The Estimator API

The Estimator API

Estimator API for consist interface

Create a model object

Fit to training Data

Predict for new data

Pipelines for complex operations

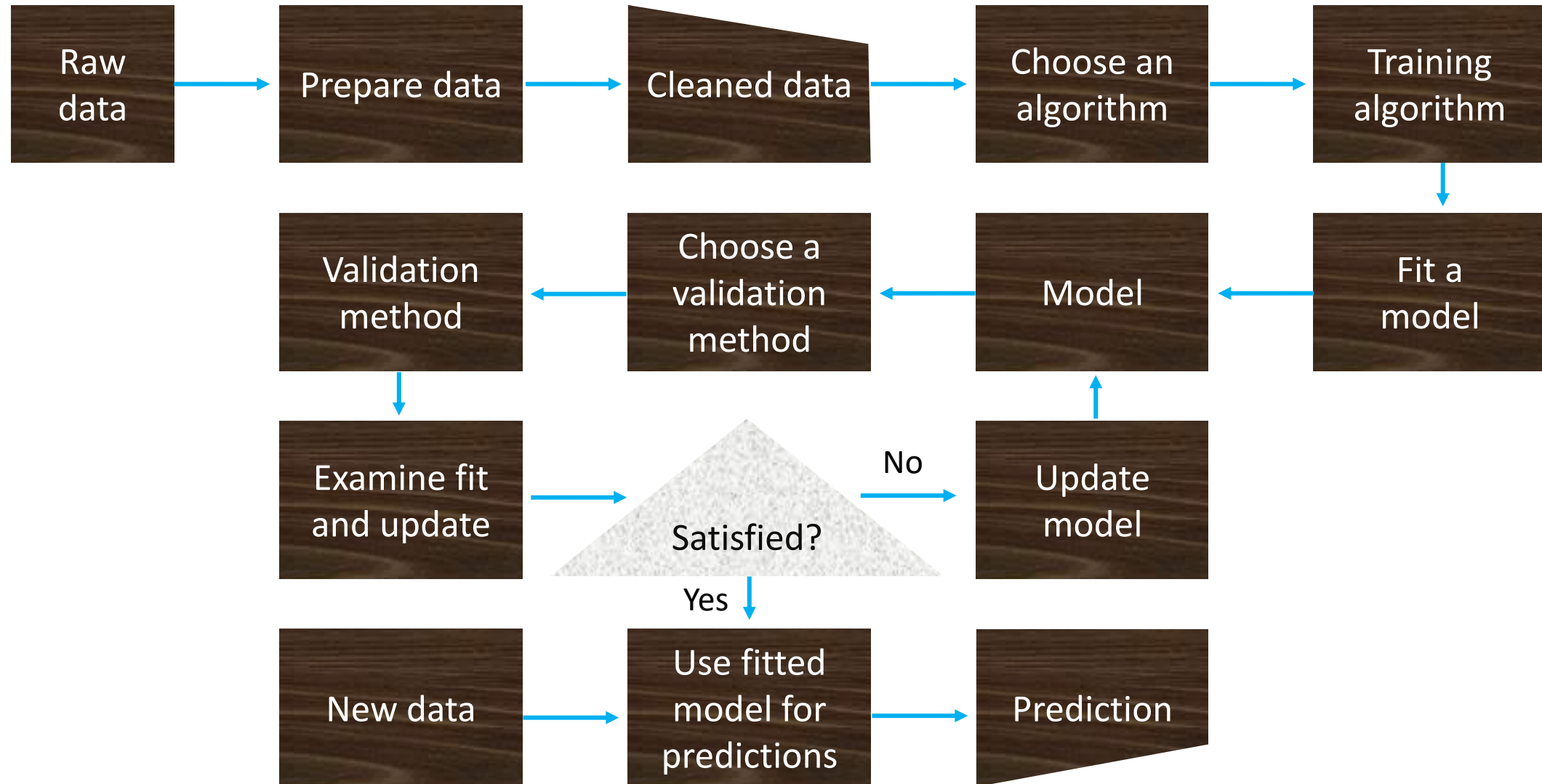# Principles Underlying Estimator APIs

Consistency

Inspection
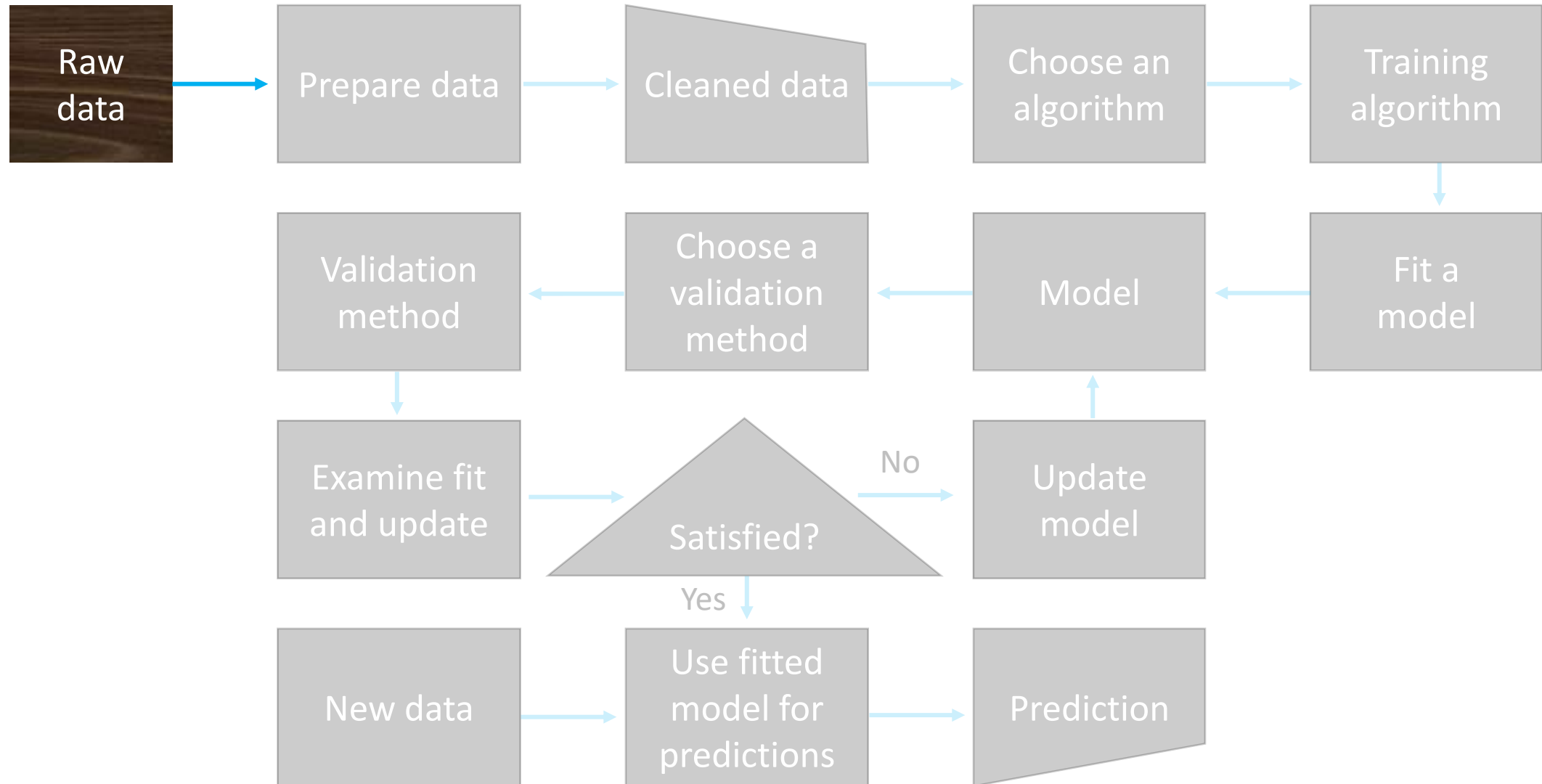
Limited object hierarchy

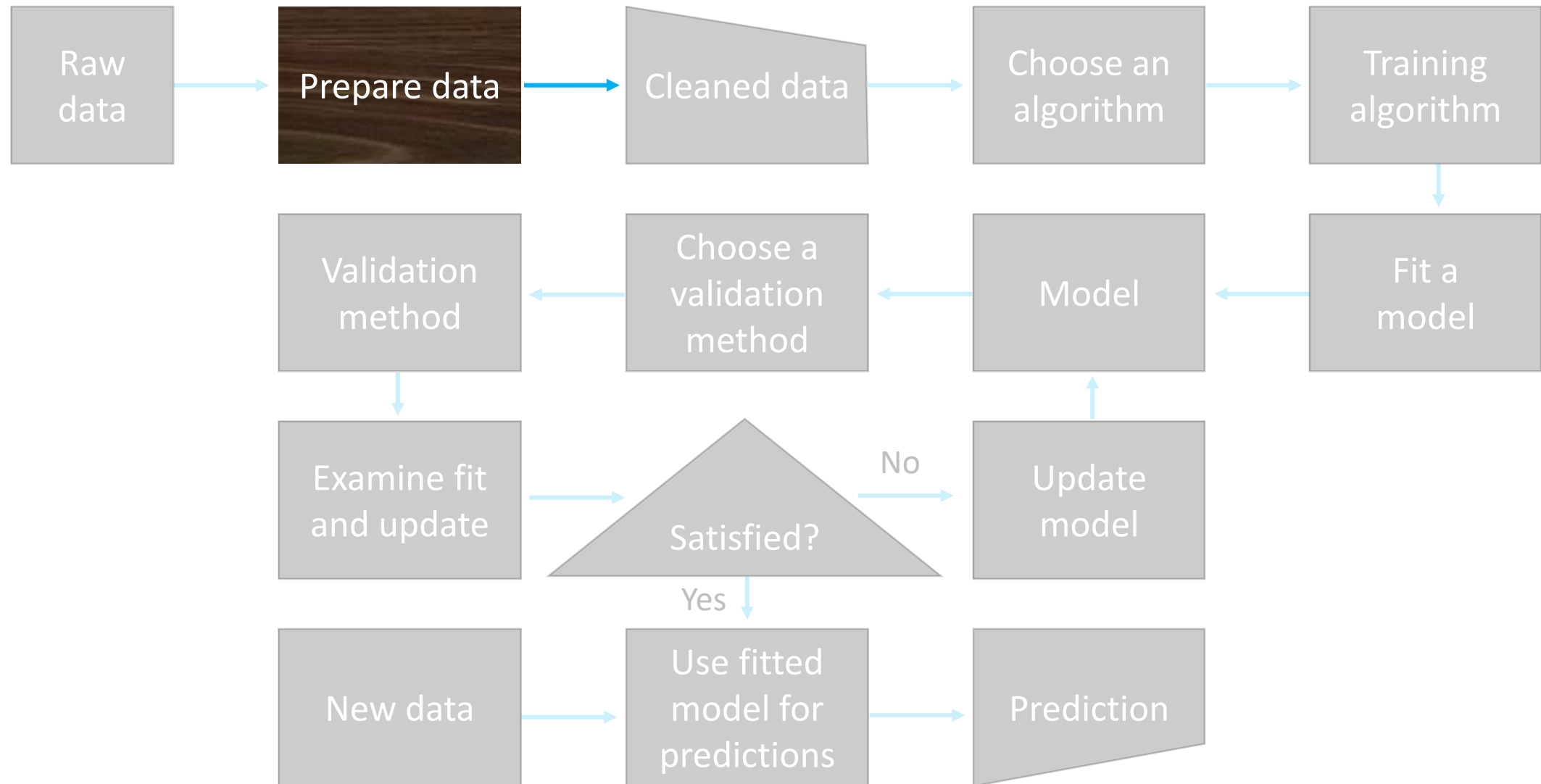Composition

Sensible defaults

Basic Machine Learning Workflow

Pandas and NumPy Inter-operability

Standardization, Normalization, Scaling

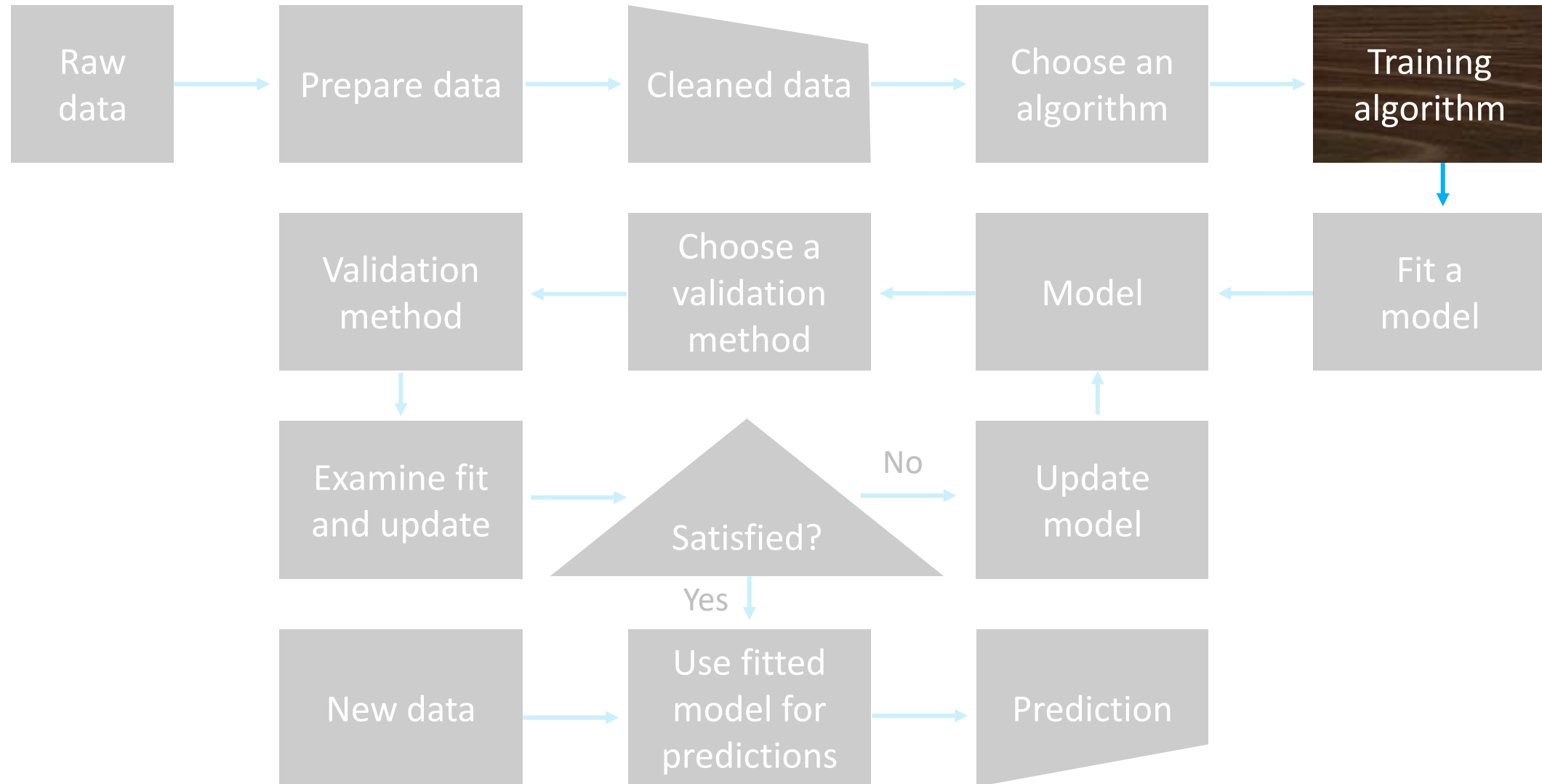```
┌──────────┐   ┌──────────┐   ┌──────────┐   ┌──────────┐   ┌──────────┐
│   Raw   │ ───▶ │ Prepare │ ───▶ │ Cleaned │ ───▶ │Choose an │ ───▶ │ Training │
│   data   │   │   data   │   │   data   │   │algorithm │   │algorithm │
└──────────┘   └──────────┘   └──────────┘   └──────────┘   └──────────┘
                                                                              │
                                                                              ▼
              ┌──────────┐   ┌──────────┐   ┌──────────┐   ┌──────────┐
              │Validation│ ◀─── │ Choose a │ ◀─── │ Model   │ ◀─── │ Fit a   │
              │ method │   │validation│   │   │   │ model   │
              └──────────┘   │ method │   └──────────┘   └──────────┘
                    │   └──────────┘
                    ▼
              ┌──────────┐   ╱╲   No   ┌──────────┐
              │Examine fit│ ───▶ ╱   ╲   ───▶ │ Update │
              │and update │   ╱Satisfied?╲   │ model   │
              └──────────┘   ╱_____╲   └──────────┘
                                      │ Yes
                                      ▼
              ┌──────────┐   ┌──────────┐   ┌──────────┐
              │ New data │ ───▶ │Use fitted│ ───▶ │Prediction│
              │   │   │model for │   │   │
              └──────────┘   │predictions│   └──────────┘
                                └──────────┘
```

# Missing Values and Outliers
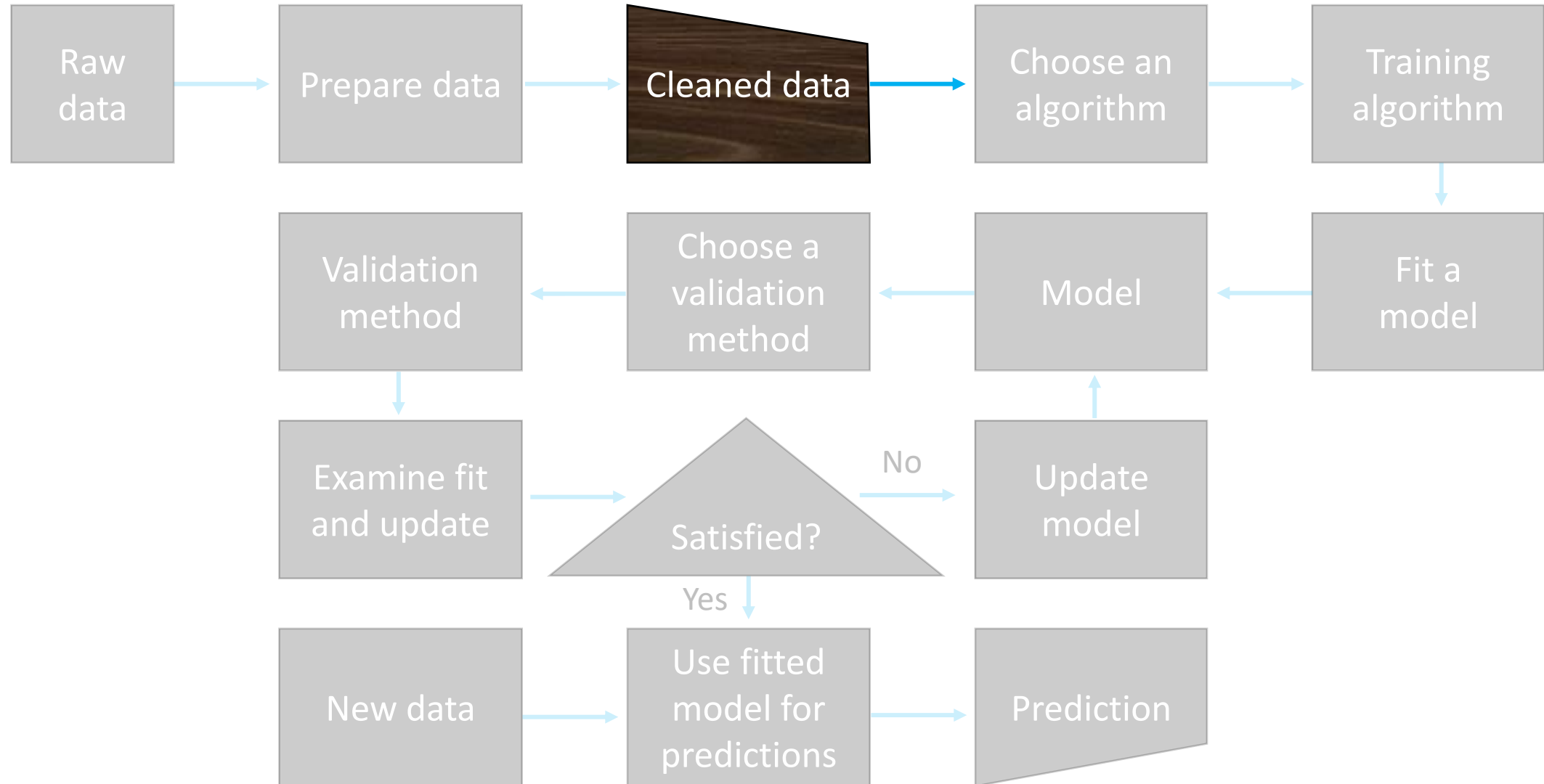
# Comprehensive Suite of Algorithms

Raw data → Prepare data → Cleaned data → Choose an algorithm → Training algorithm

Validation method ← Choose a validation method ← Model ← Fit a model ← Training algorithm

Validation method → Examine fit and update → Satisfied? — No → Update model → Model

Satisfied? — Yes → Use fitted model for predictions

New data → Use fitted model for predictions → Prediction
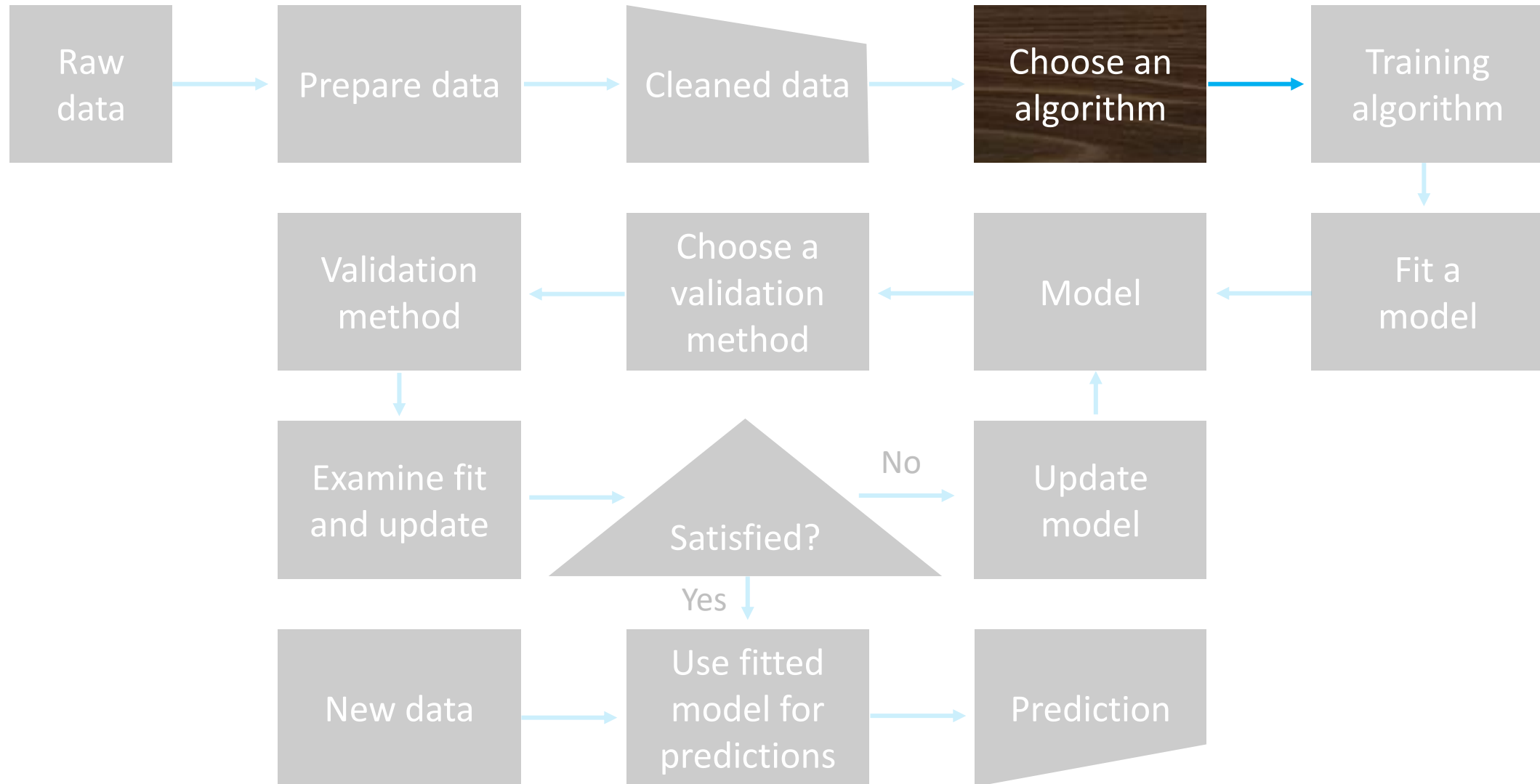
# Find Best Model Parameters

Raw data → Prepare data → Cleaned data → Choose an algorithm → Training algorithm

Training algorithm → Fit a model → Model → Choose a validation method → Validation method

Validation method → Examine fit and update → Satisfied?

Satisfied? → No → Update model → Model

Satisfied? → Yes → Use fitted model for predictions

New data → Use fitted model for predictions → Prediction

# Invoke fit() Or fit_transform()

Raw data → Prepare data → Cleaned data → Choose an algorithm → Training algorithm

Training algorithm → Fit a model → Model → Choose a validation method → Validation method

Validation method → Examine fit and update → Satisfied?

Satisfied? → No → Update model → Model

Satisfied? → Yes → Use fitted model for predictions

New data → Use fitted model for predictions → Prediction
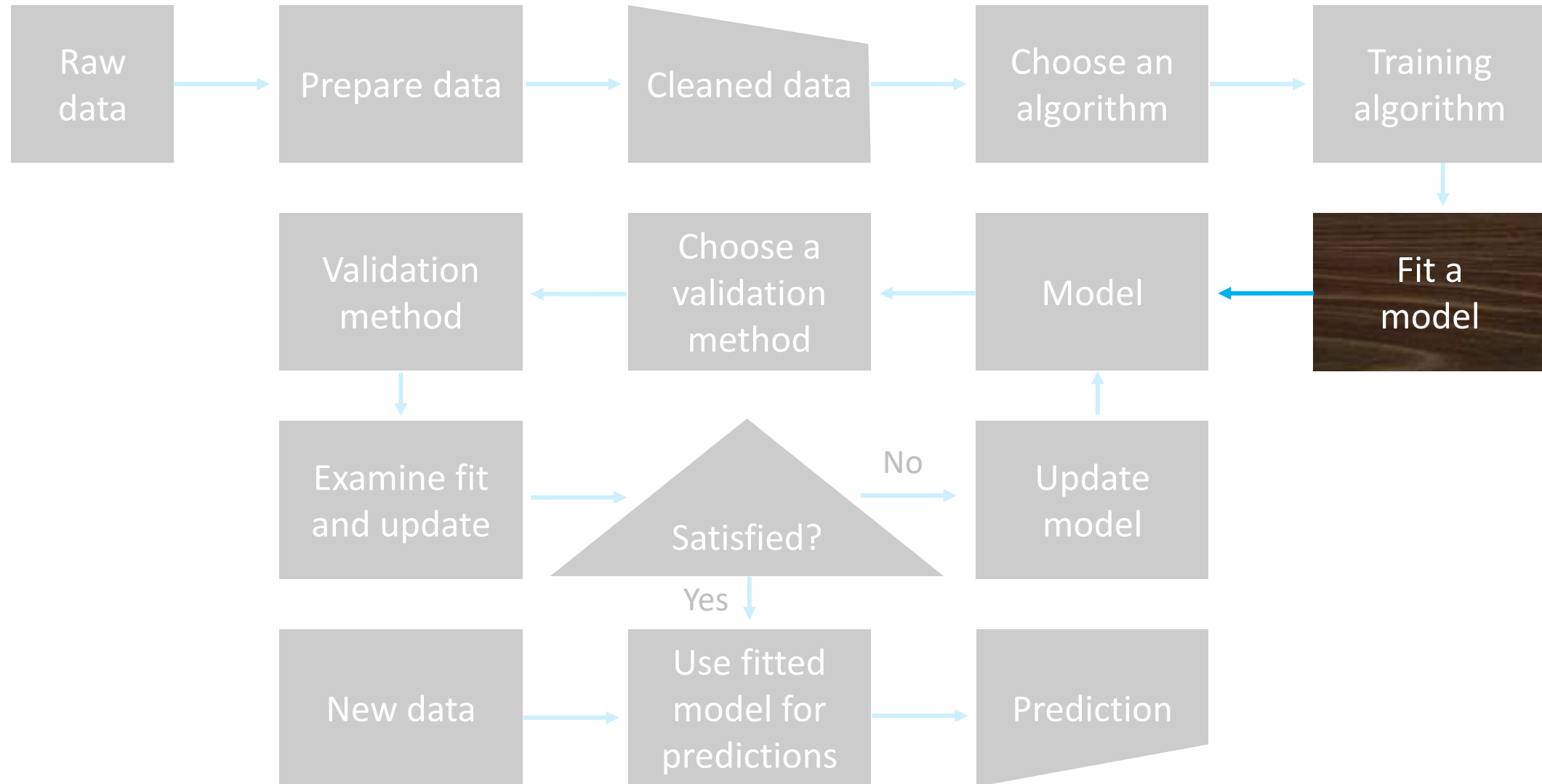
# Trained Model Available
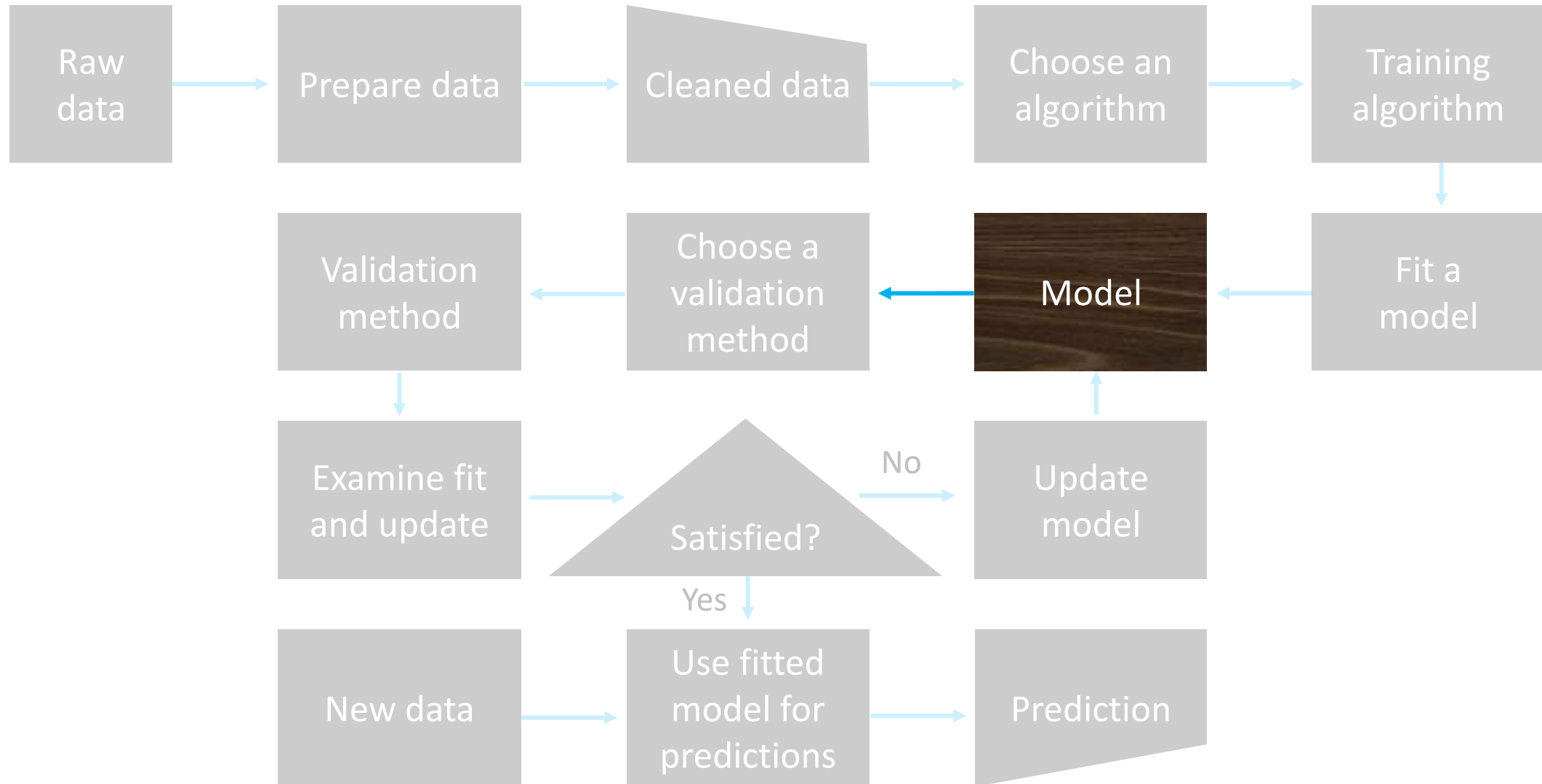
# Comprehensive Suits of Cross-validation Tools
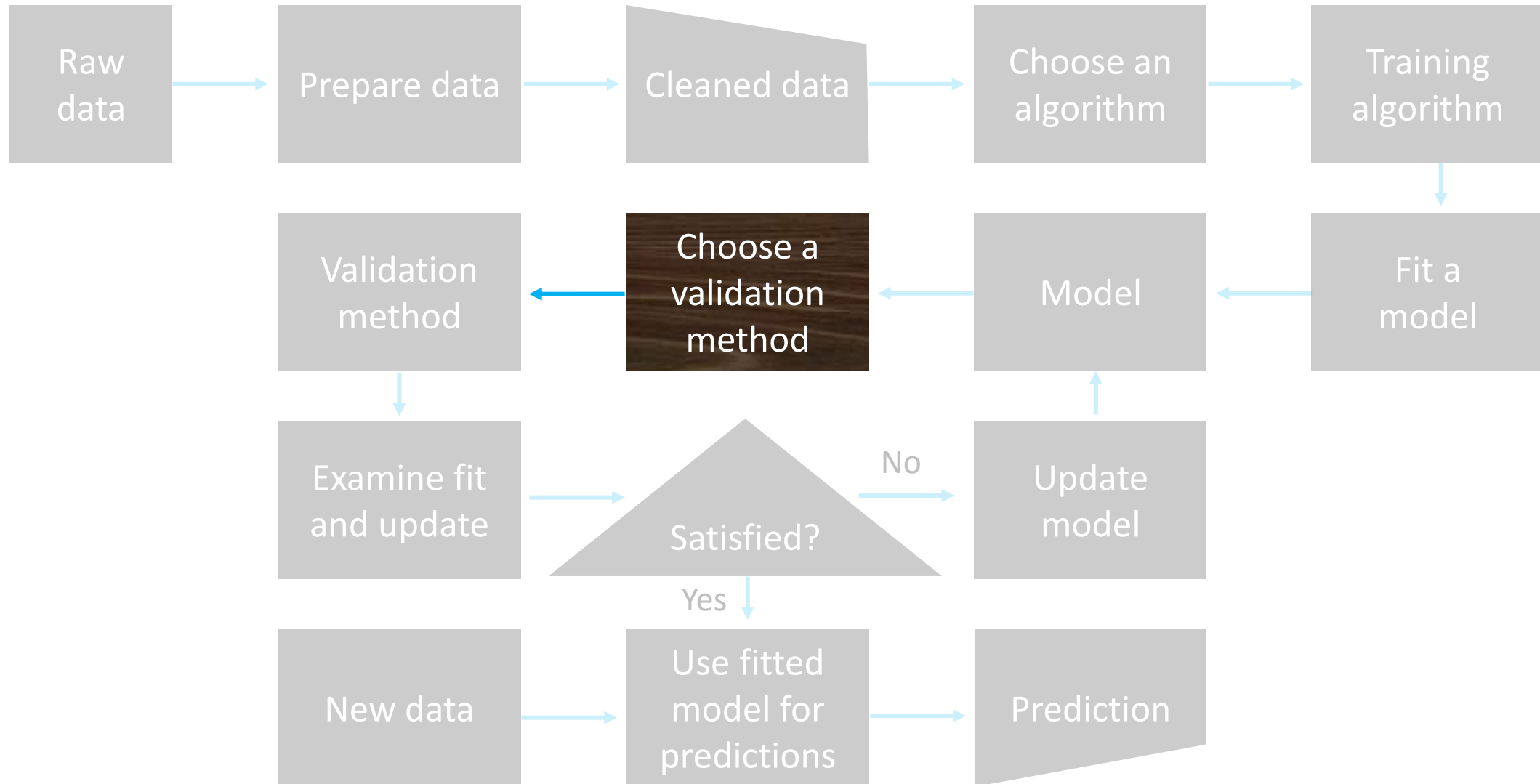
# Cross-validation, K-fold, Group K-fold

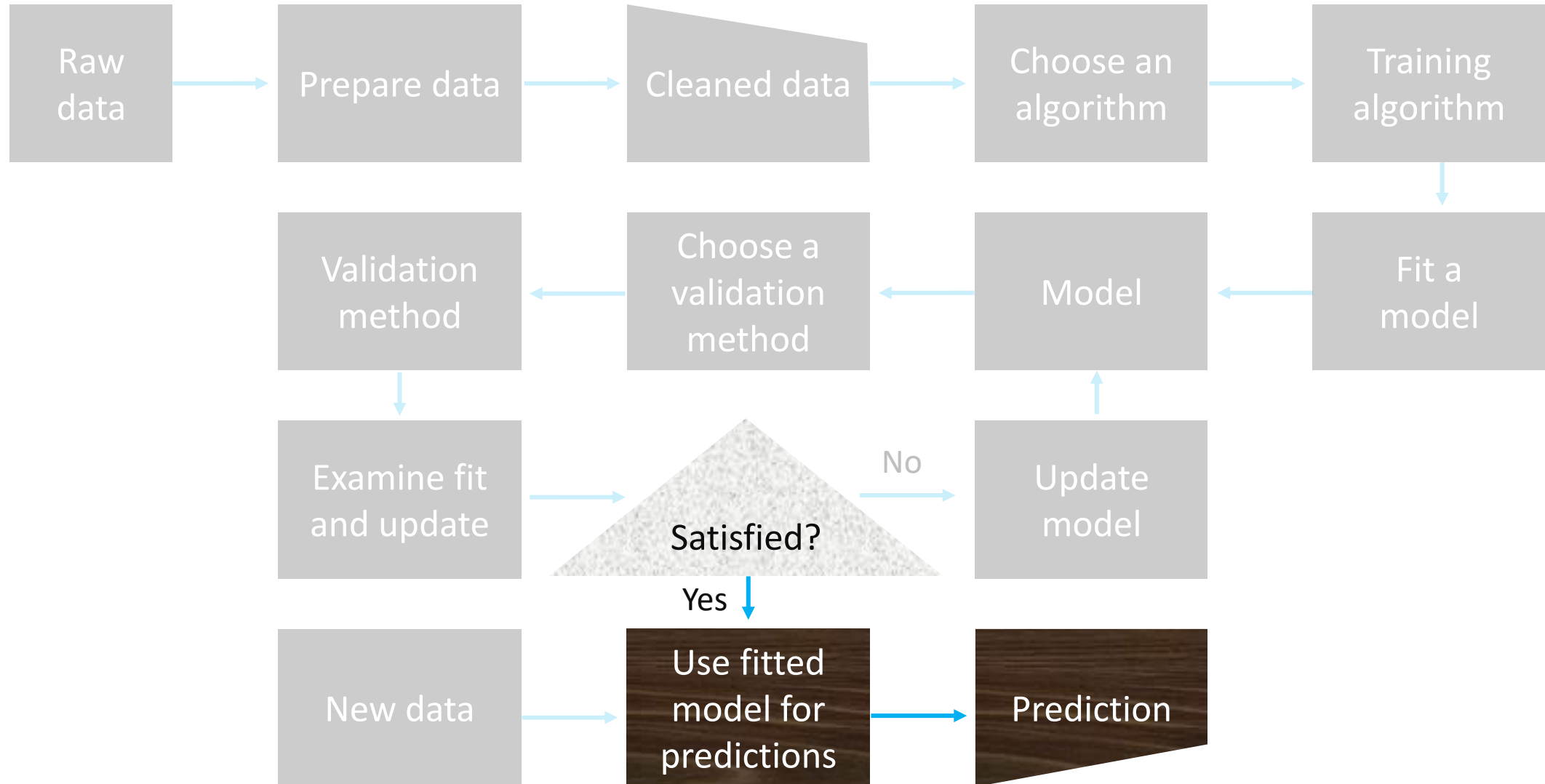Raw data → Prepare data → Cleaned data → Choose an algorithm → Training algorithm

Validation method ← Choose a validation method ← Model ← Fit a model

Validation method → Examine fit and update → Satisfied?

Satisfied? — No → Update model → Model

Satisfied? — Yes → Use fitted model for predictions

New data → Use fitted model for predictions → Prediction
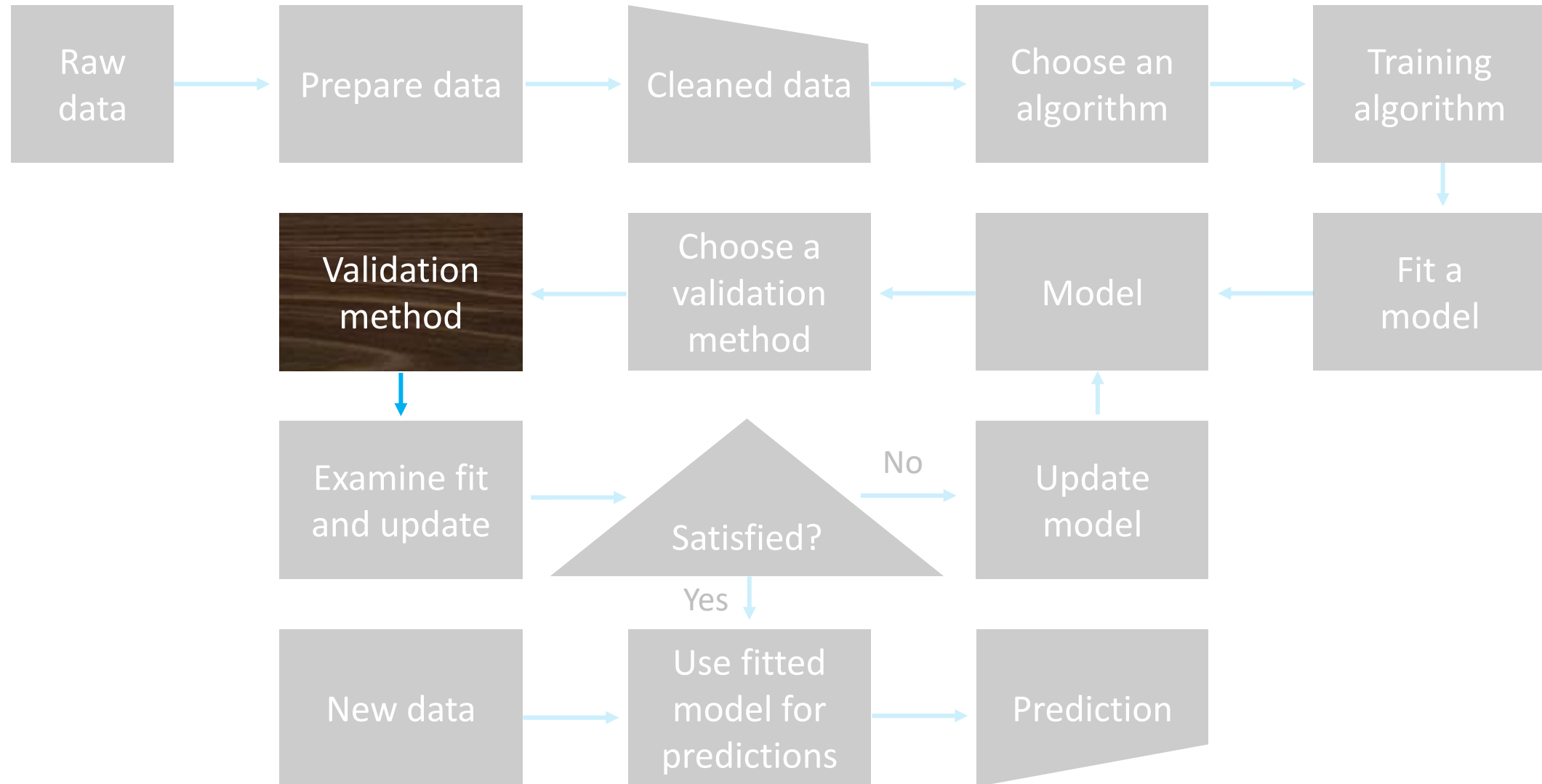
# Metrics for Model Evaluation
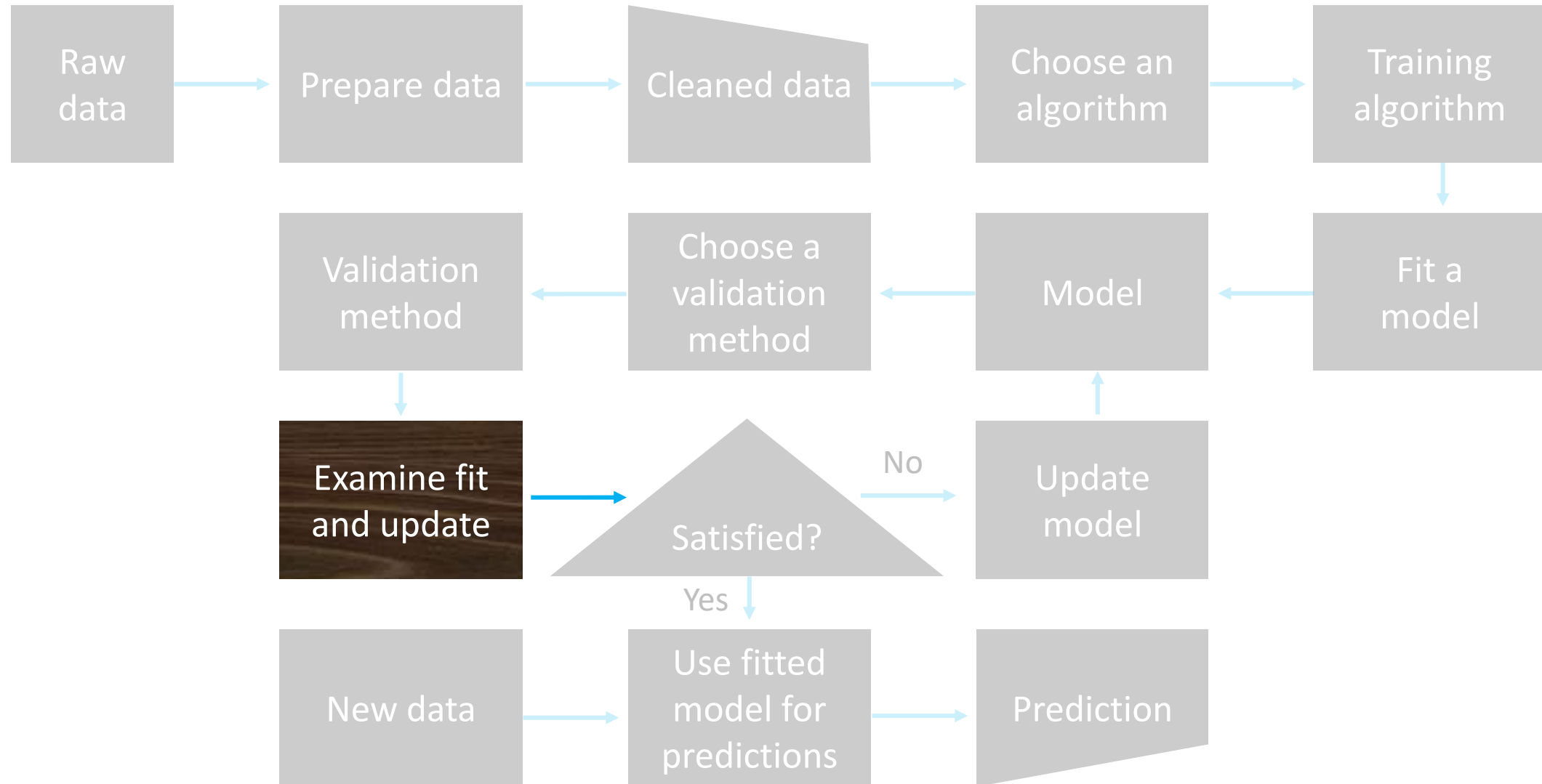
# Rinse and Repeat
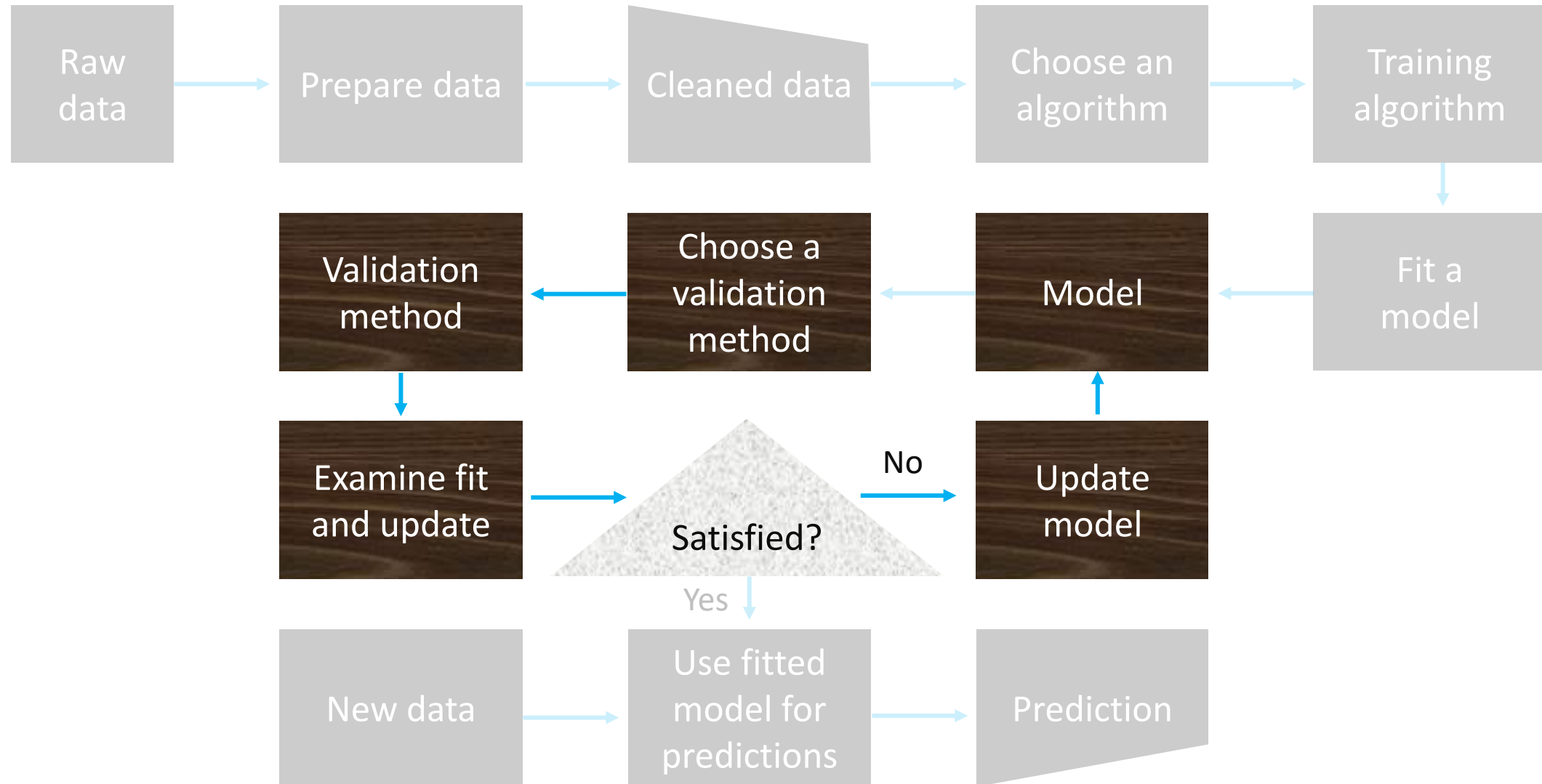
# Fitted Model for Production Use

| Raw data | → | Prepare data | → | Cleaned data | → | Choose an algorithm | → | Training algorithm |
|---|---|---|---|---|---|---|---|---|

| Validation method | ← | Choose a validation method | ← | Model | ← | Fit a model |
|---|---|---|---|---|---|---|

| Examine fit and update | → | Satisfied? | —No→ | Update model |
|---|---|---|---|---|

Yes ↓

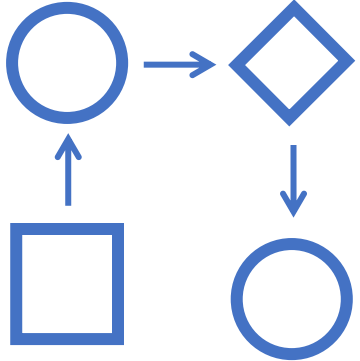| New data | → | Use fitted model for predictions | → | Prediction |
|---|---|---|---|---|

# Invoke Predict () Method
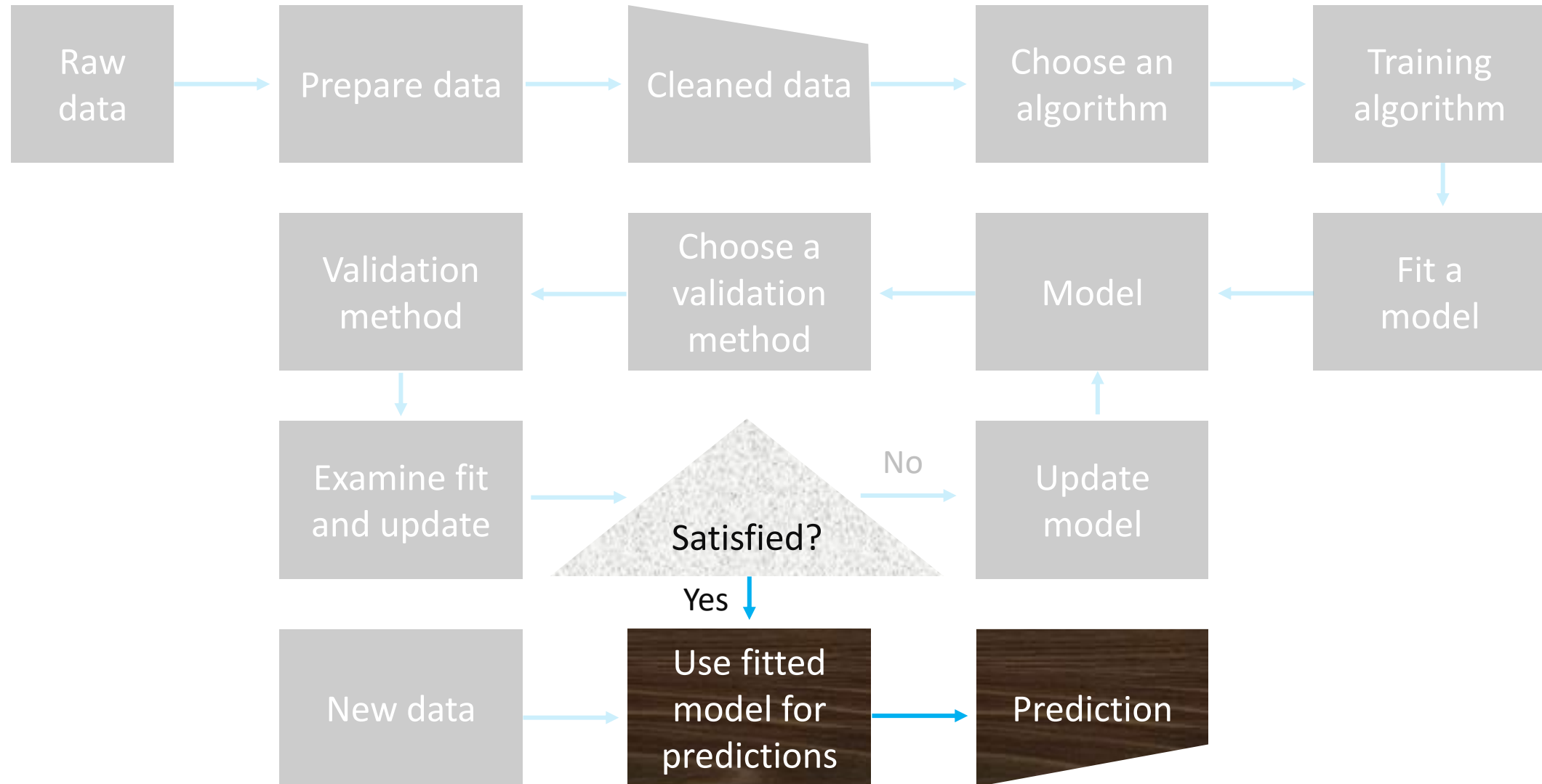
# Invoke Predict () Method
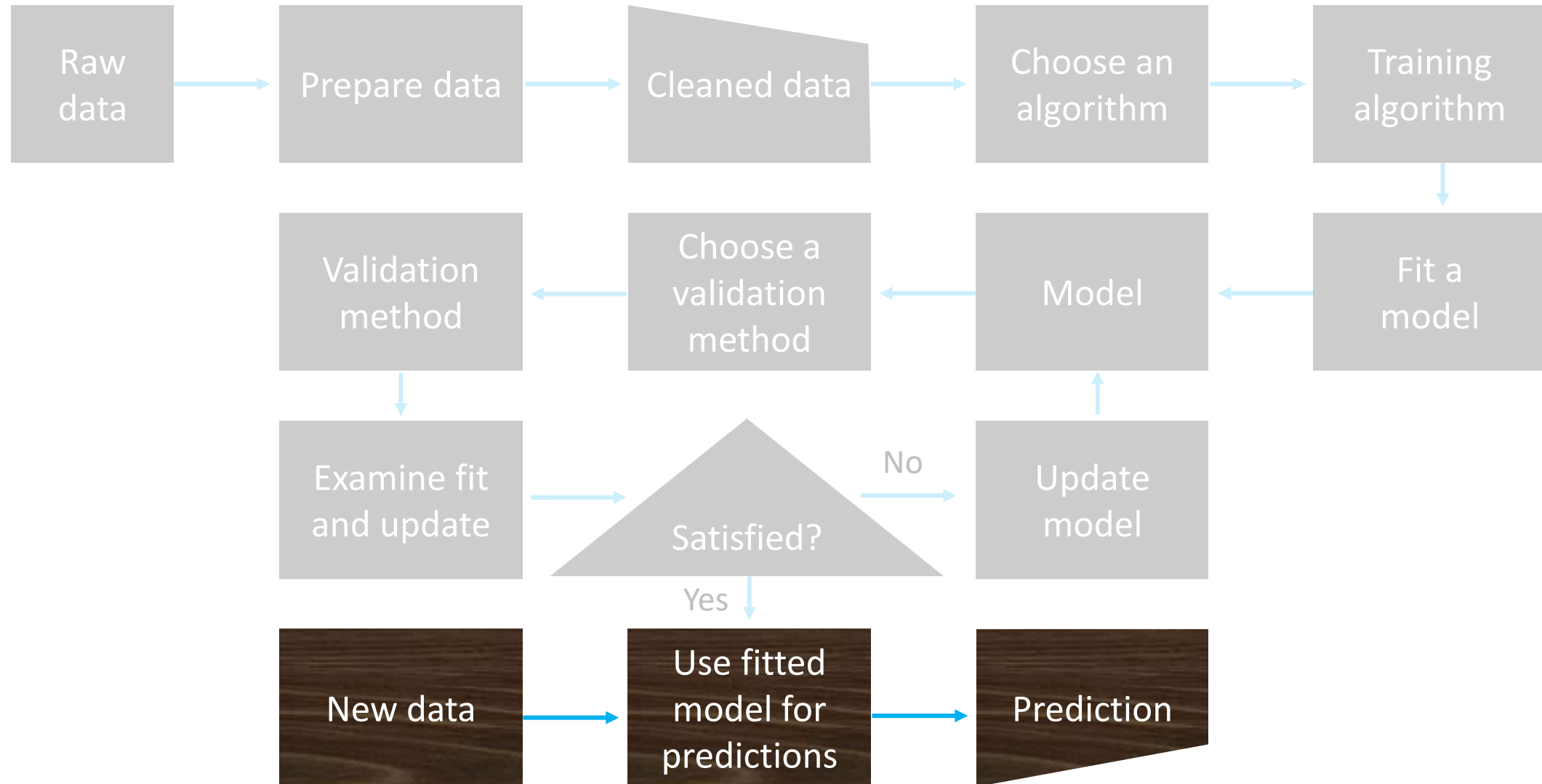
# Scikit-learn Pipeline

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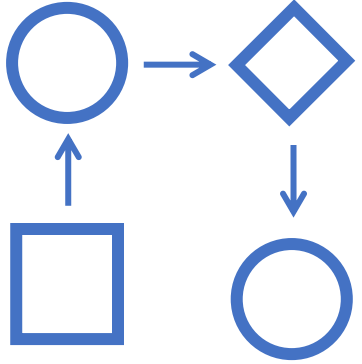
# Pipelines

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