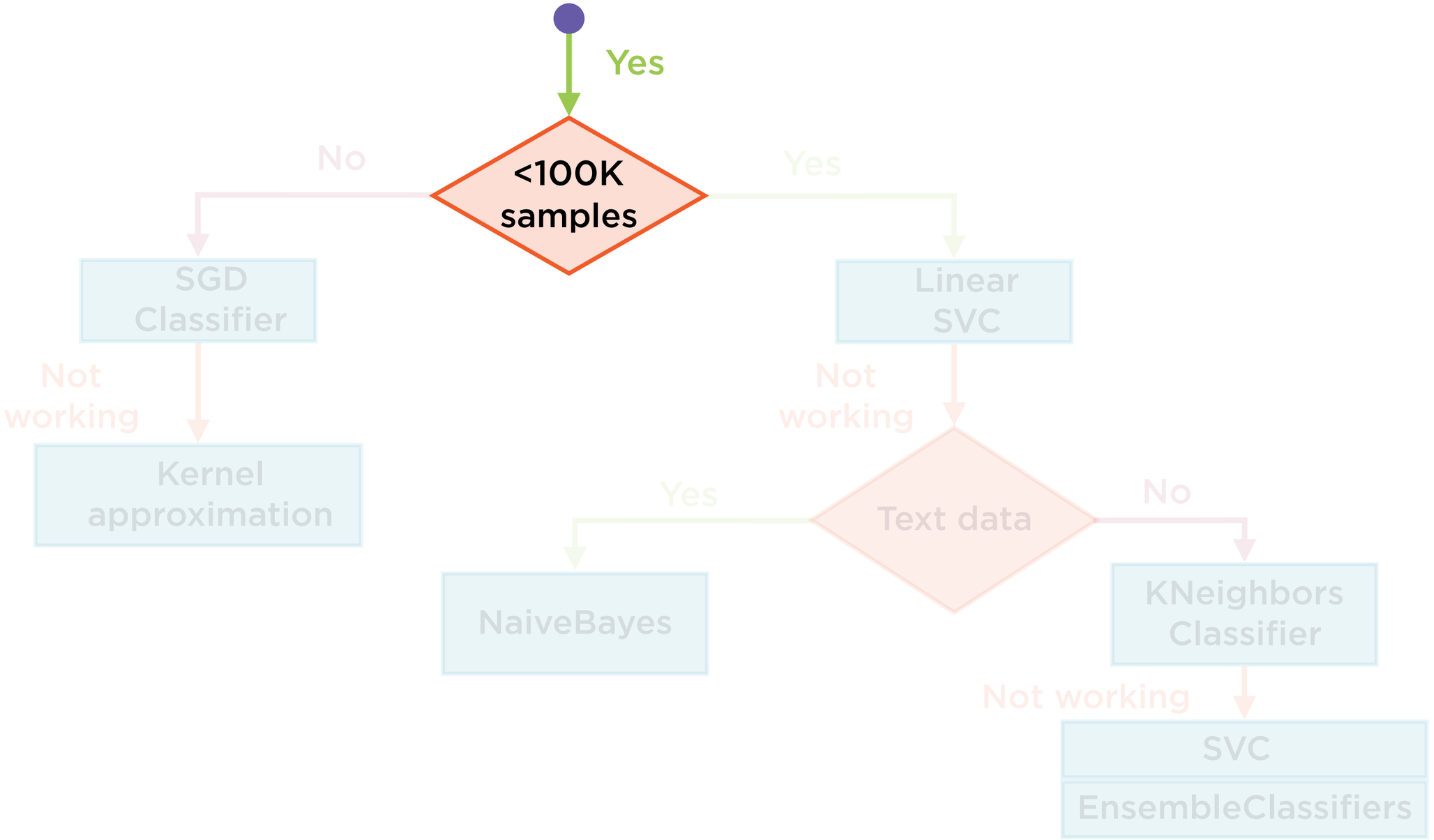
Classification

Regression

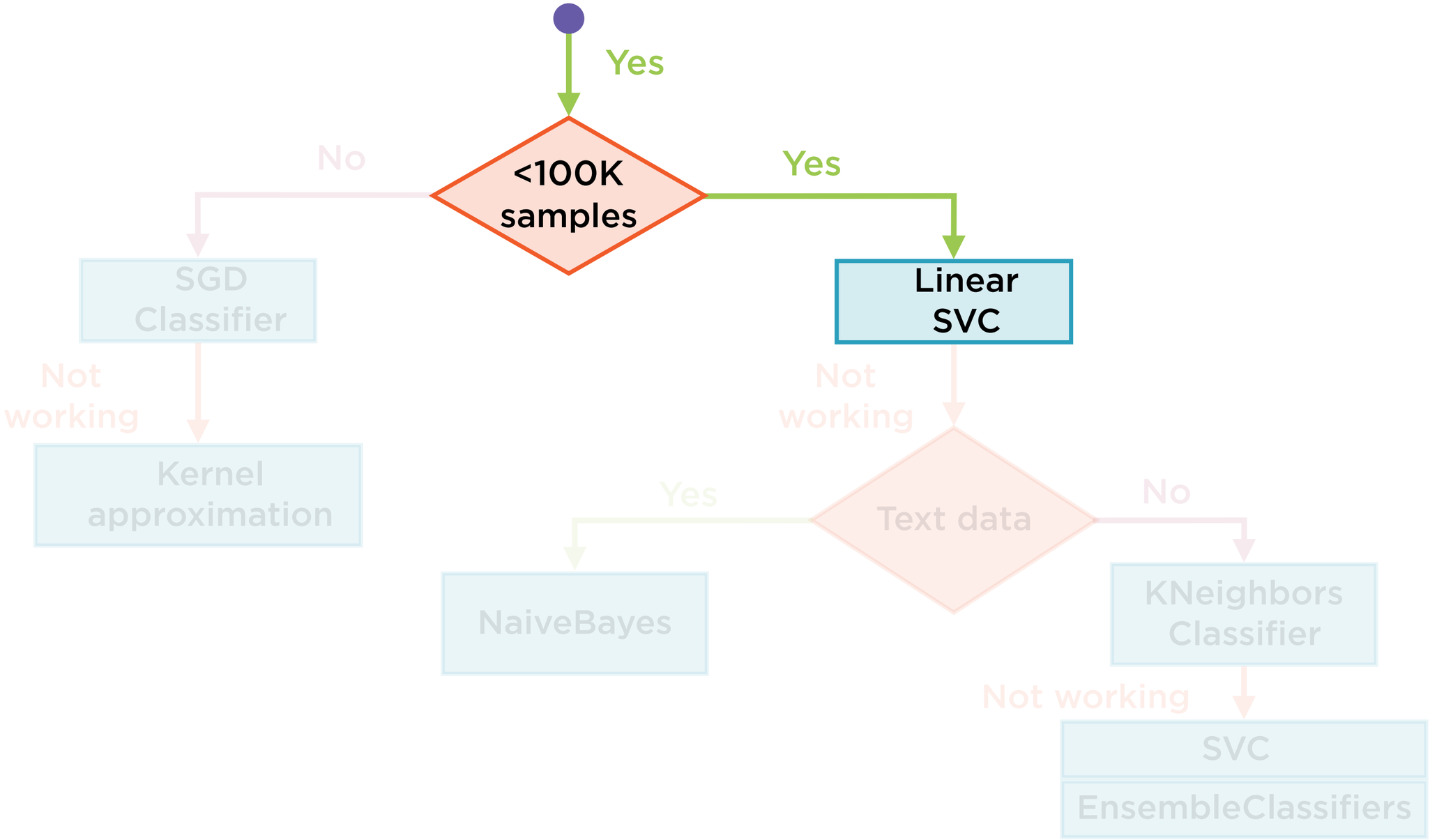
Clustering

Dimensionality reduction

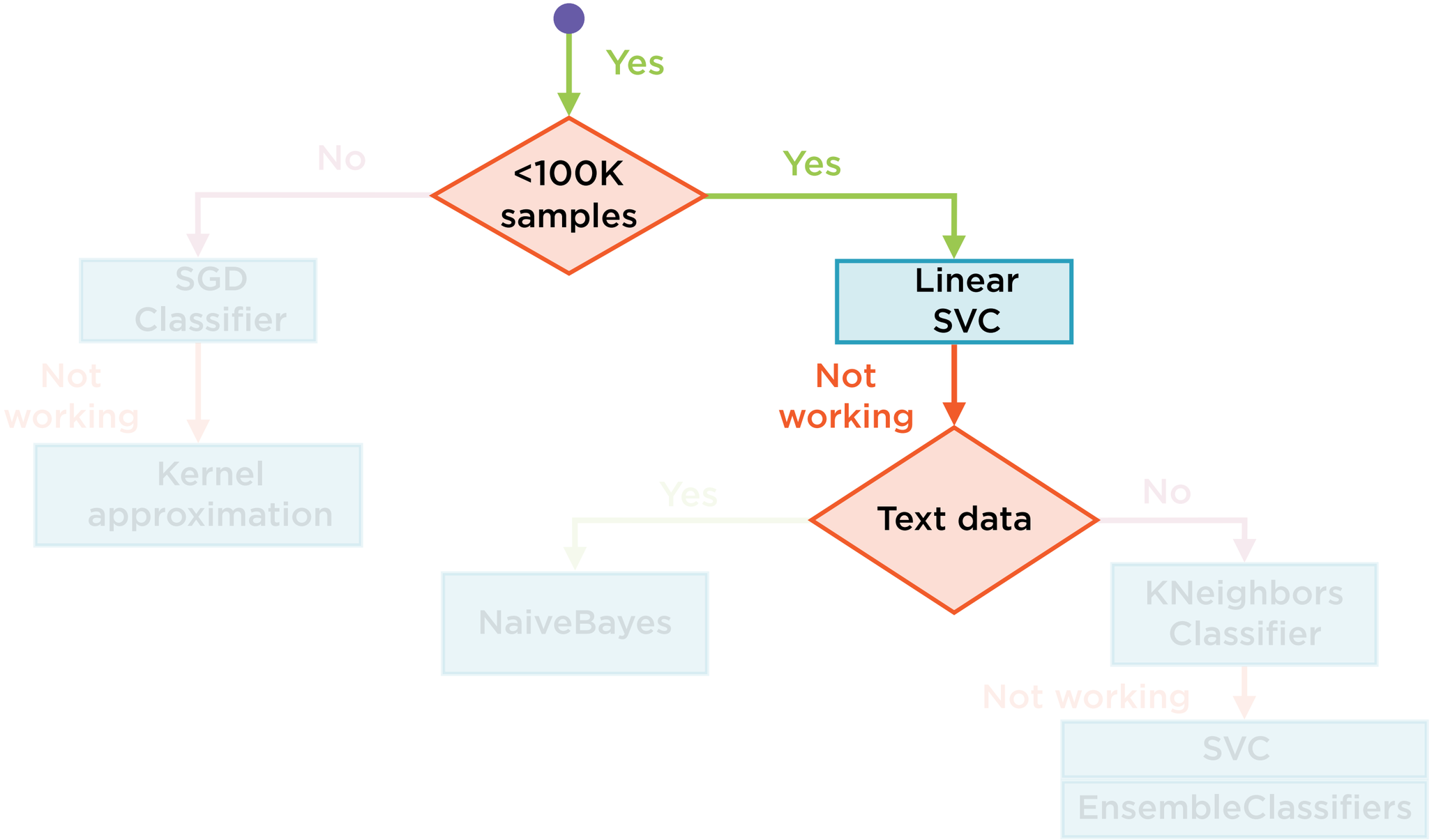
Classification



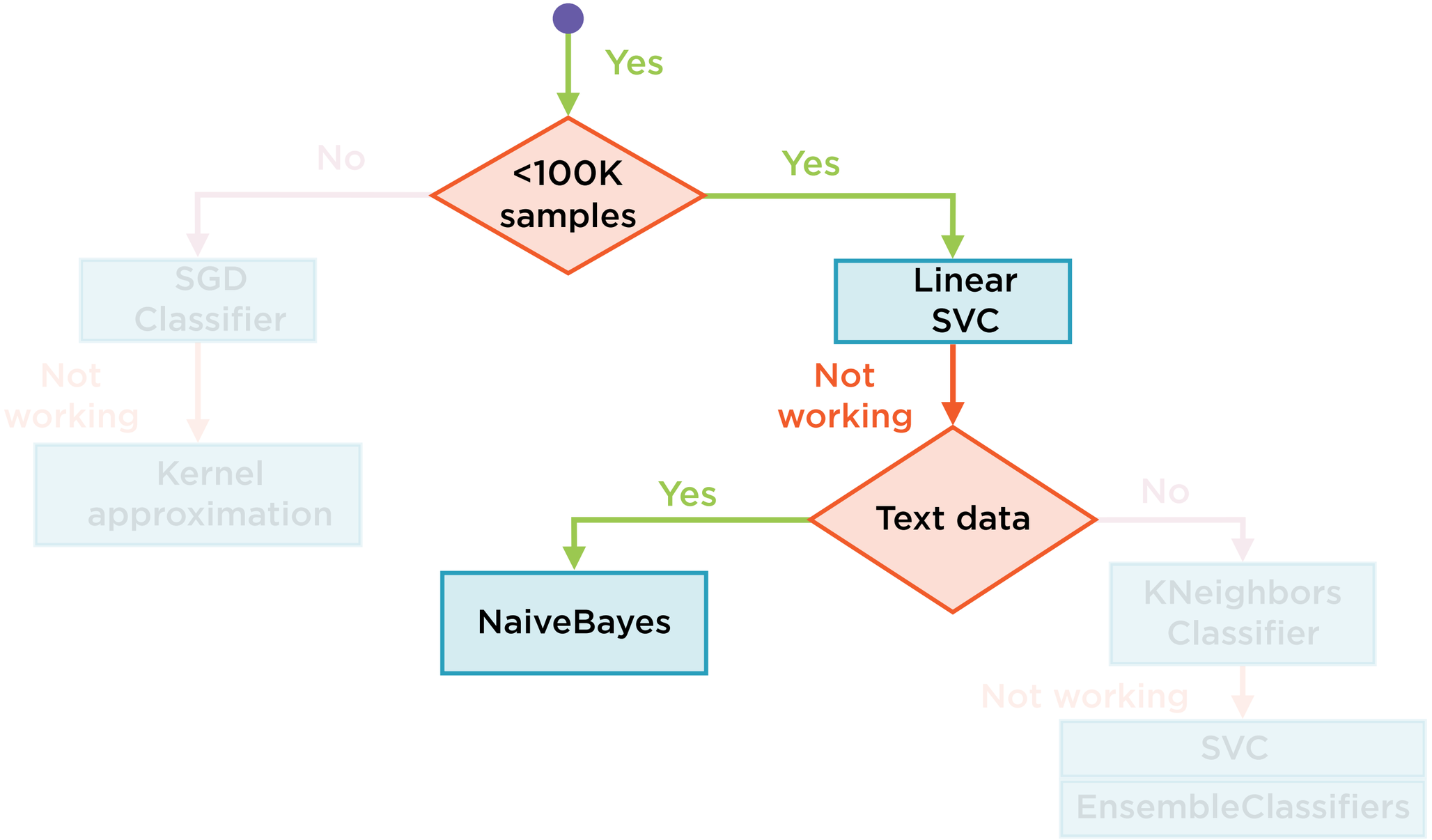
Classification



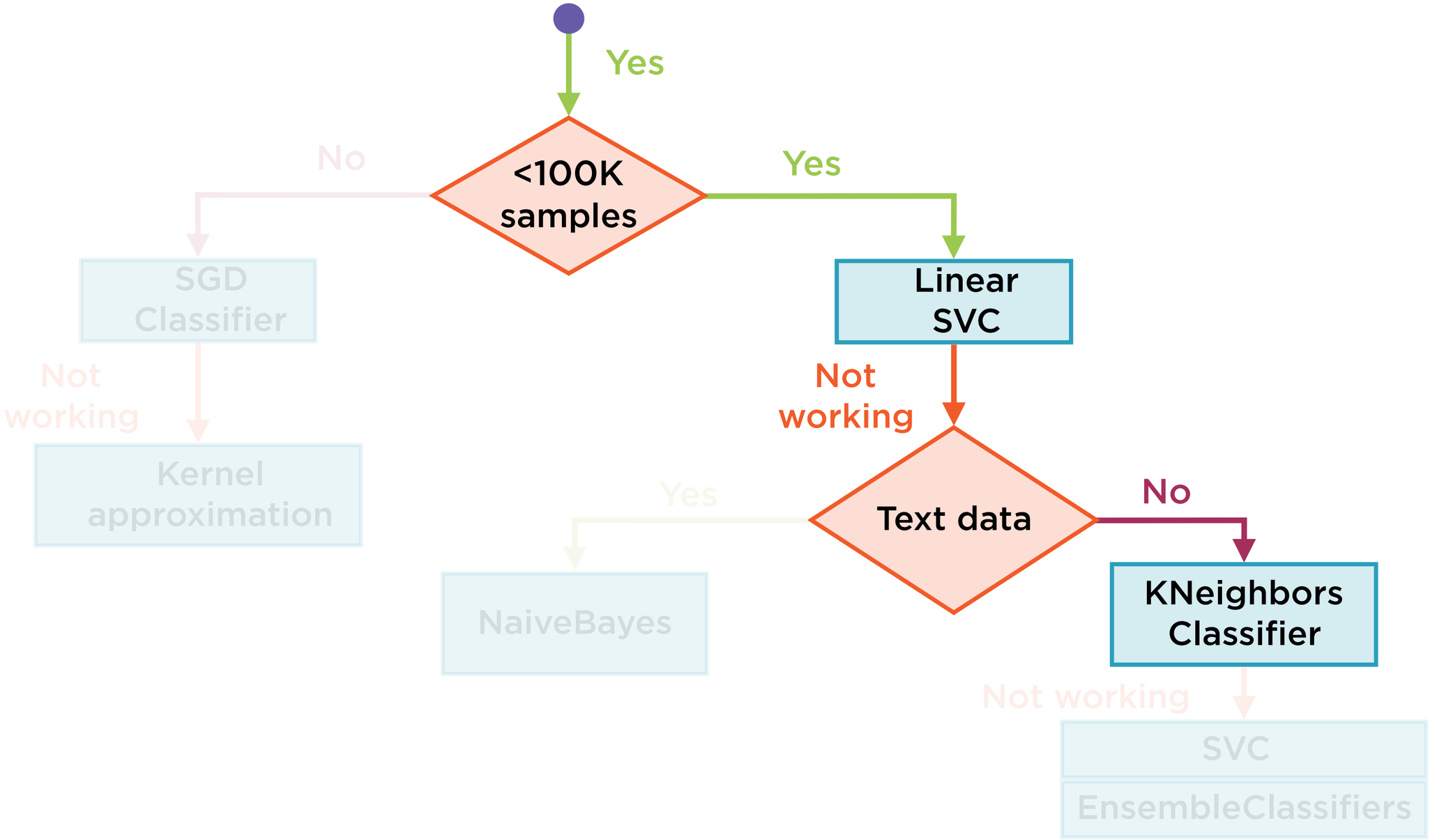
Classification



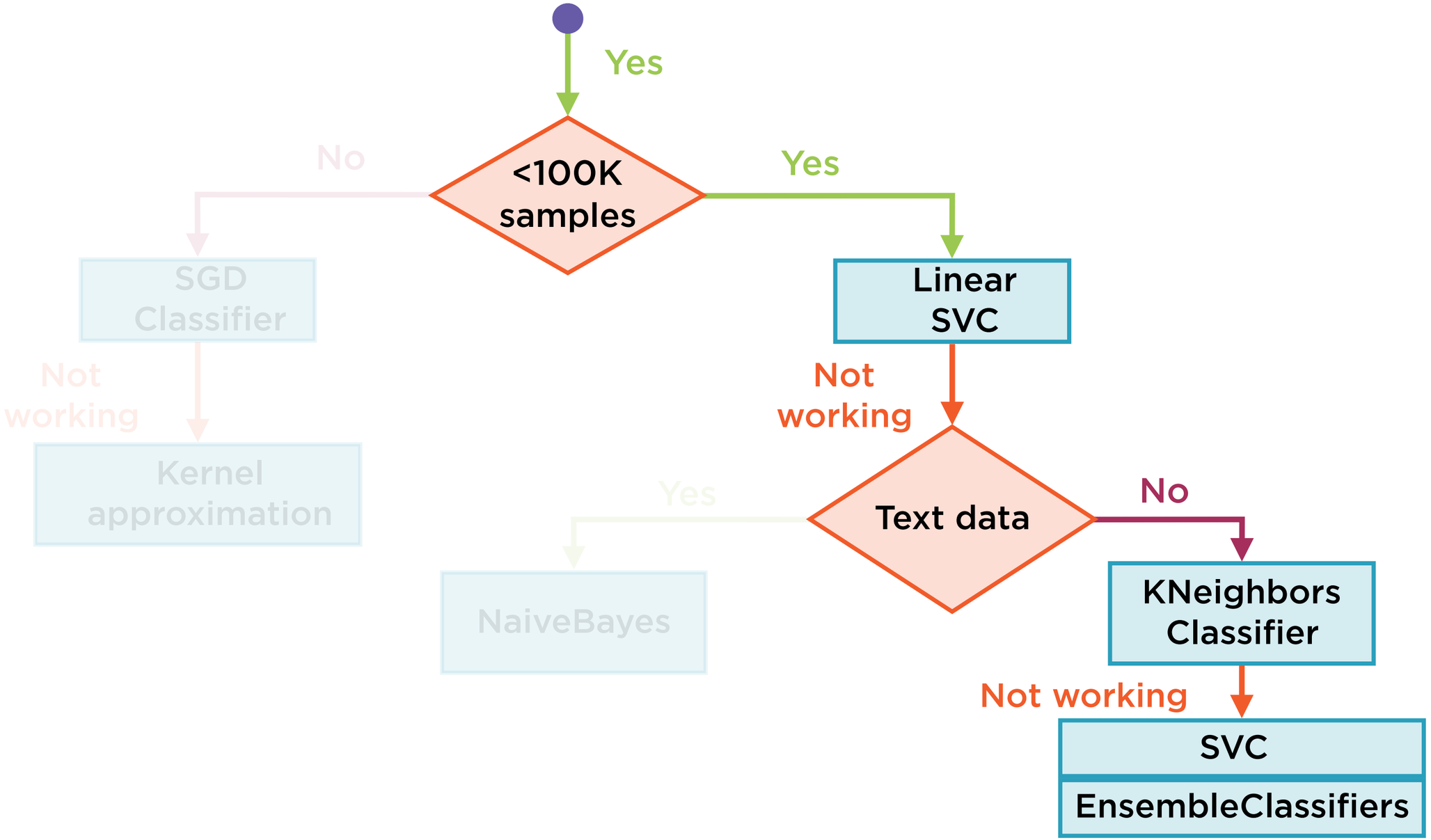
Classification



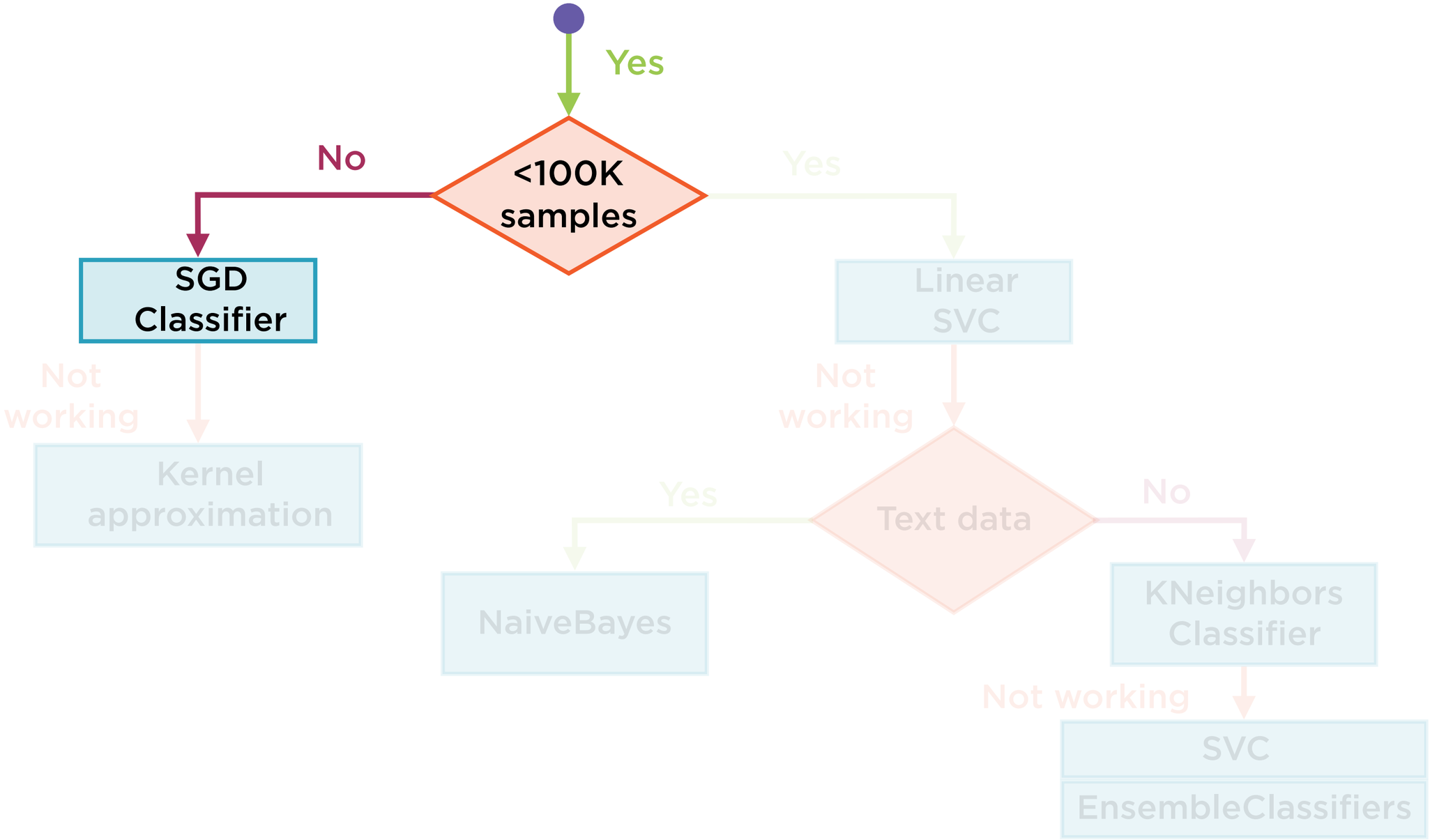
Classification



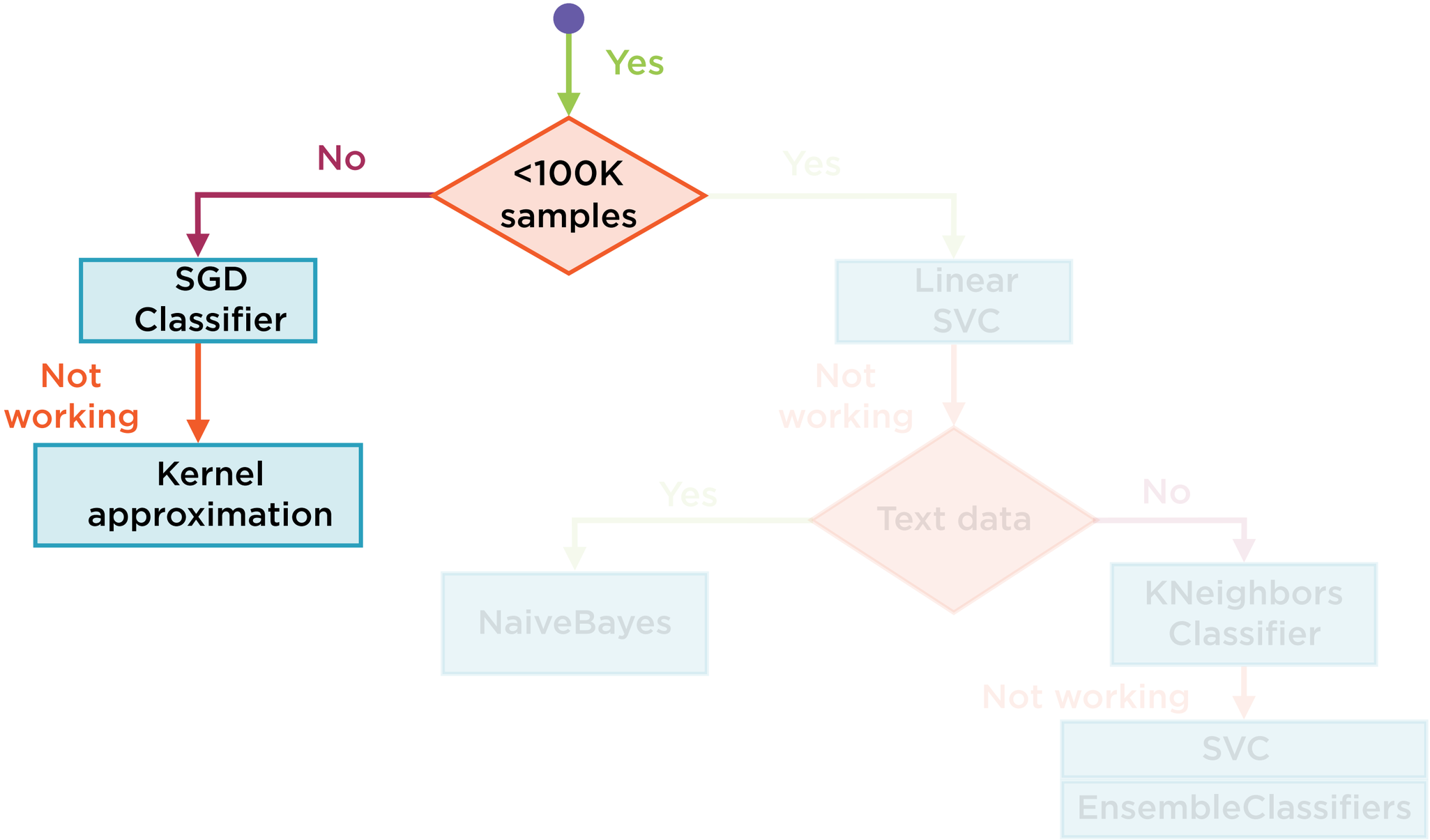
Classification



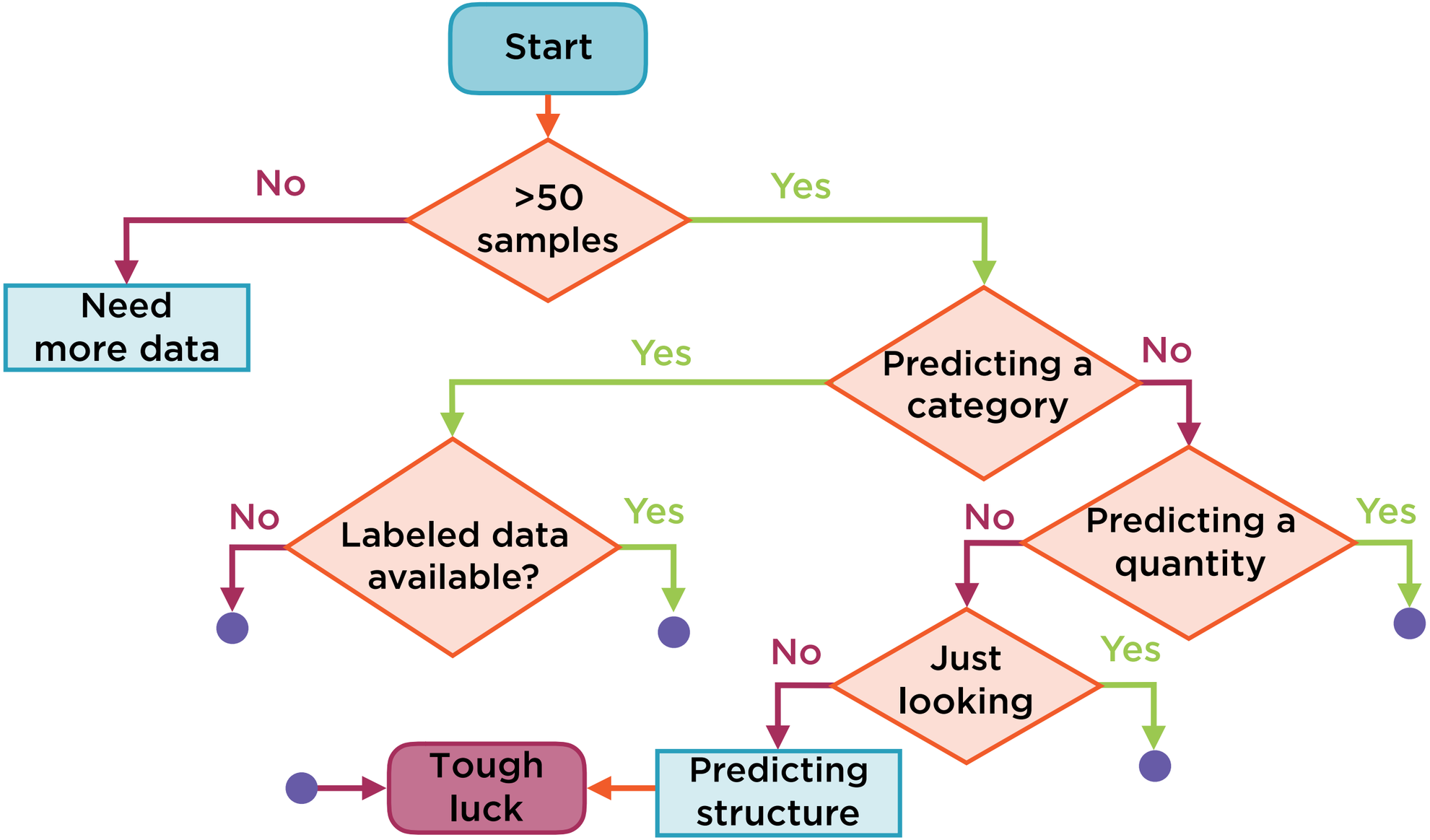
Classification



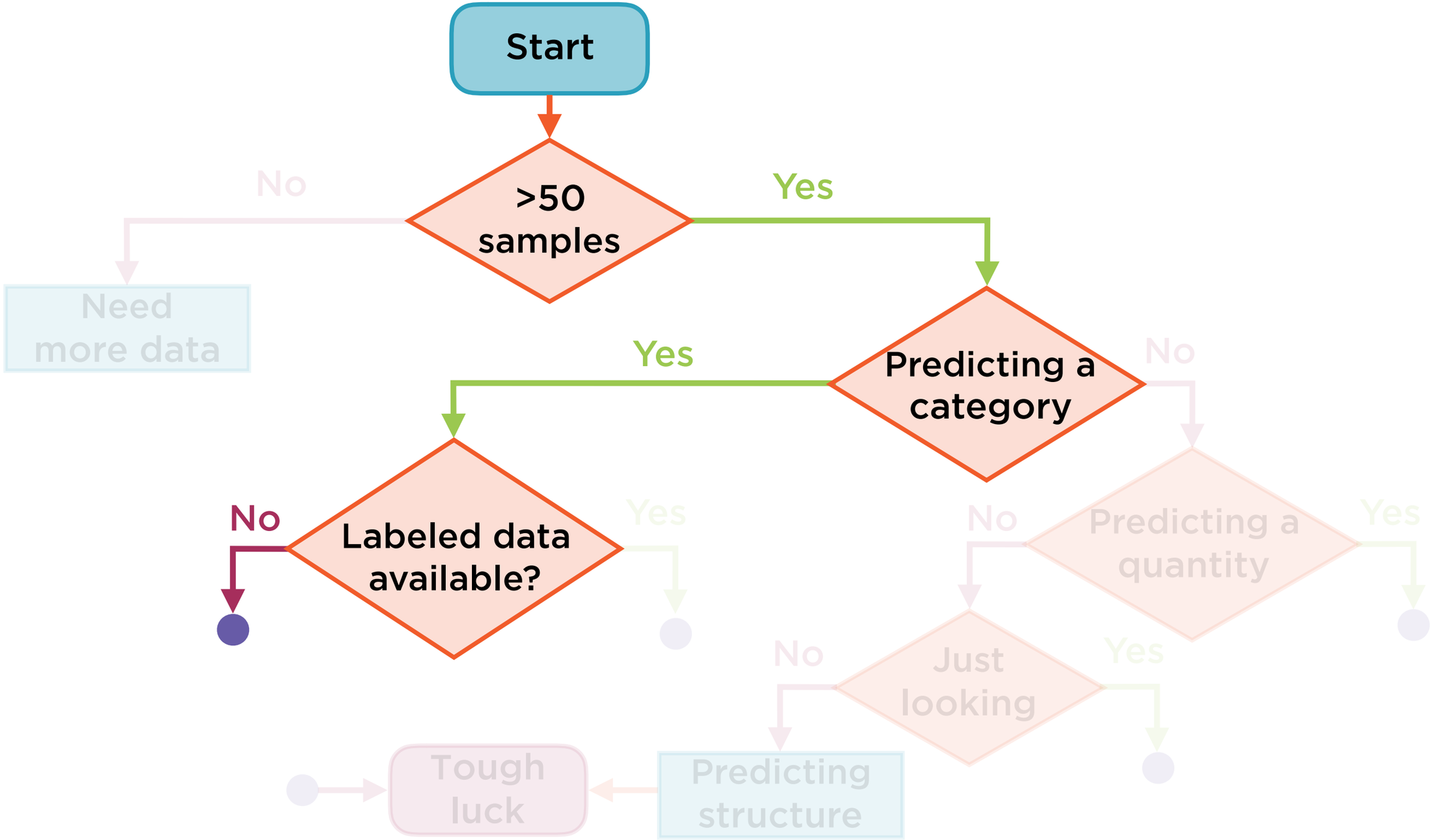
Classification



Choosing the Right Estimator



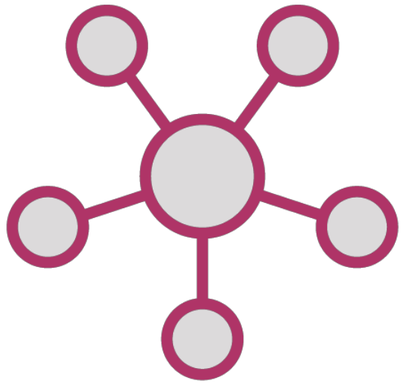
Choosing the Right Estimator



Types of Machine Learning Problems

Classification

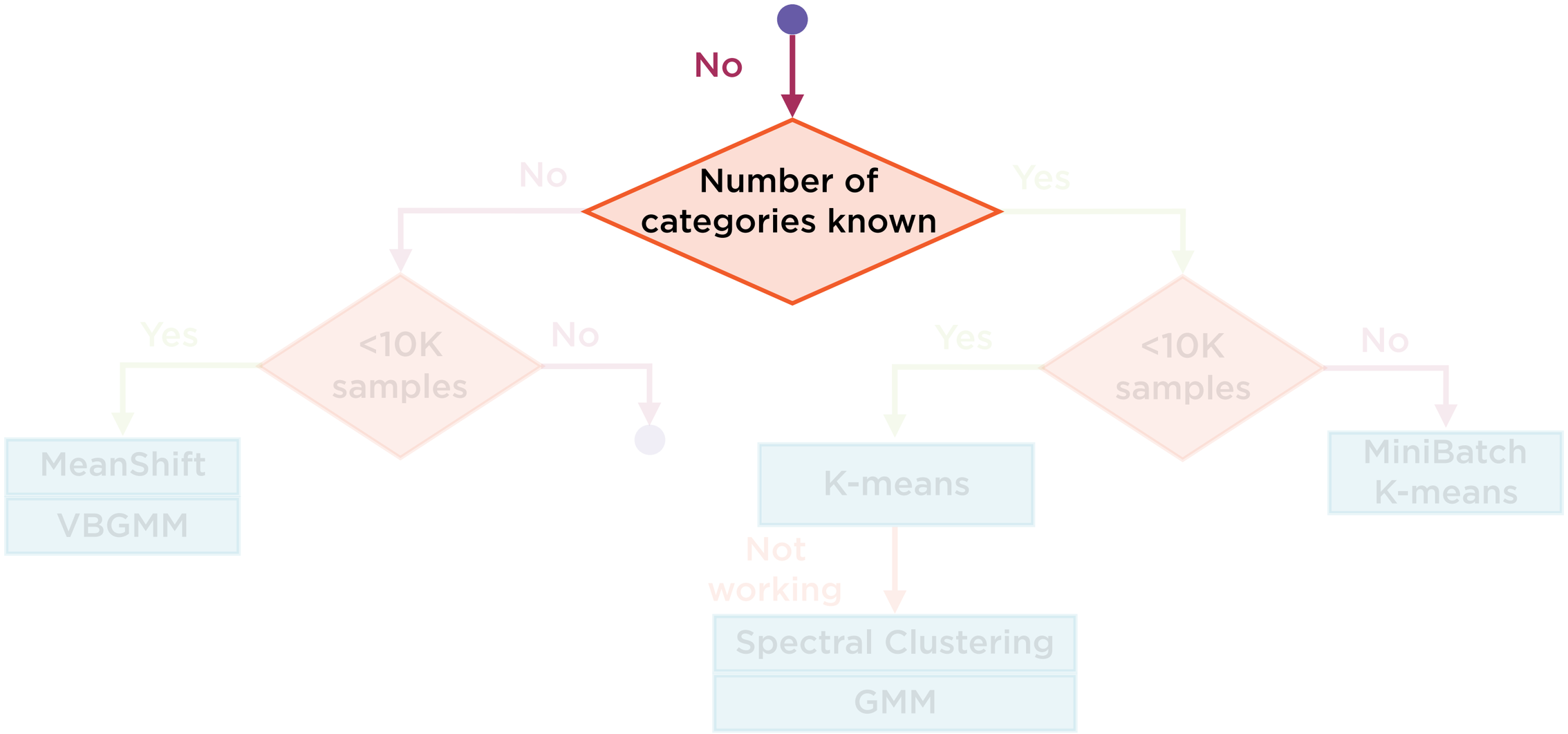
Regression



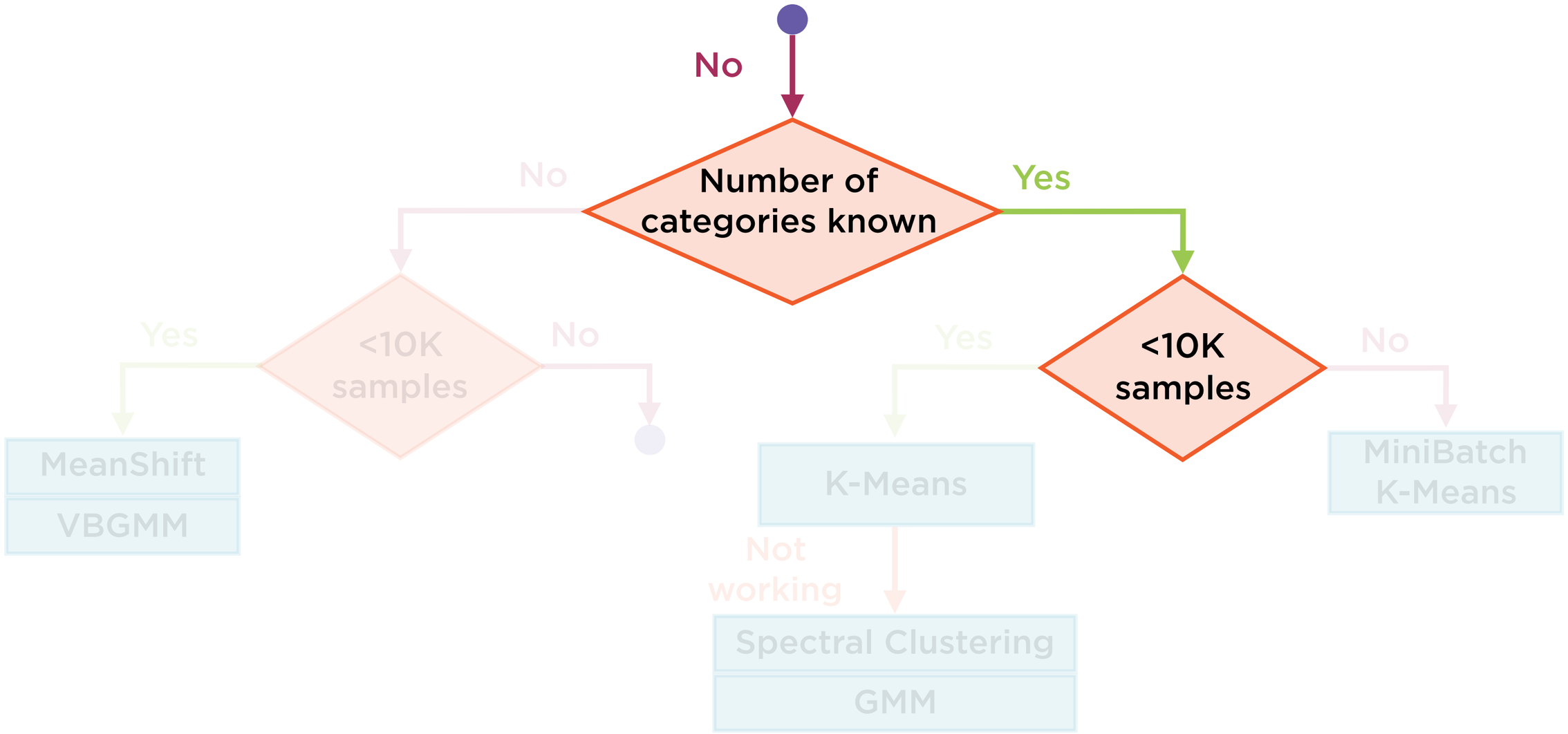
Clustering

Dimensionality reduction

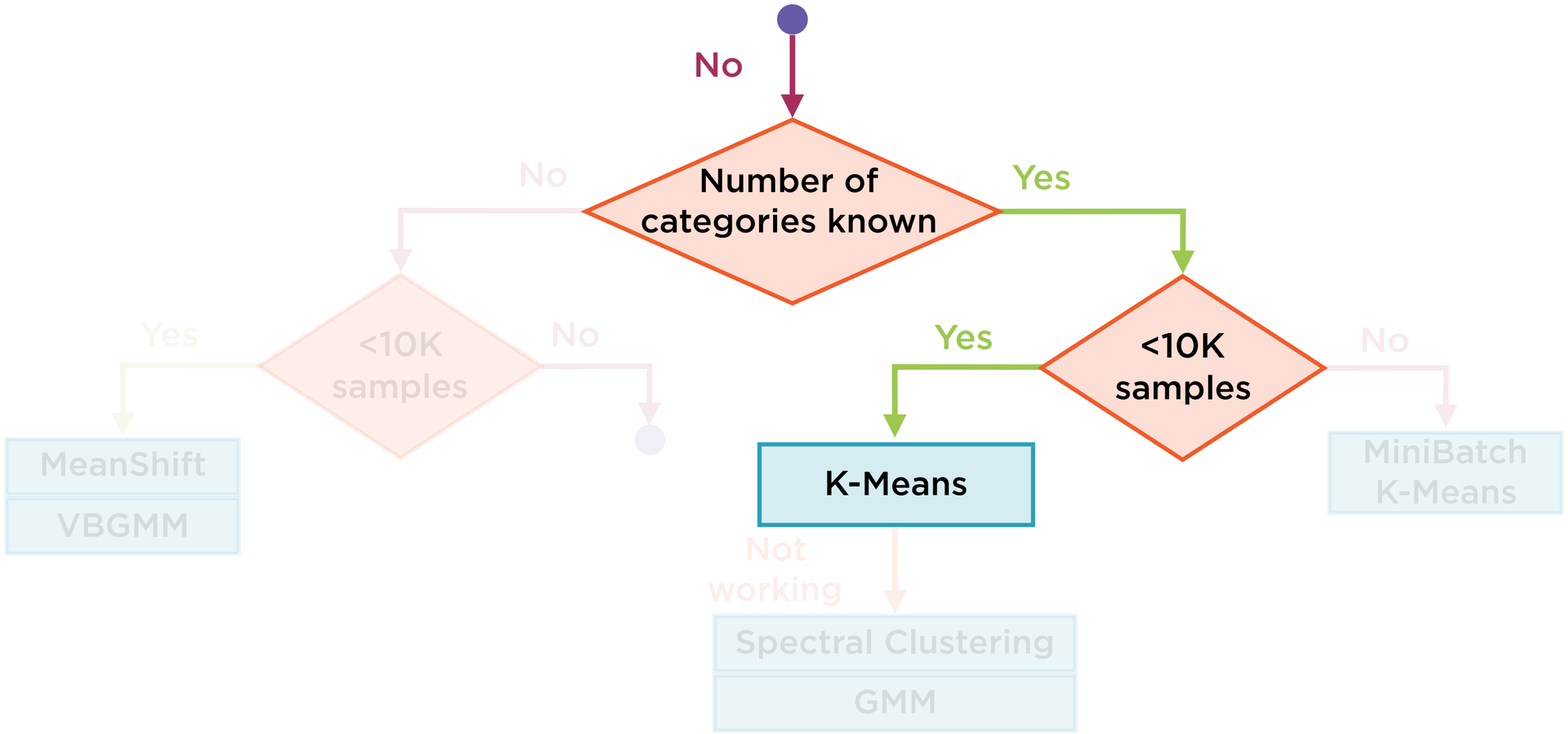
Clustering


No
Number of categories known
No
Yes
<10K samples
Yes
No
MeanShift
VBGMM
<10K samples
Yes
No
K-means
MiniBatch K-means
Not working
Spectral Clustering
GMM

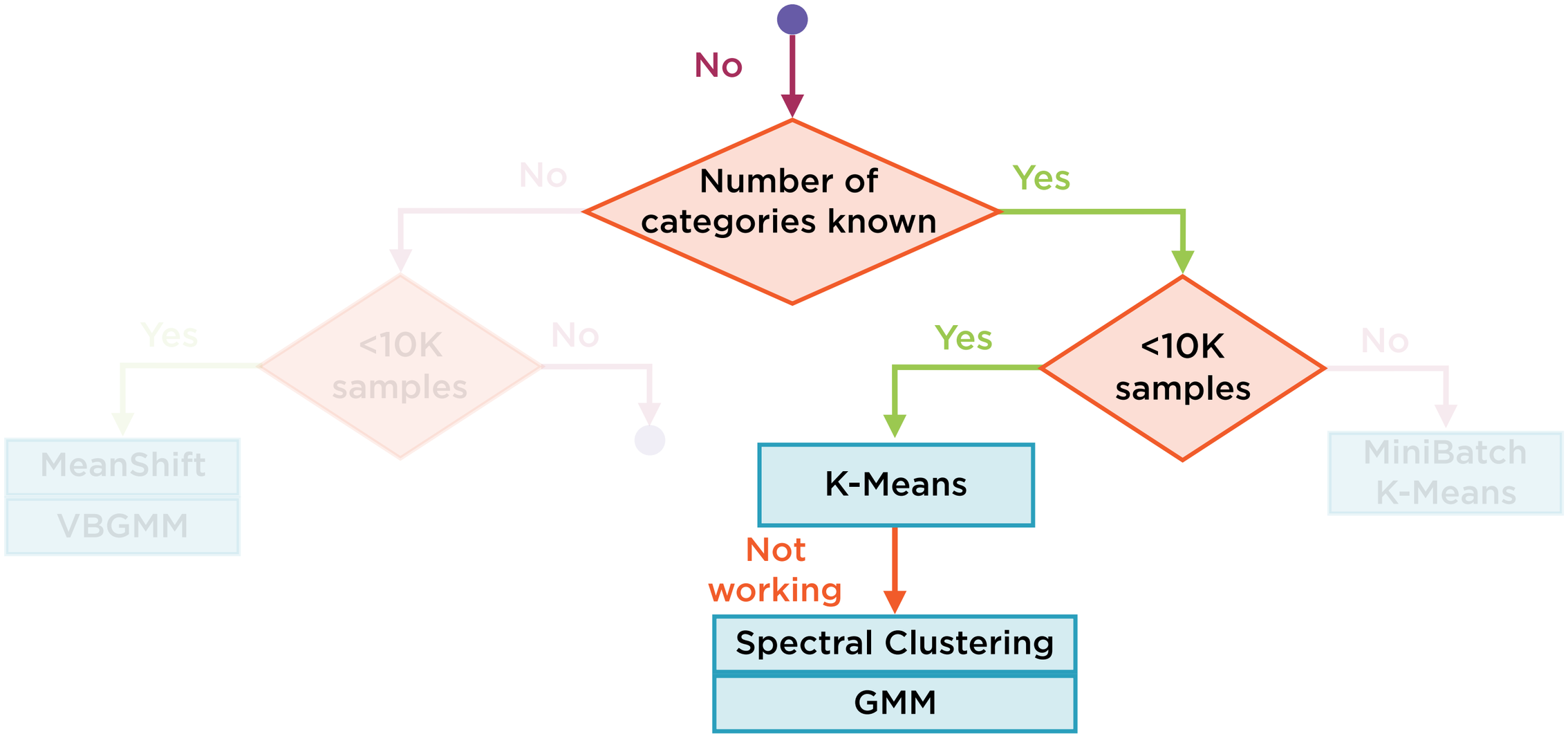
Clustering


No
Number of categories known
No
Yes
<10K samples
Yes
No
MeanShift
VBGMM
<10K samples
Yes
No
K-Means
MiniBatch K-Means
Not working
Spectral Clustering
GMM

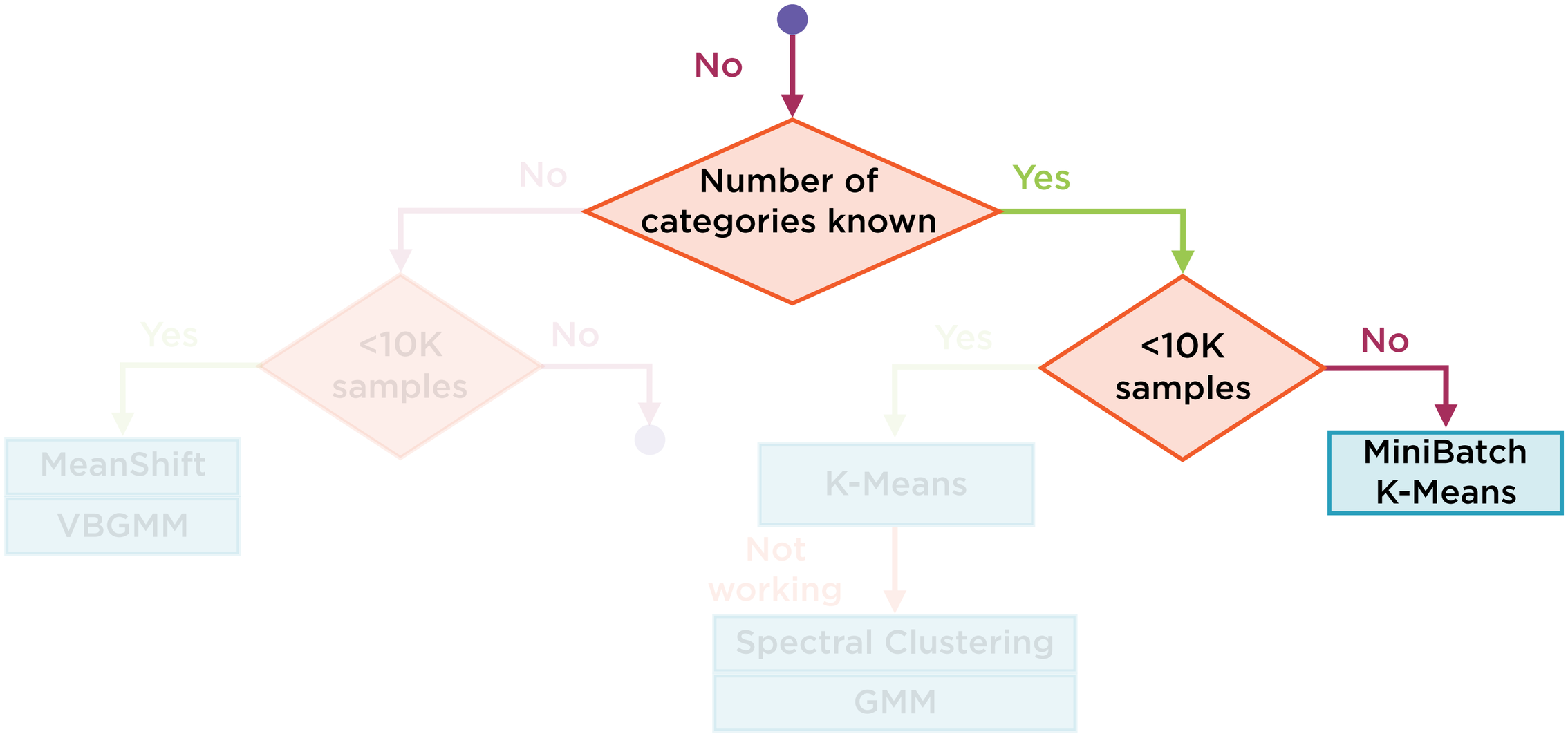
Clustering


No
Number of categories known
No
Yes
<10K samples
Yes
No
MeanShift
VBGMM
Yes
<10K samples
No
K-Means
MiniBatch K-Means
Not working
Spectral Clustering
GMM

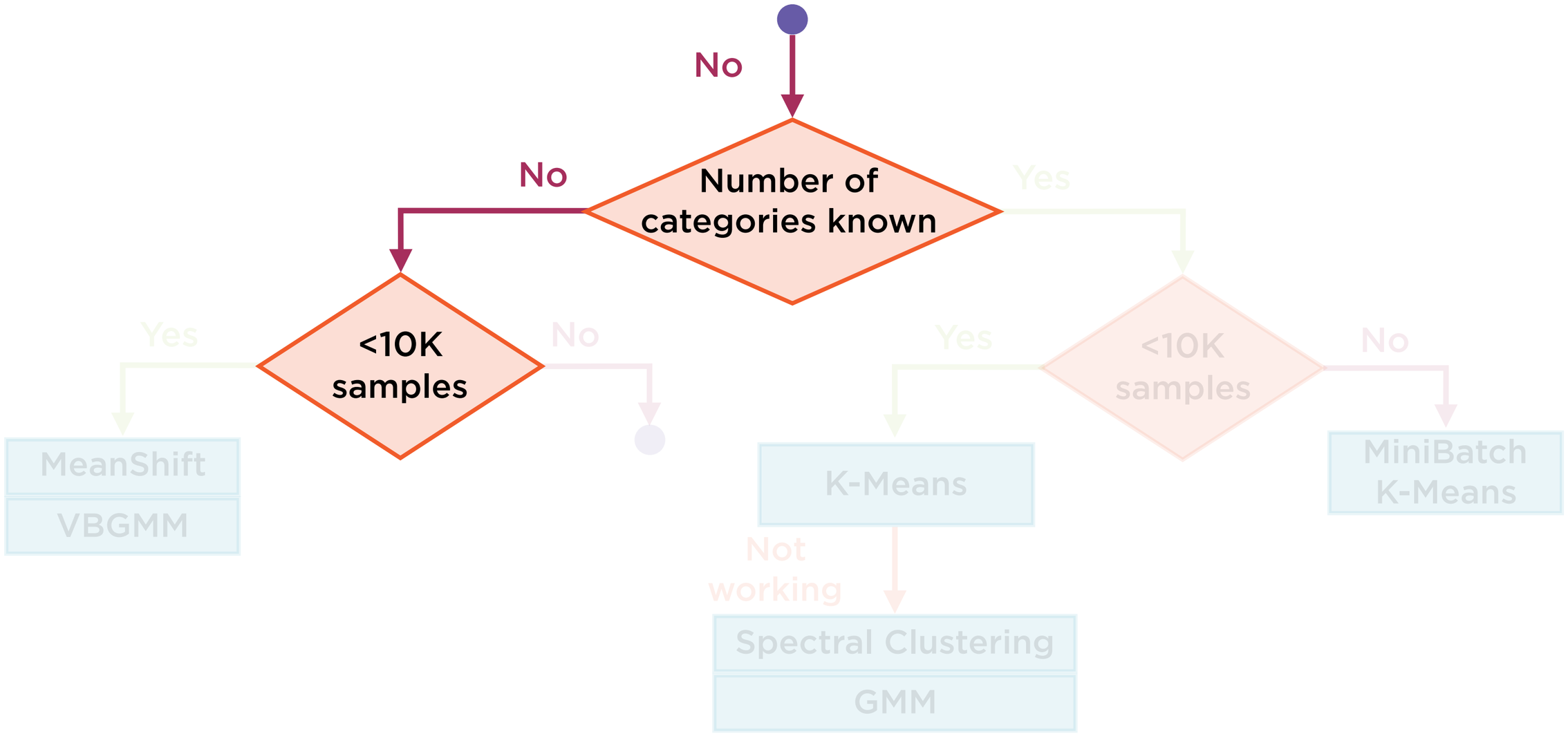
Clustering


No
Number of categories known
No
Yes
Yes
<10K samples
No
MeanShift
VBGMM
Yes
<10K samples
No
K-Means
MiniBatch K-Means
Not working
Spectral Clustering
GMM

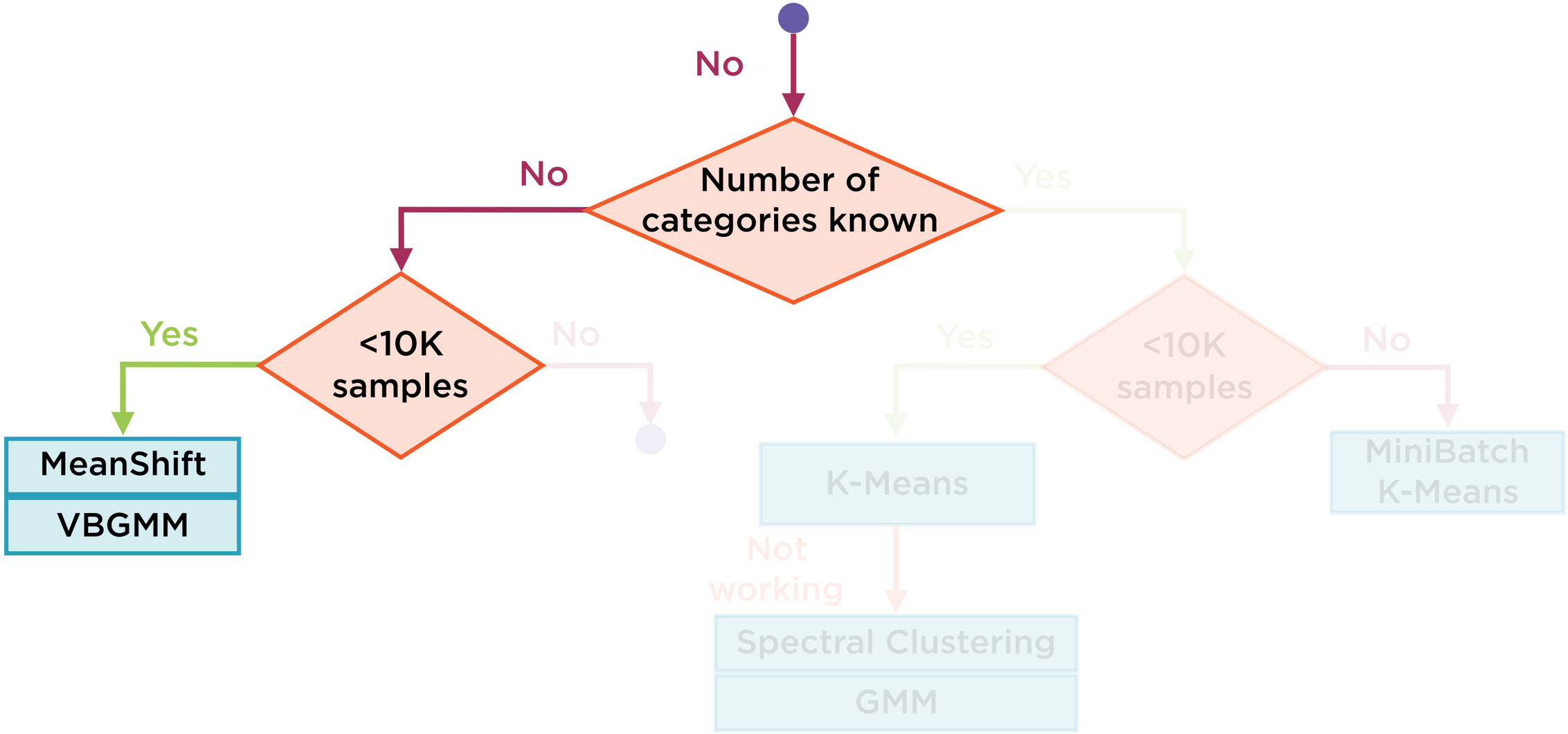
Clustering


No
Number of categories known
Yes
No
<10K samples
Yes
No
MeanShift
VBGMM
Yes
<10K samples
No
MiniBatch K-Means
K-Means
Not working
Spectral Clustering
GMM

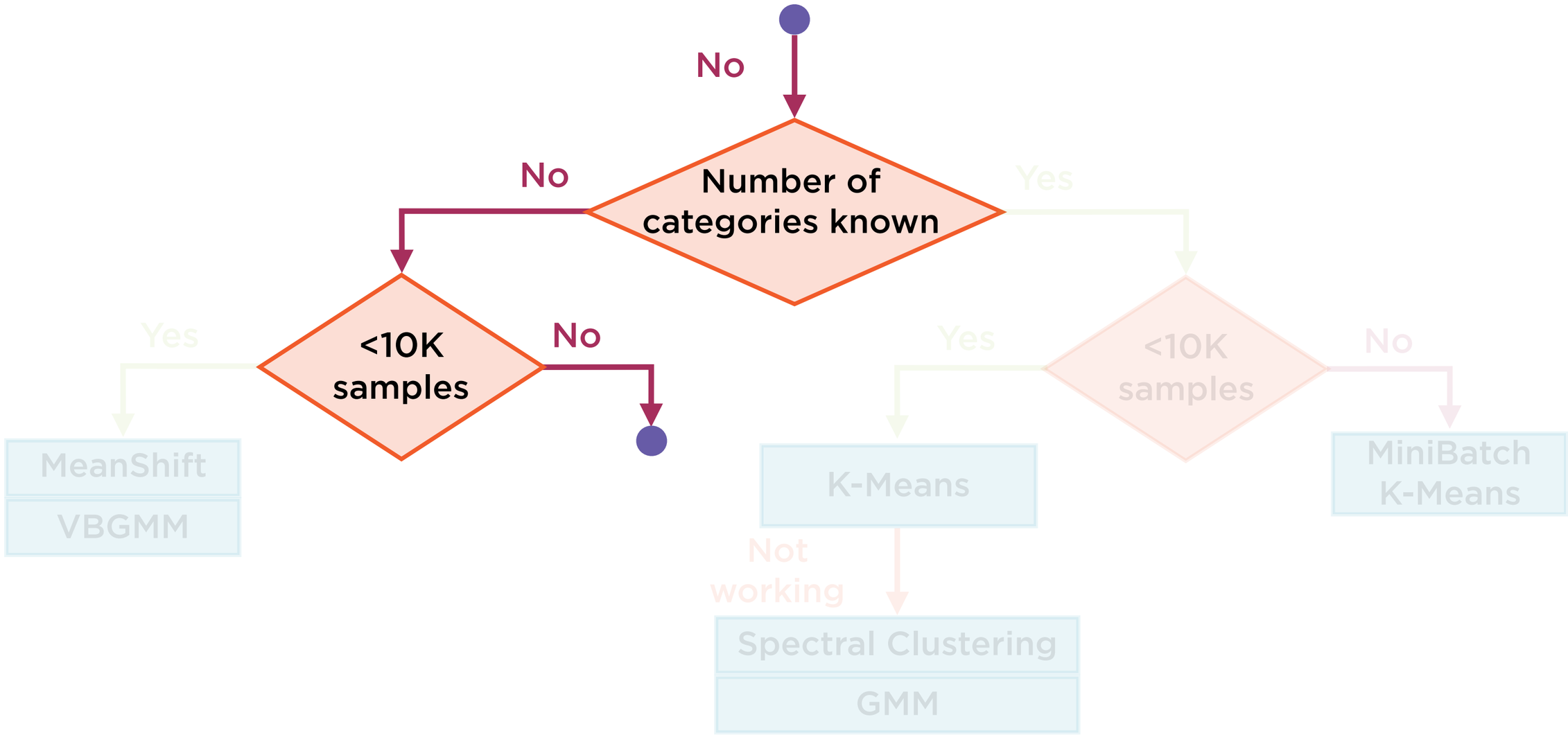
Clustering


No
Number of categories known
No
Yes
<10K samples
Yes
No
MeanShift
VBGMM
<10K samples
Yes
No
K-Means
MiniBatch K-Means
Not working
Spectral Clustering
GMM

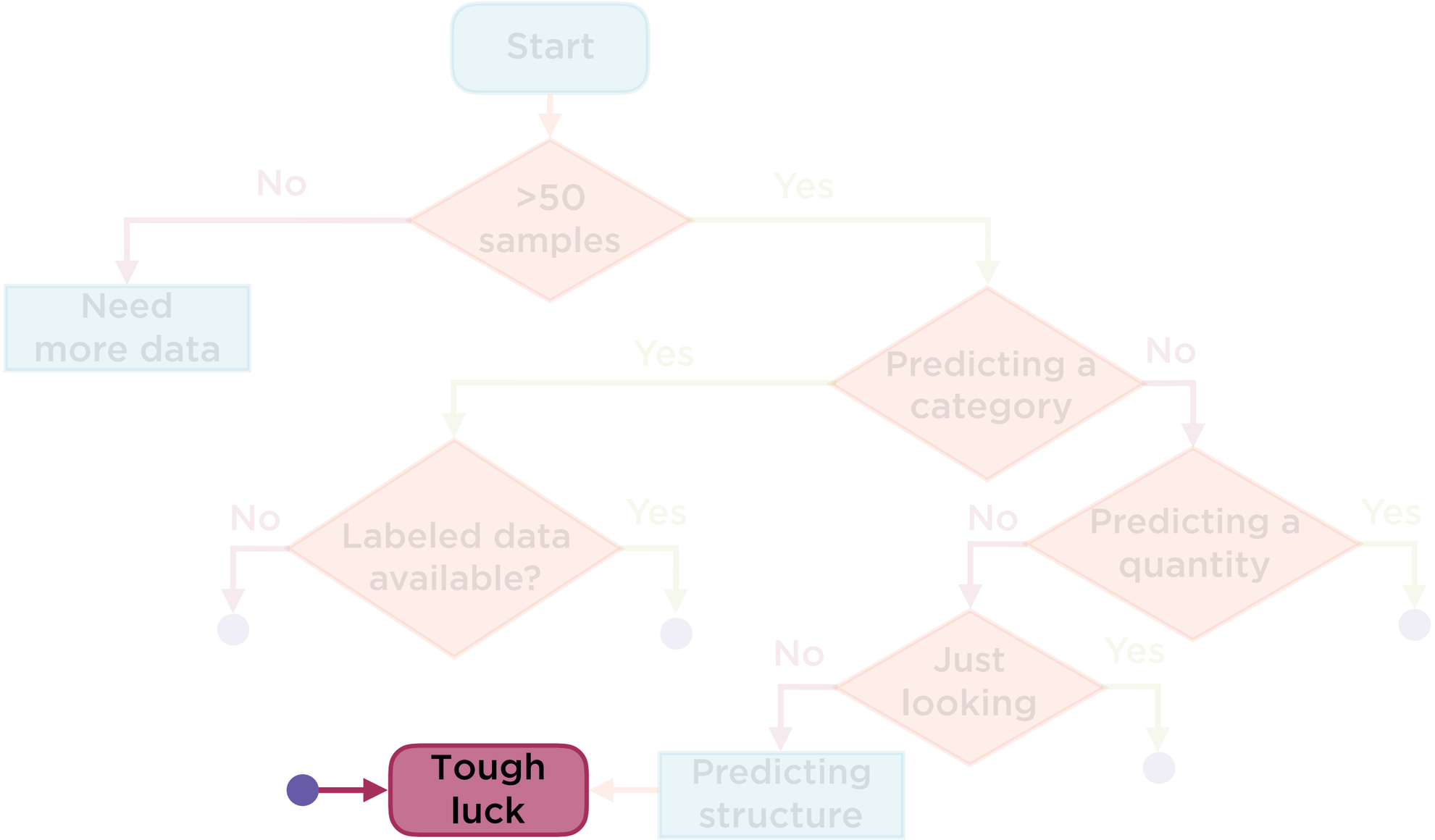
Clustering


No
Number of categories known
No
Yes
<10K samples
Yes
No
MeanShift
VBGMM
Yes
<10K samples
No
K-Means
MiniBatch K-Means
Not working
Spectral Clustering
GMM

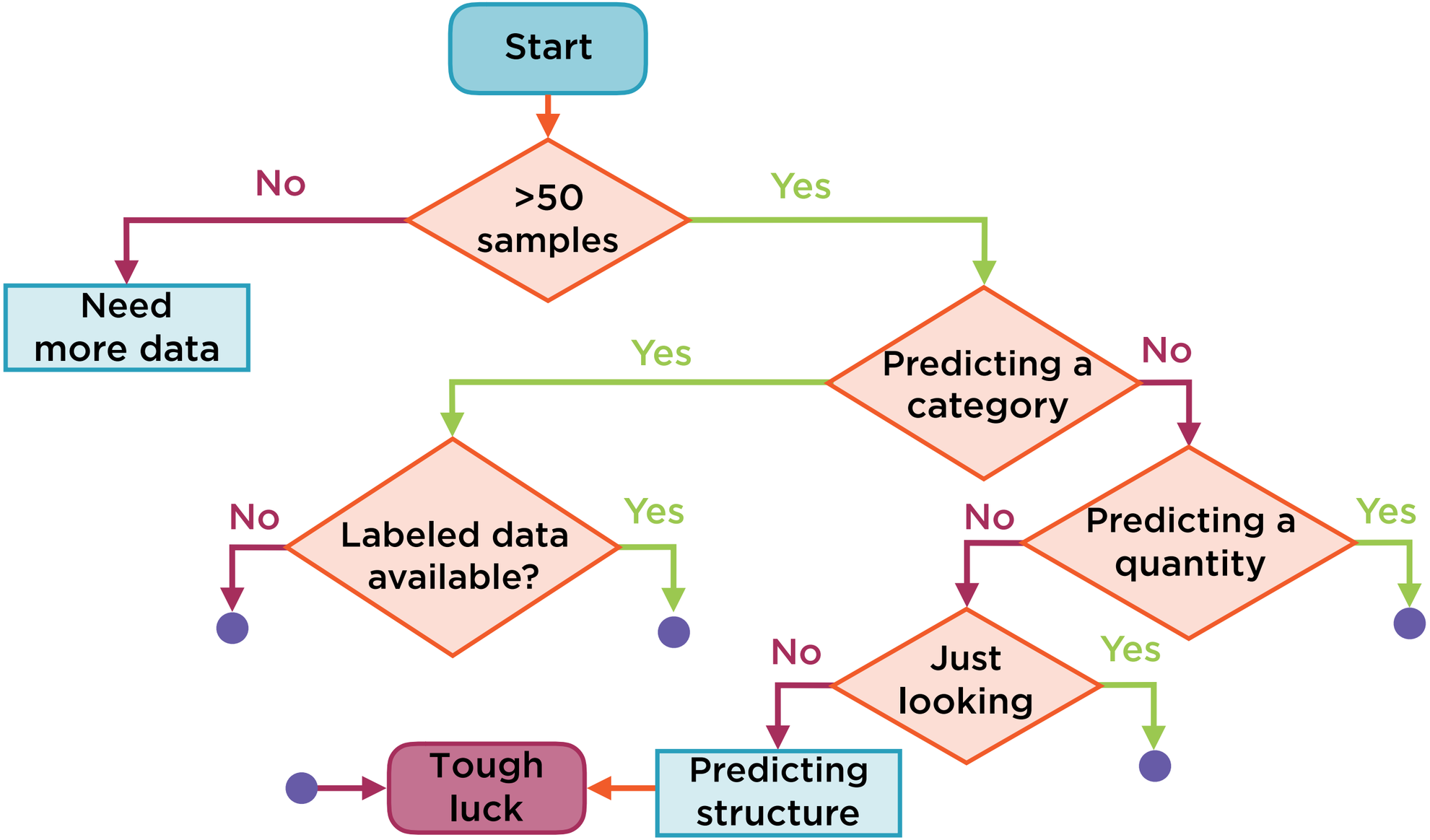
Clustering


No
Number of categories known
No
Yes
<10K samples
Yes
No
MeanShift
VBGMM
Yes
<10K samples
No
K-Means
MiniBatch K-Means
Not working
Spectral Clustering
GMM

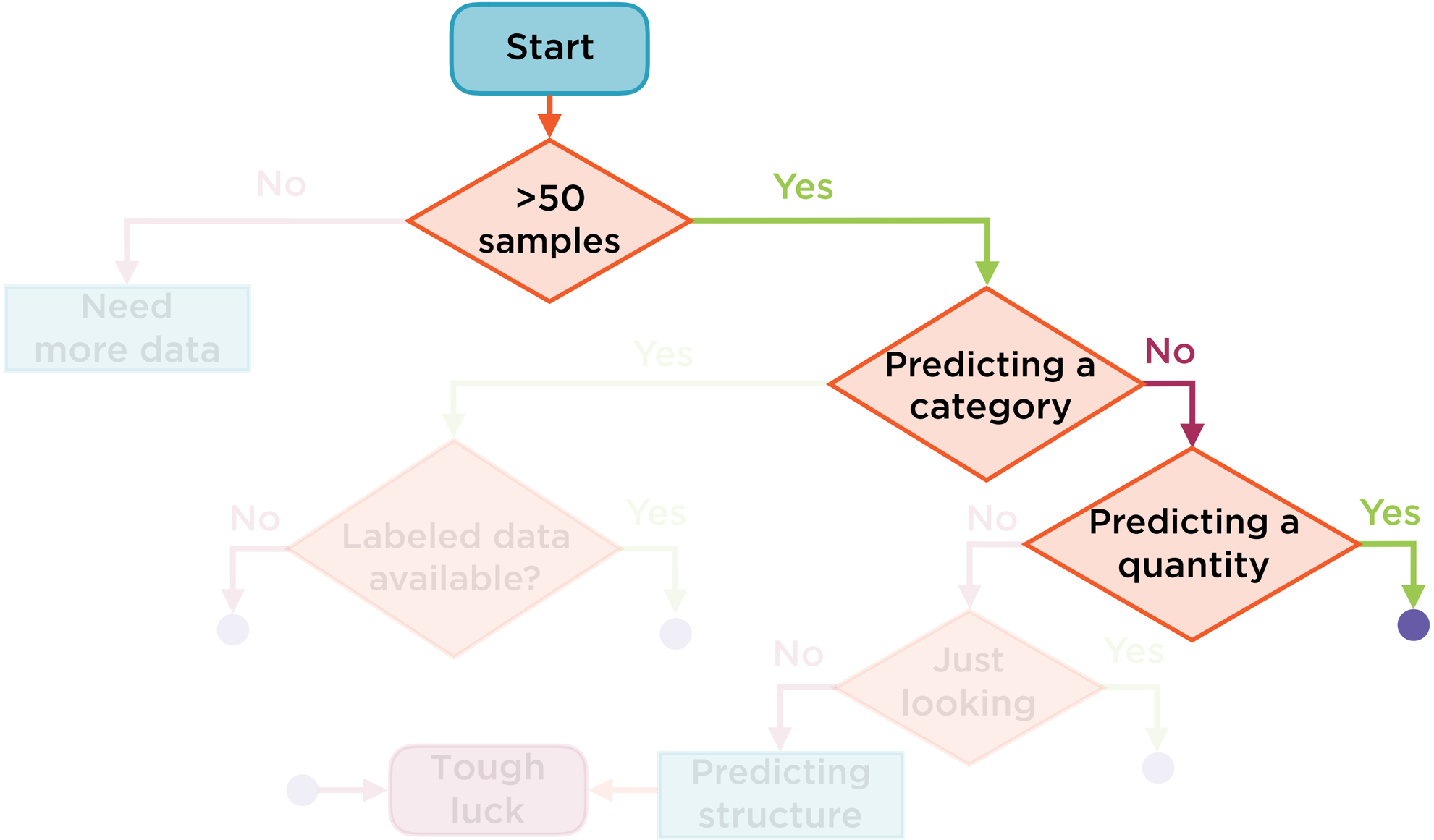
Choosing the Right Estimator



Choosing the Right Estimator



Choosing the Right Estimator



Types of Machine Learning Problems

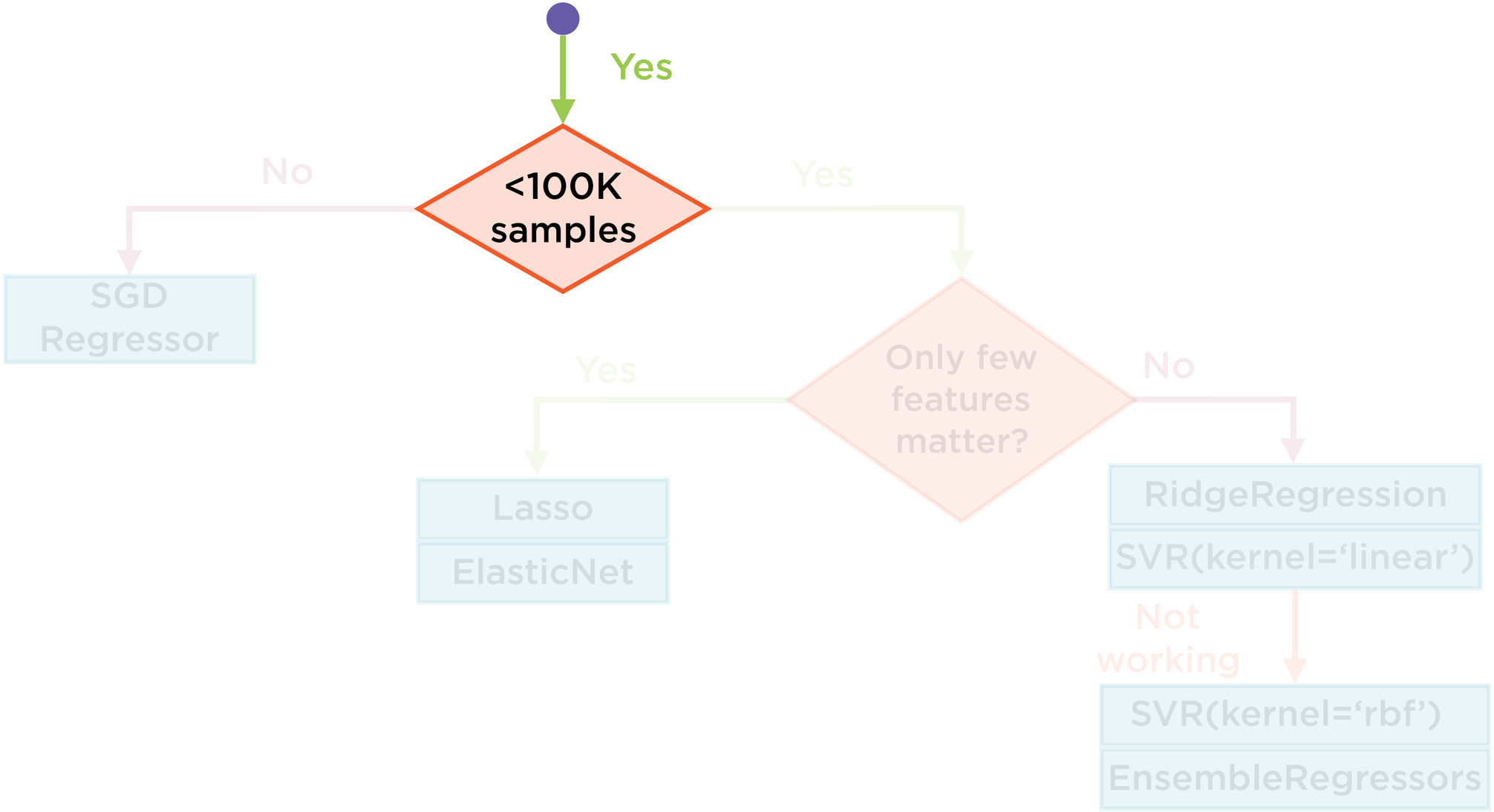
Classification

Regression

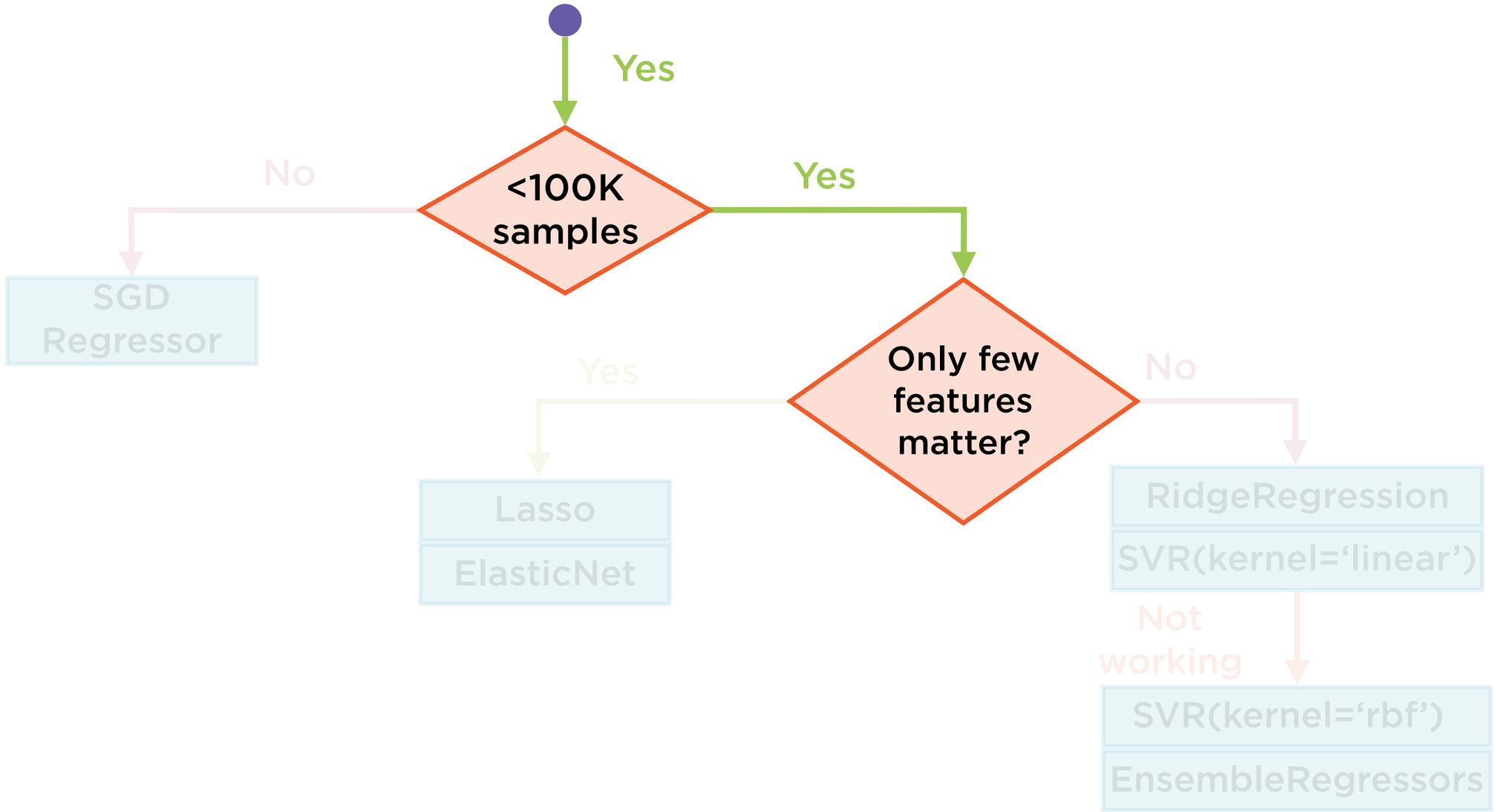
Clustering

Dimensionality reduction

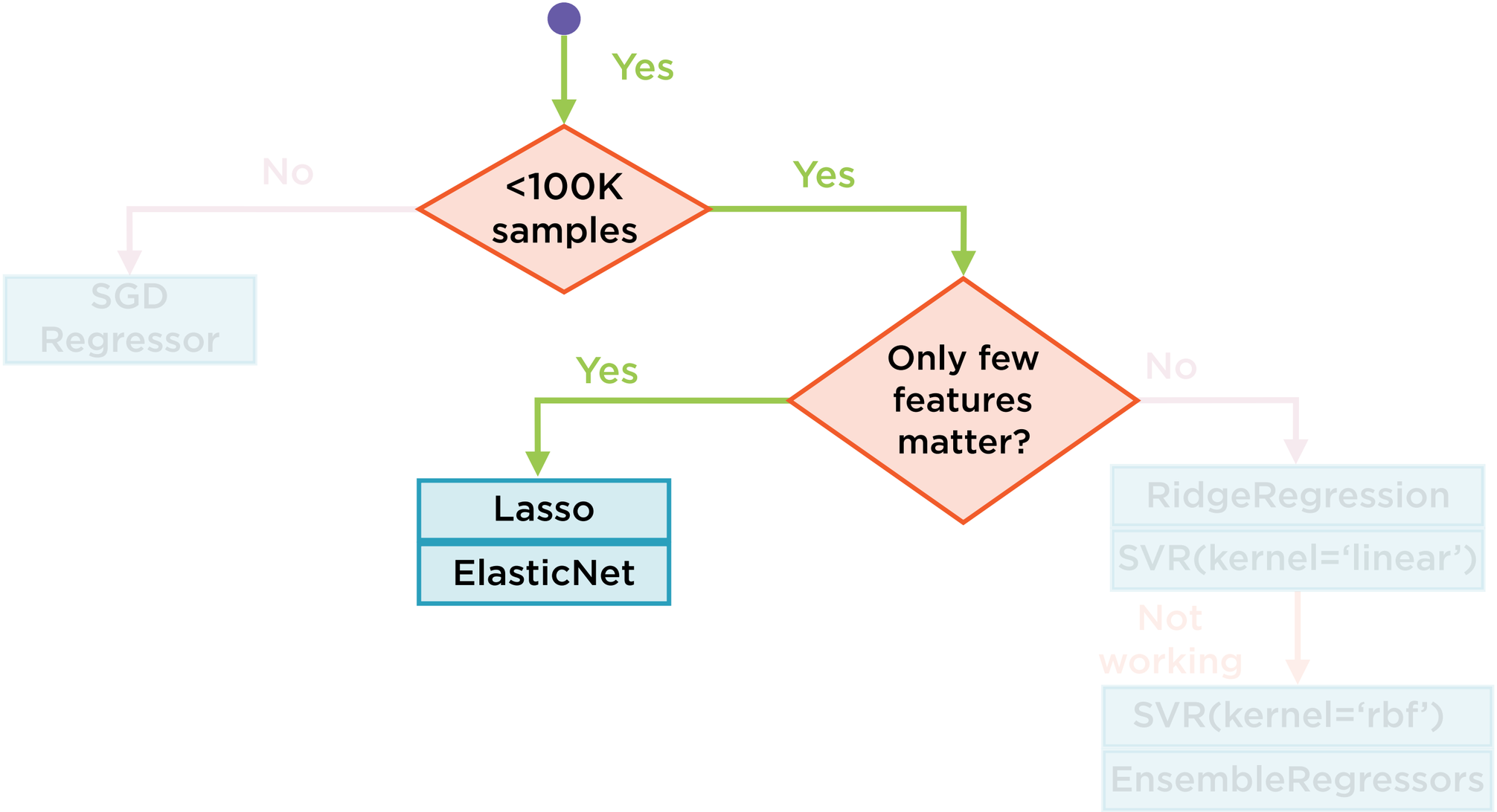
Regression


Yes
<100K samples
No
SGD Regressor
Yes
Only few features matter?
Yes
Lasso
ElasticNet
No
RidgeRegression
SVR(kernel='linear')
Not working
SVR(kernel='rbf')
EnsembleRegressors

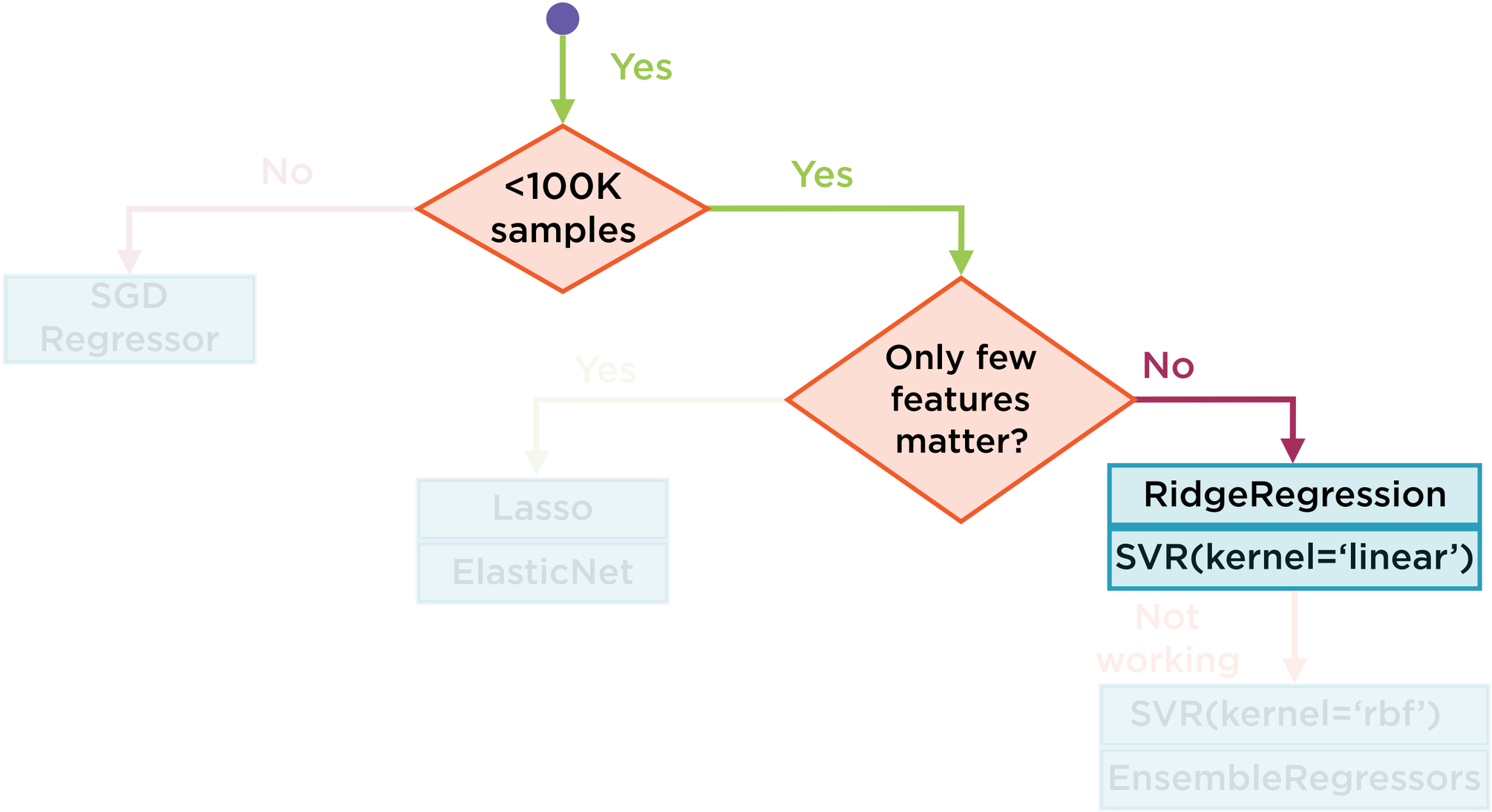
Regression


Yes
<100K samples
No
SGD Regressor
Yes
Only few features matter?
Yes
Lasso
ElasticNet
No
RidgeRegression
SVR(kernel='linear')
Not working
SVR(kernel='rbf')
EnsembleRegressors

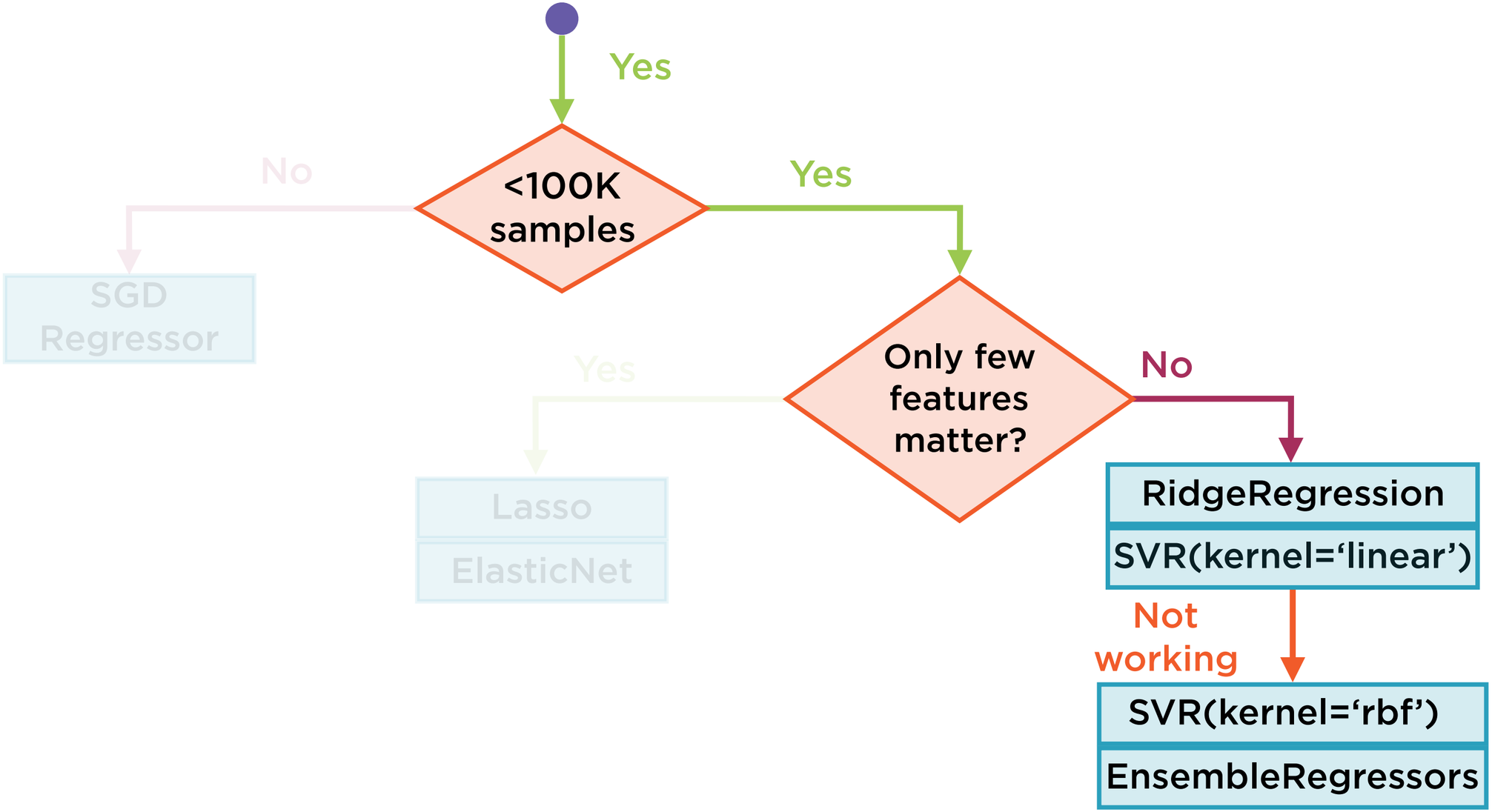
Regression


Yes
No
<100K samples
Yes
SGD Regressor
Yes
Only few features matter?
No
Lasso
ElasticNet
RidgeRegression
SVR(kernel='linear')
Not working
SVR(kernel='rbf')
EnsembleRegressors

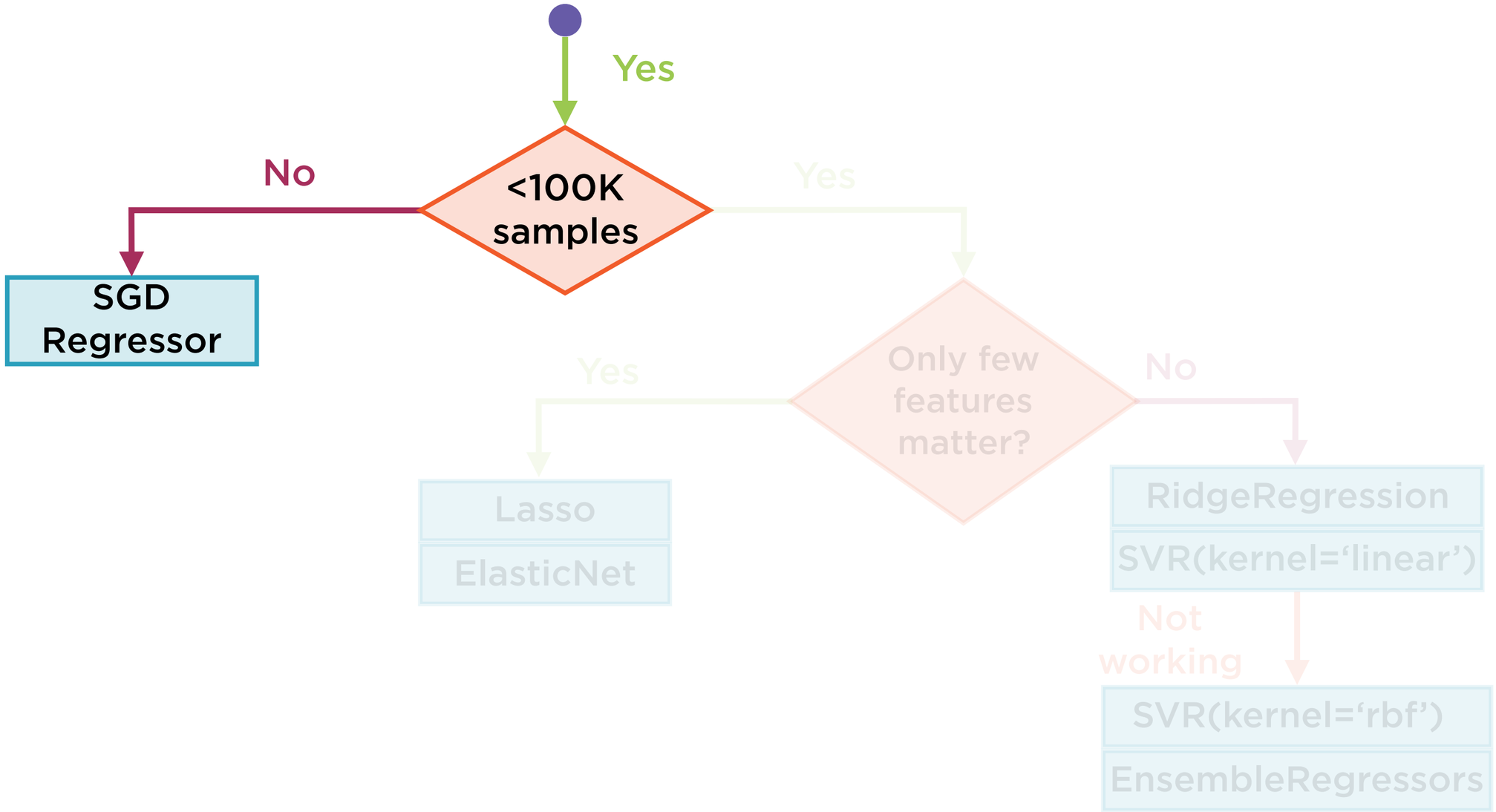
Regression


Yes
<100K samples
No
SGD Regressor
Yes
Only few features matter?
Yes
Lasso
ElasticNet
No
RidgeRegression
SVR(kernel='linear')
Not working
SVR(kernel='rbf')
EnsembleRegressors

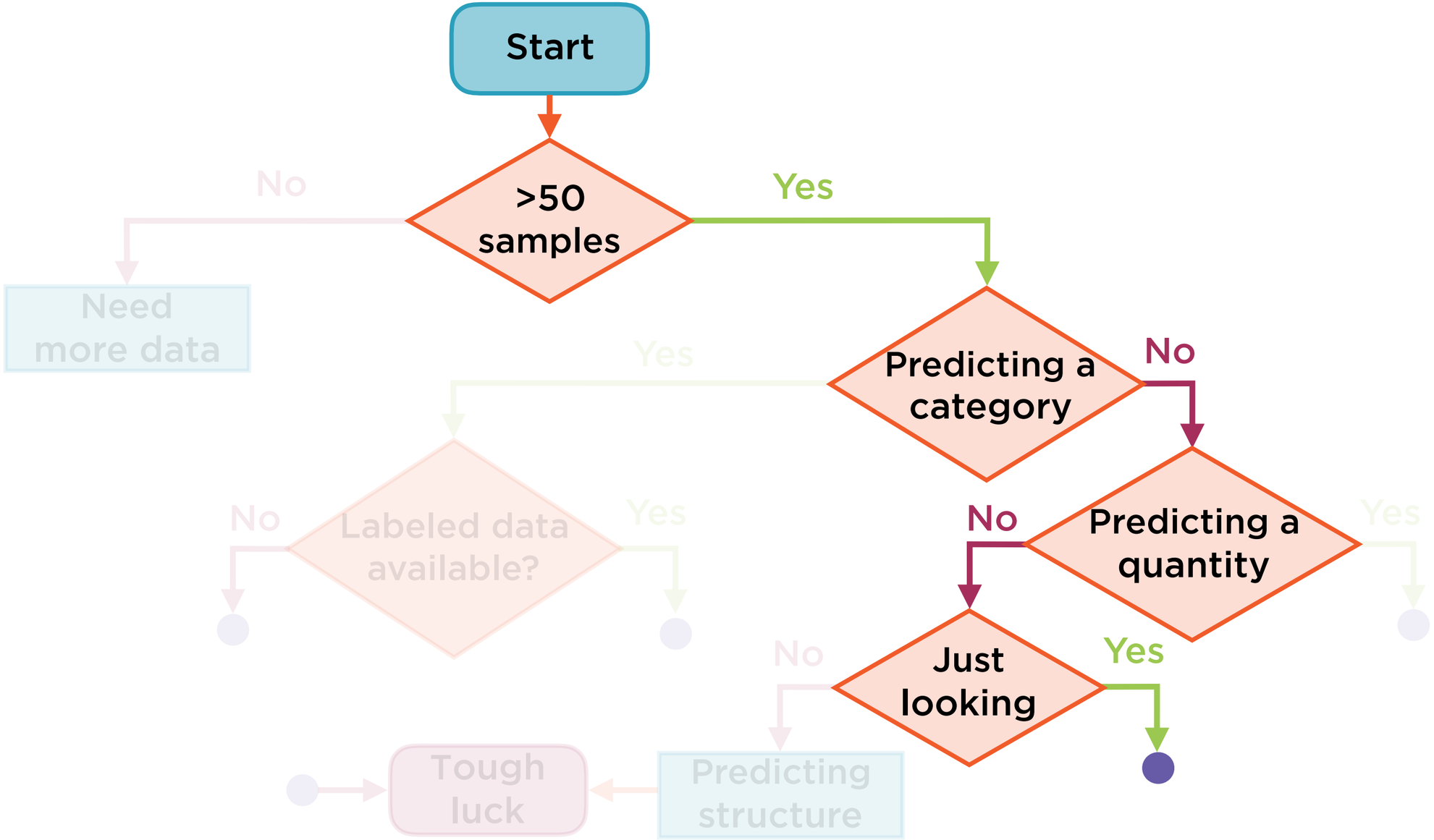
Regression


Yes
<100K samples
No
SGD Regressor
Yes
Only few features matter?
Yes
Lasso
ElasticNet
No
RidgeRegression
SVR(kernel='linear')
Not working
SVR(kernel='rbf')
EnsembleRegressors

Regression


Yes
No
<100K samples
SGD Regressor
Yes
Only few features matter?
Yes
Lasso
ElasticNet
No
RidgeRegression
SVR(kernel='linear')
Not working
SVR(kernel='rbf')
EnsembleRegressors

Choosing the Right Estimator

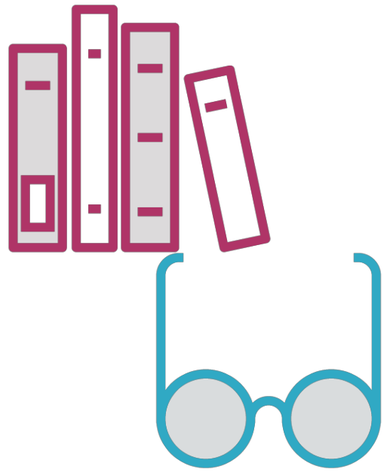


Types of Machine Learning Problems

Classification

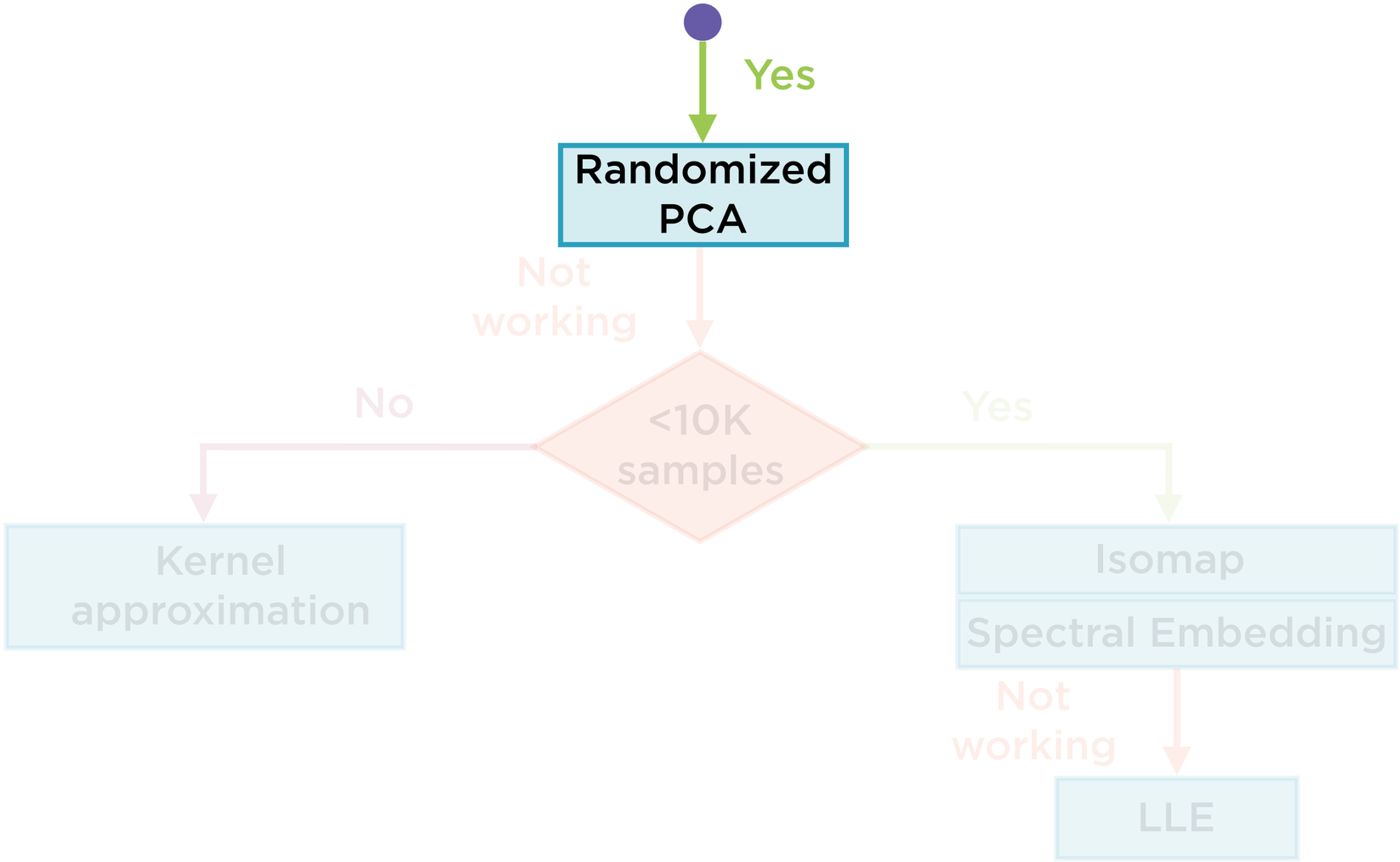
Regression

Clustering

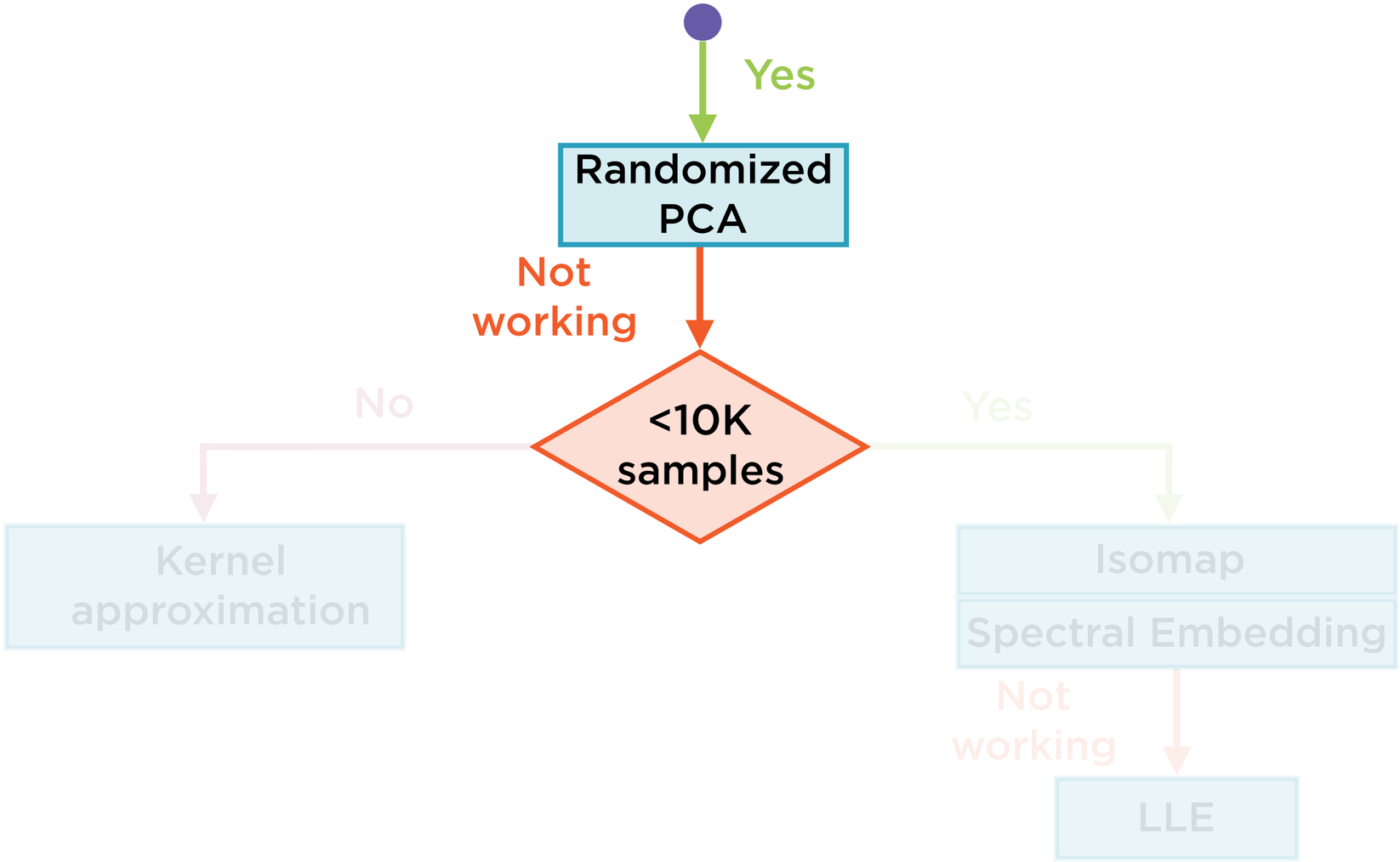


Dimensionality reduction

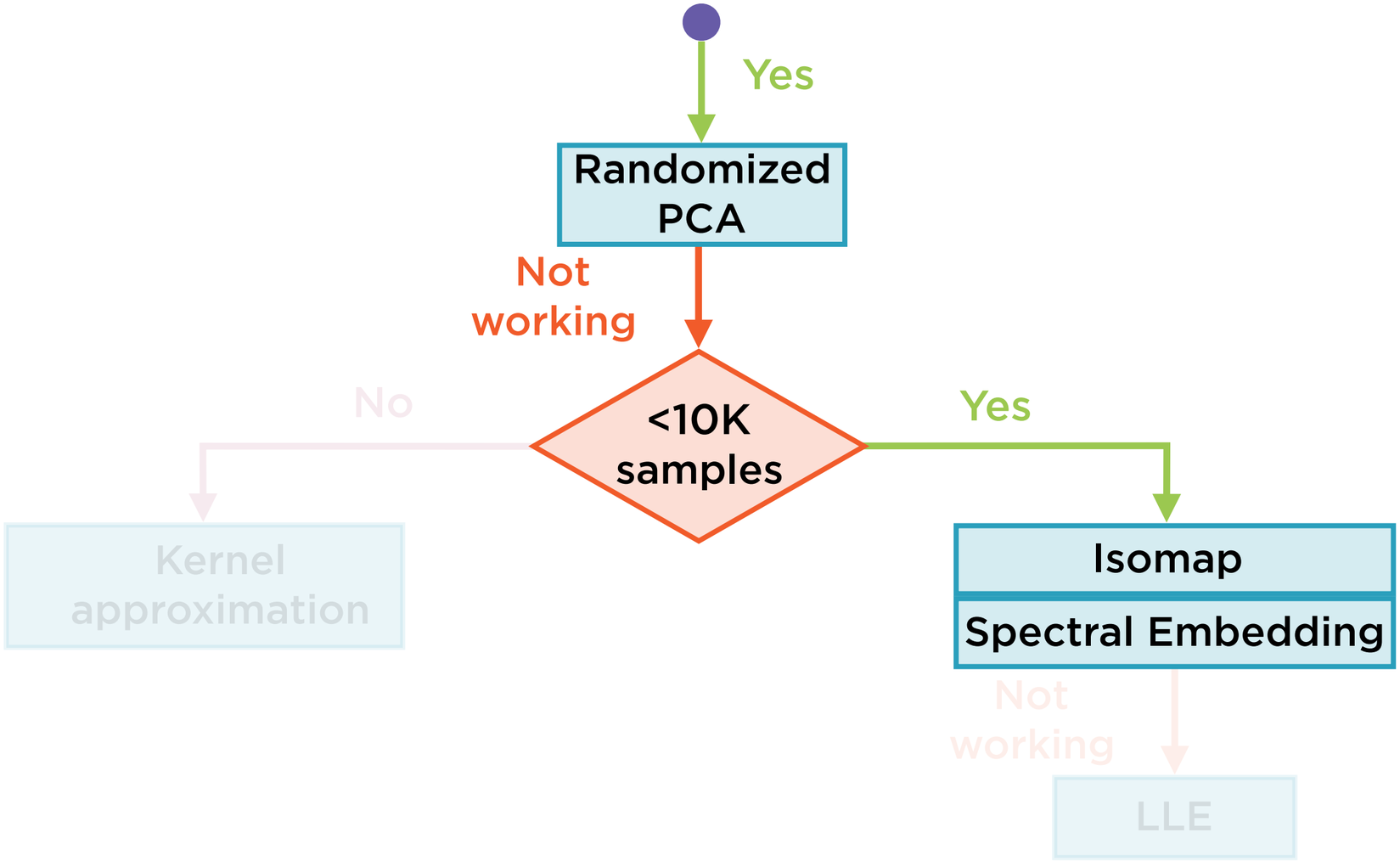
Dimensionality Reduction



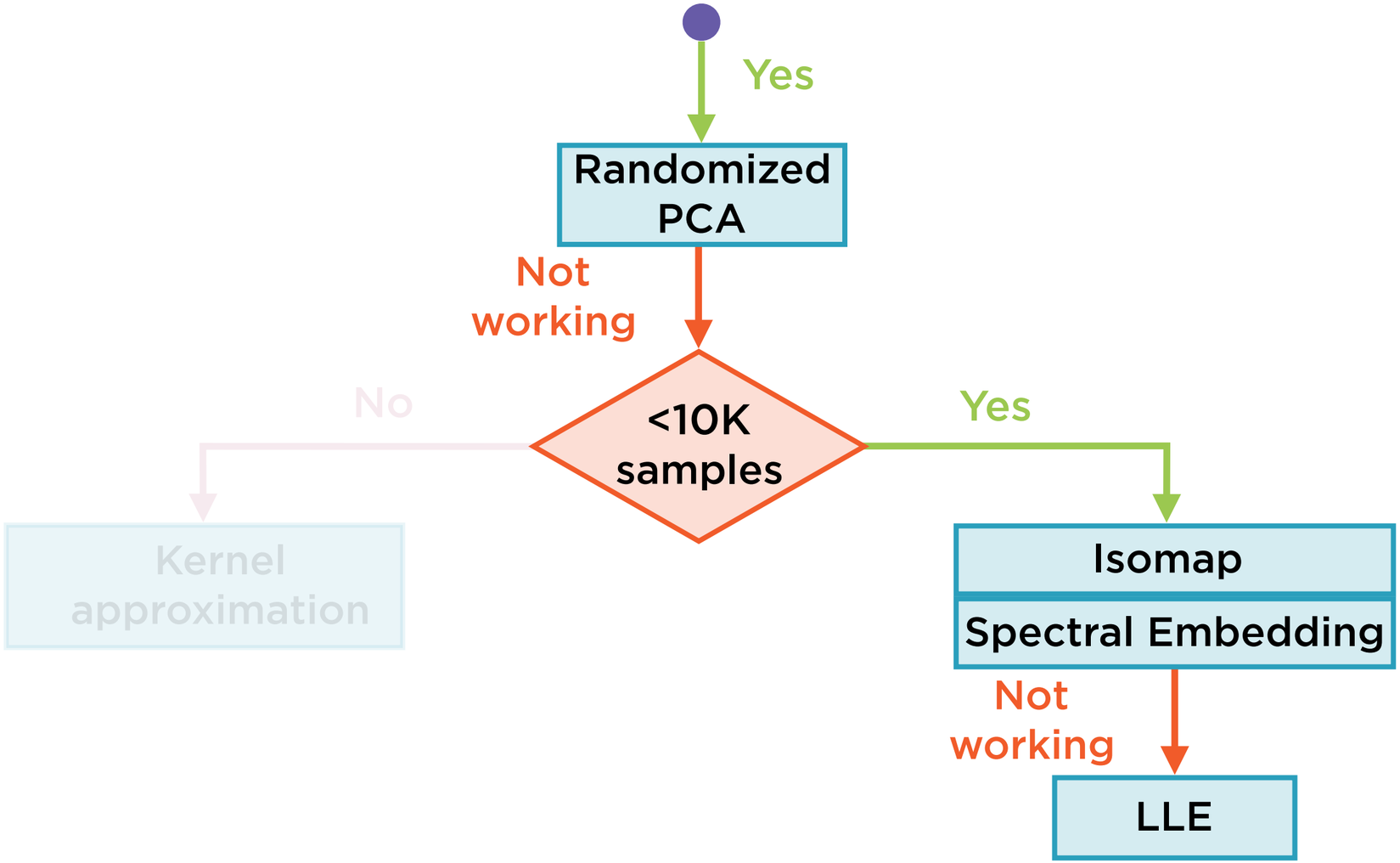
Dimensionality Reduction



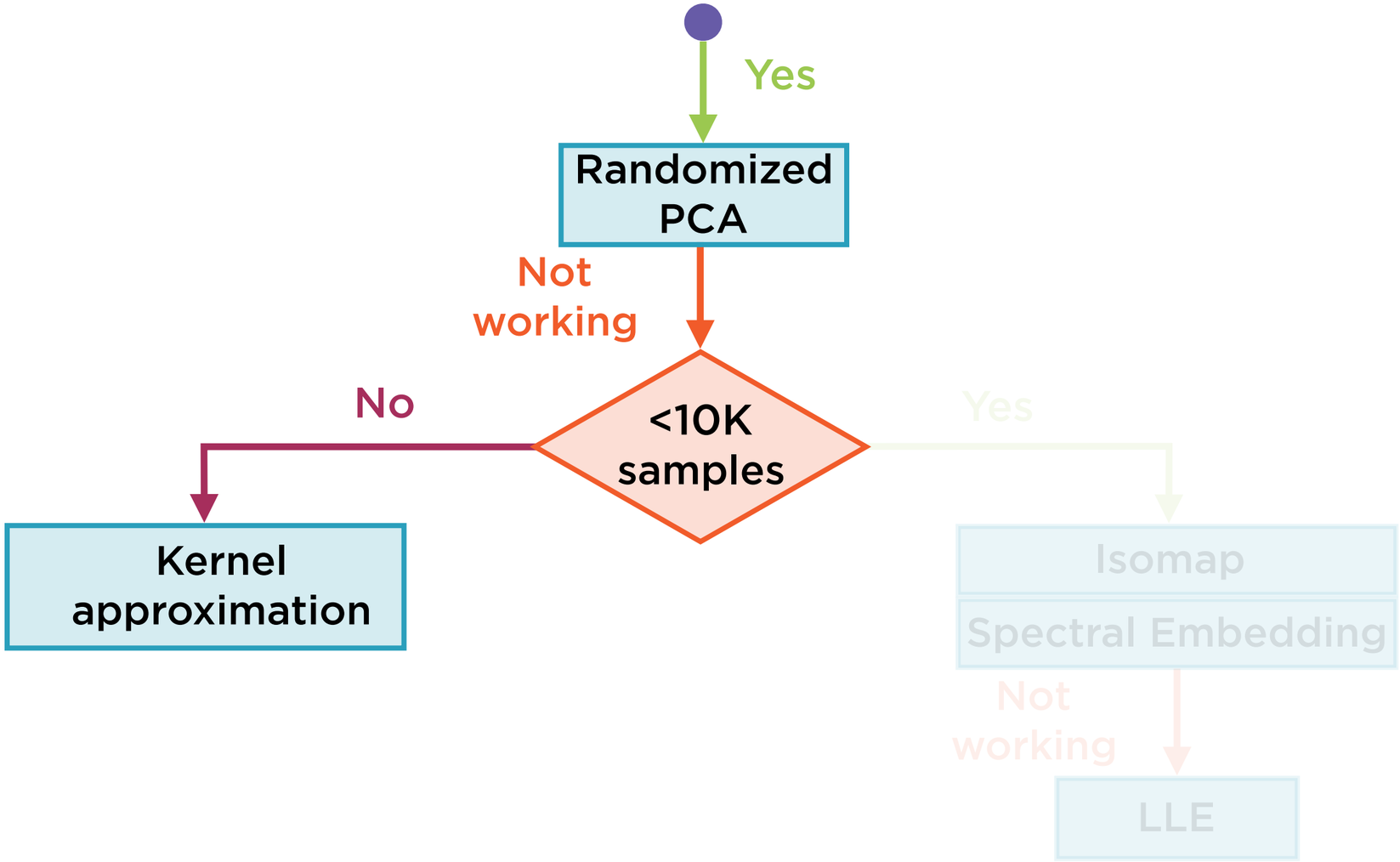
Dimensionality Reduction



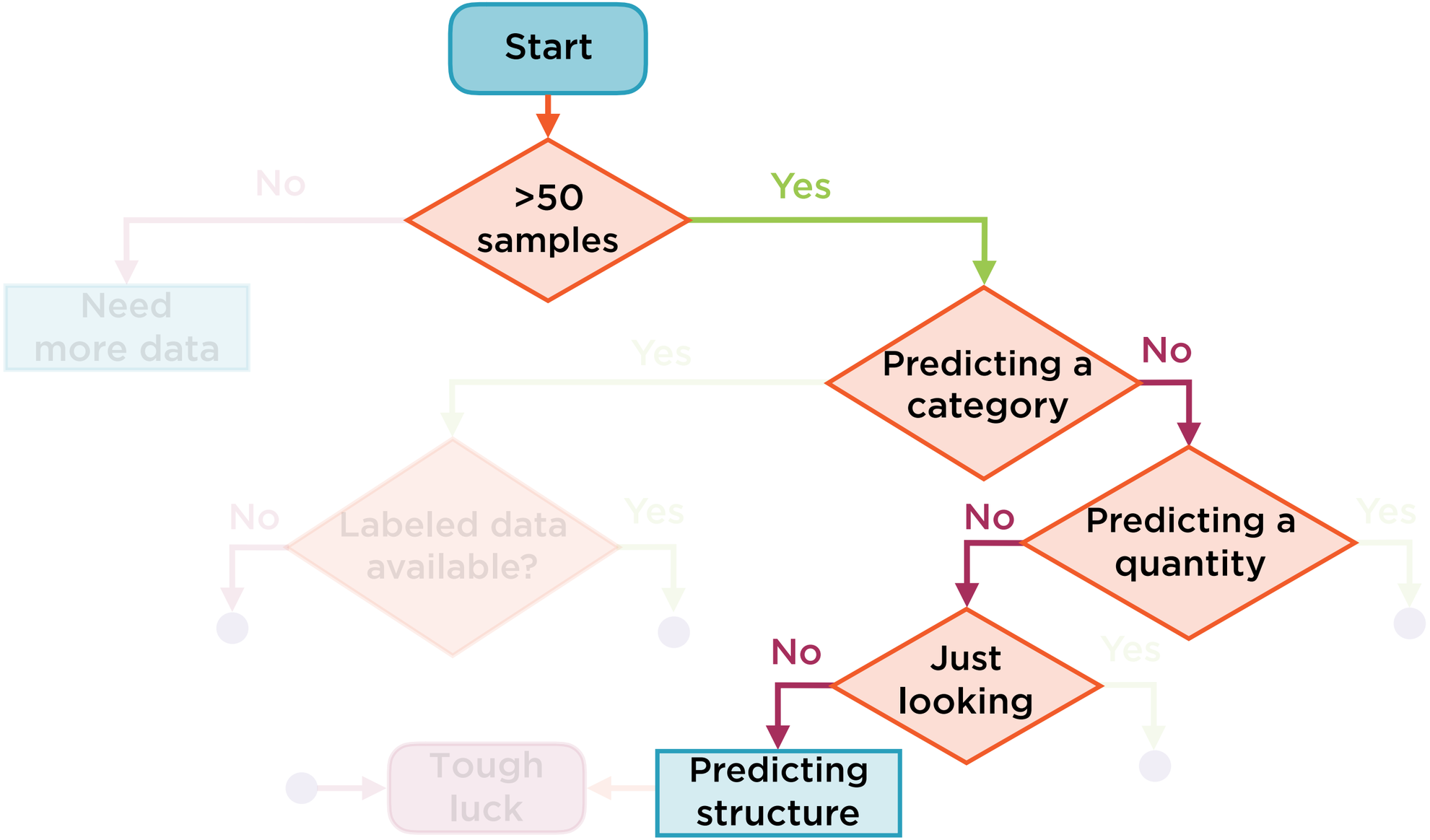
Dimensionality Reduction



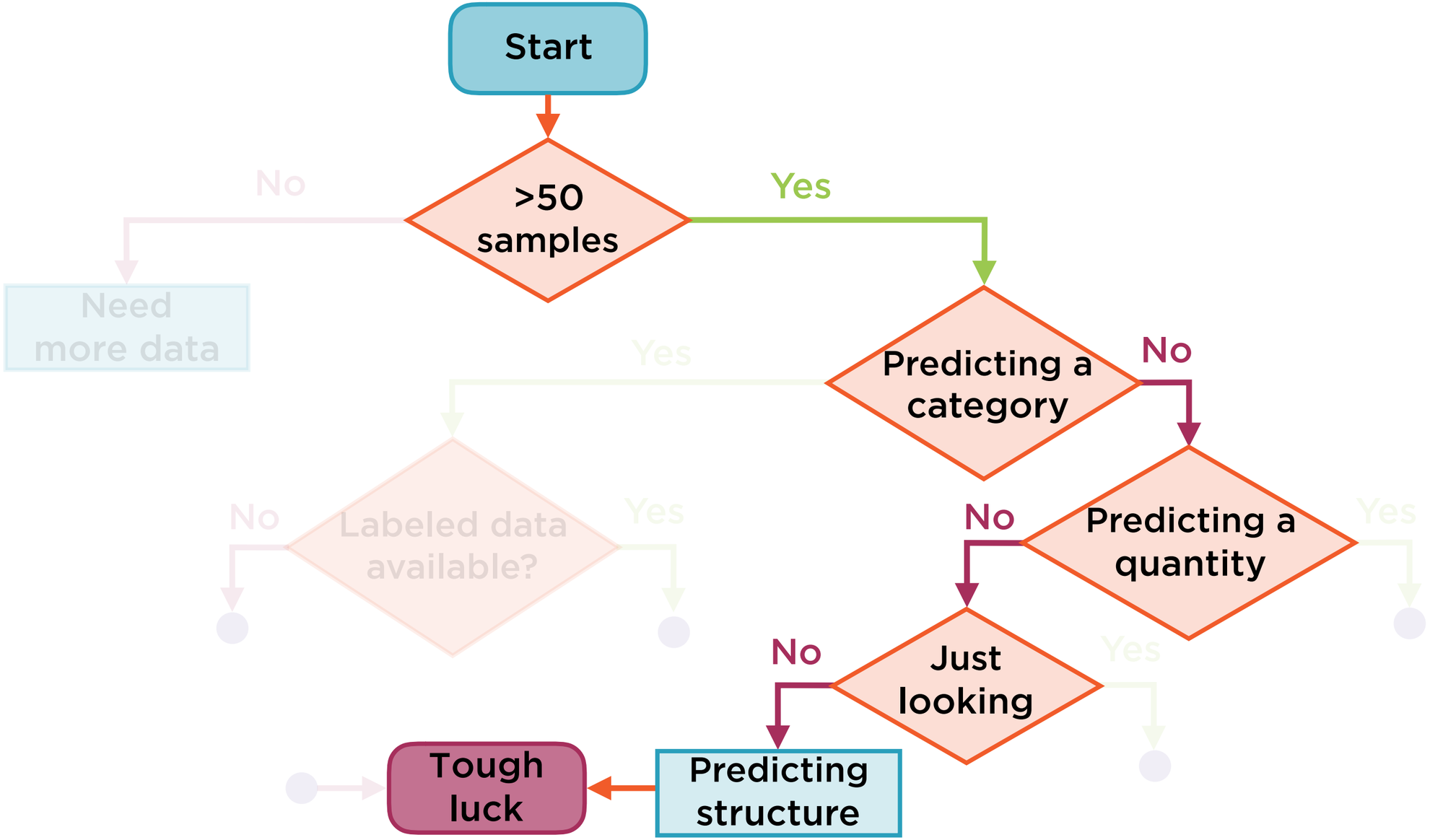
Dimensionality Reduction



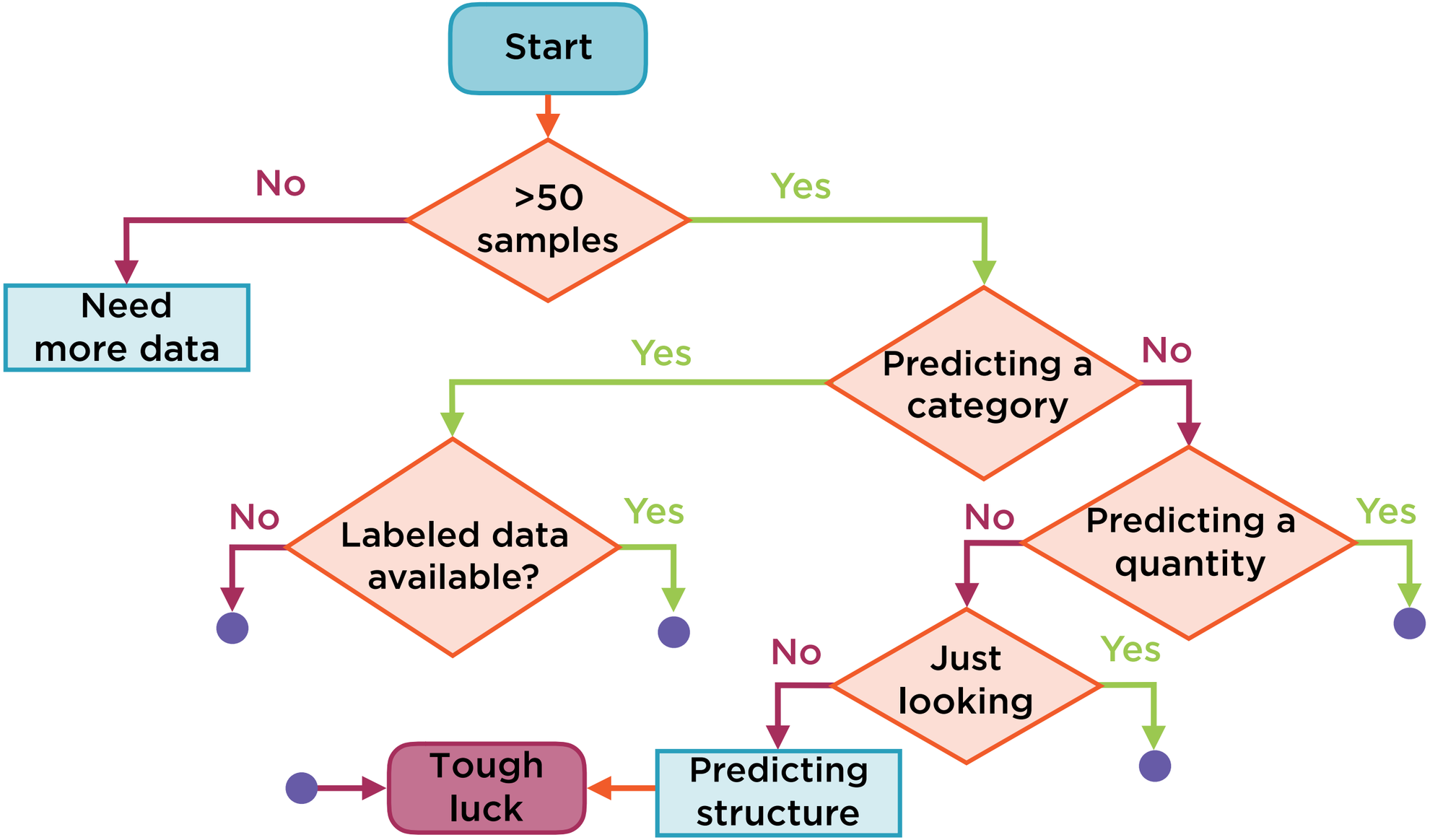
Choosing the Right Estimator



Choosing the Right Estimator



Choosing the Right Estimator



Demo

Exploring built-in datasets in scikit-learn

Demo

Loading and working with external datasets in scikit-learn

Summary

scikit-learn in the typical ML workflow

Estimators and pipelines

Model evaluation and transformation

Loading, cleaning, transforming, and visualizing datasets