Experimental Design for Data Analysis

DESIGNING AN EXPERIMENT FOR DATA ANALYSIS

Janani Ravi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Hypothesis testing to evaluate proposed explanations for phenomenon

Understanding the T-test to test for differences between categories

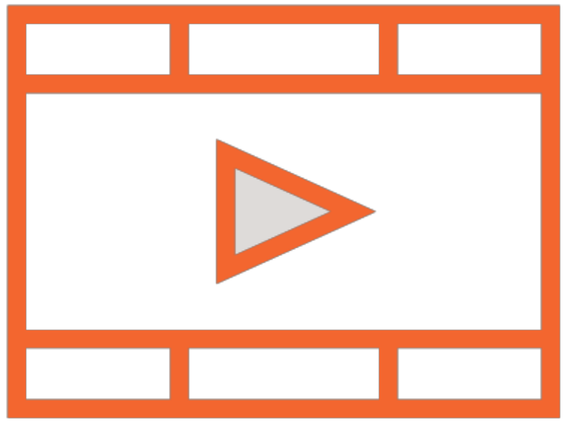
Using ANOVA to test differences across multiple groups

Choosing an algorithm based on prediction target

Understanding the steps involved in building a model

Prerequisites and Course Outline

Prerequisites



Basic Python programming

Basic understanding of ML models

High school math

Course Outline

Positing and testing hypothesis

Framing experiments to build models

Accounting for data biases

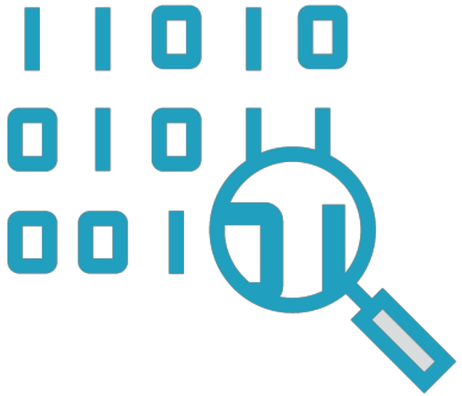
Validating models

Refining models

“My mind is made up. Don’t confuse me with the facts.”

Some powerful person

Thoughtful, Fact-based Point of View



Fact-based

Built with painstakingly collected data

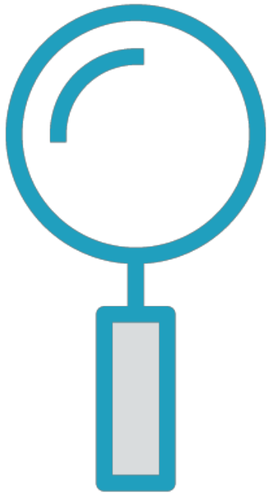
Thoughtful

Balanced, weighing pros and cons

Point of View

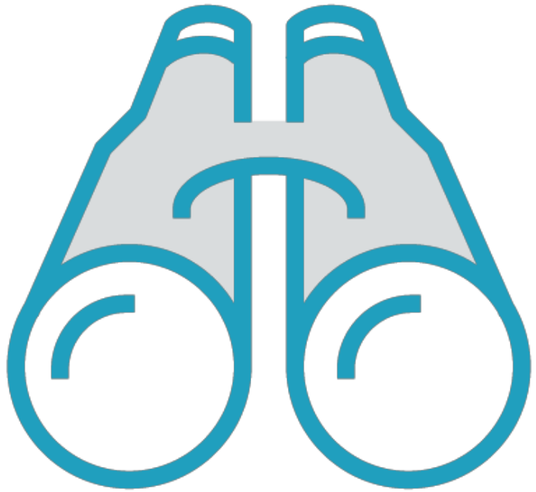
Prediction, recommendation, call to action

Two Sets of Statistical Tools



Descriptive Statistics

Identify important elements in a dataset



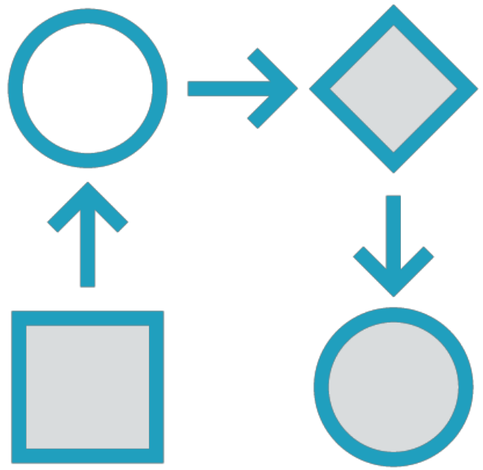
Inferential Statistics

Explain those elements via relationships with other elements

Two Hats of a Data Professional

Find the Dots

Identify important elements in a dataset



Connect the Dots

Explain those elements via relationships with other elements

Connecting the Dots

Explore and pre-process data

Posit hypotheses and build models

Link to real-world data and scenarios

Related Courses on Pluralsight

Explore and pre-process data

Representing, Processing and Preparing Data

Summarizing Data and Deducing Probabilities

Combining and Shaping Data

Related Courses on Pluralsight

Posit hypotheses and build models

This course

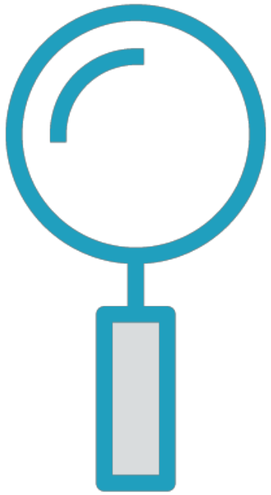
Related Courses on Pluralsight

Link to real-world data and scenarios

Communicating Data Insights

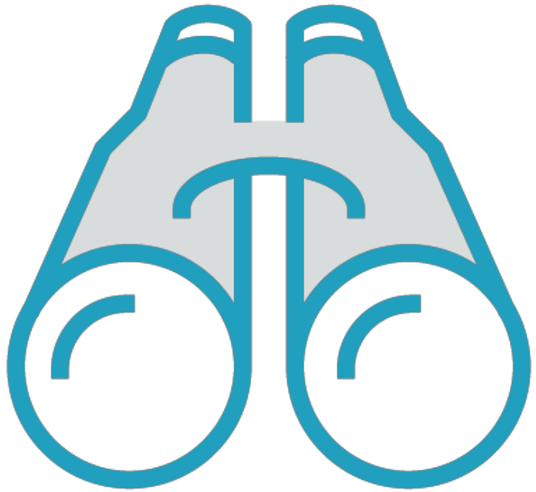
Hypothesis Testing

Two Sets of Statistical Tools



Descriptive Statistics

Identify important elements in a dataset



Inferential Statistics

Explain those elements via relationships with other elements

From Statistics to ML

Descriptive Statistics

Explore the data

No points-of-view yet

Rule-based Learning Models

Frame rules based on the data

Performed by experts - risk of too much certainty

Inferential Statistics

Frame hypotheses and test them

Tentatively evaluating many points-of-view

Machine Learning Models

Build models that change with the data

Full circle - back to no points-of-view

Hypothesis

Proposed explanation for a phenomenon

Hypothesis

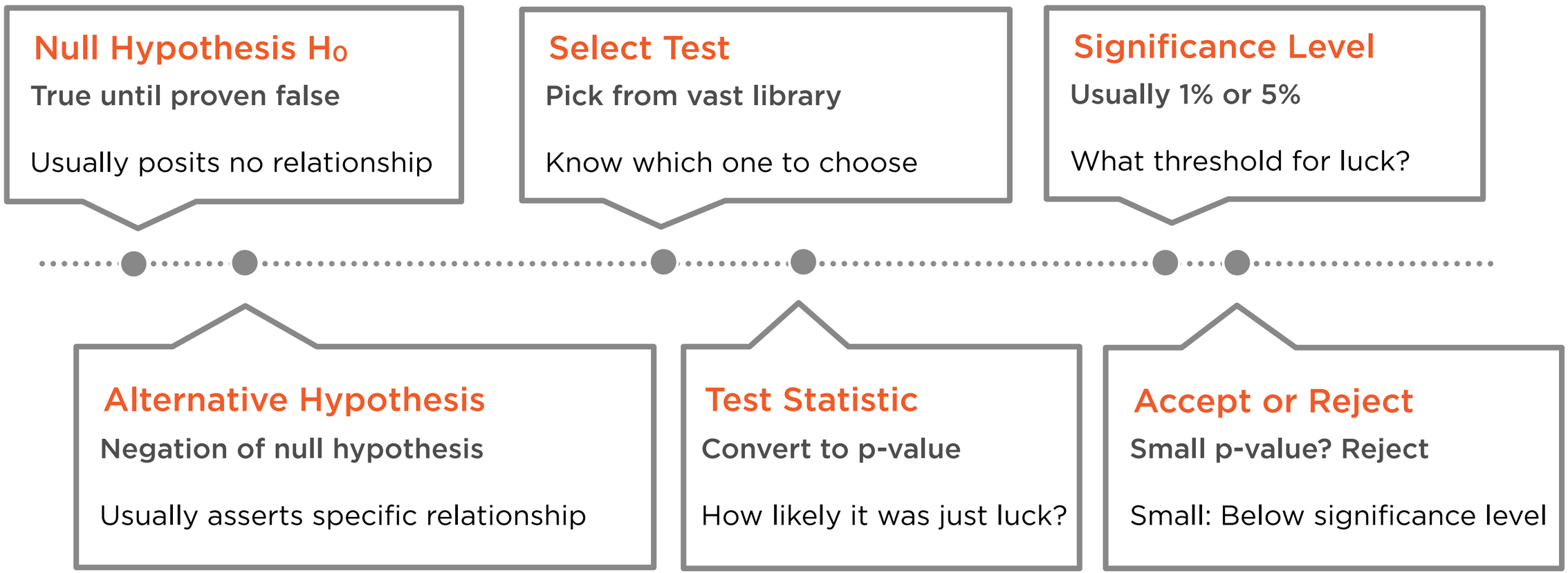
Proposed explanation

Objectively testable

Singular - hypothesis

Plural - hypotheses

Hypothesis Testing



Lady Tasting Tea

Lady tasting tea: famous experiment

Was tea added before or after milk?

Muriel Bristol claimed she could tell

Lady Tasting Tea

Null Hypothesis

(H_0)

The lady cannot tell if milk was poured first

Alternate Hypothesis

(H_1)

The lady can tell if milk was poured first

Lady Tasting Tea

Null Hypothesis

The lady **cannot** tell if the milk was poured first

Alternate Hypothesis

The lady **can** tell if the milk was poured first

It is good practice to assume that the null hypothesis is correct unless proven otherwise

Lady Tasting Tea

Null Hypothesis

The lady **cannot** tell if the milk was poured first

Alternate Hypothesis

The lady **can** tell if the milk was poured first

It is good practice to assume that the null hypothesis is correct unless proven otherwise

Lady Tasting Tea

Null Hypothesis H_0

"Lady cannot tell difference"

Can't tell if milk poured first

Select Test

8 cups, 4 of each type

Lady got all 8 correct

Significance Level

Choose 5% significance level

Part of design of experiment

Alternative Hypothesis

"Lady can tell difference"

Can indeed discern if milk poured first

Test Statistic

p-value = 1/70 = 1.4%

$^{8}C_{4}$ = 70 combinations

Accept or Reject

1.4% < 5% => Reject H_0

Lady can indeed tell difference

Lady Tasting Tea

Experiment proved that she could

Conducted by Sir Ronald Fisher

(considered founder of modern statistics)

Errors in Hypothesis Testing

		Decision about Null Hypothesis	
		REJECT	DON'T REJECT
Null Hypothesis is actually	TRUE	Type I error	Correct Inference
	FALSE	Correct Inference	Type II error

Errors in Hypothesis Testing

		Decision about Null Hypothesis	
		REJECT	DON'T REJECT
Null Hypothesis is actually	TRUE	Type I error	Correct Inference
	FALSE	Correct Inference	Type II error

Claim the lady can tell the difference based on spurious test results which are not statistically significant

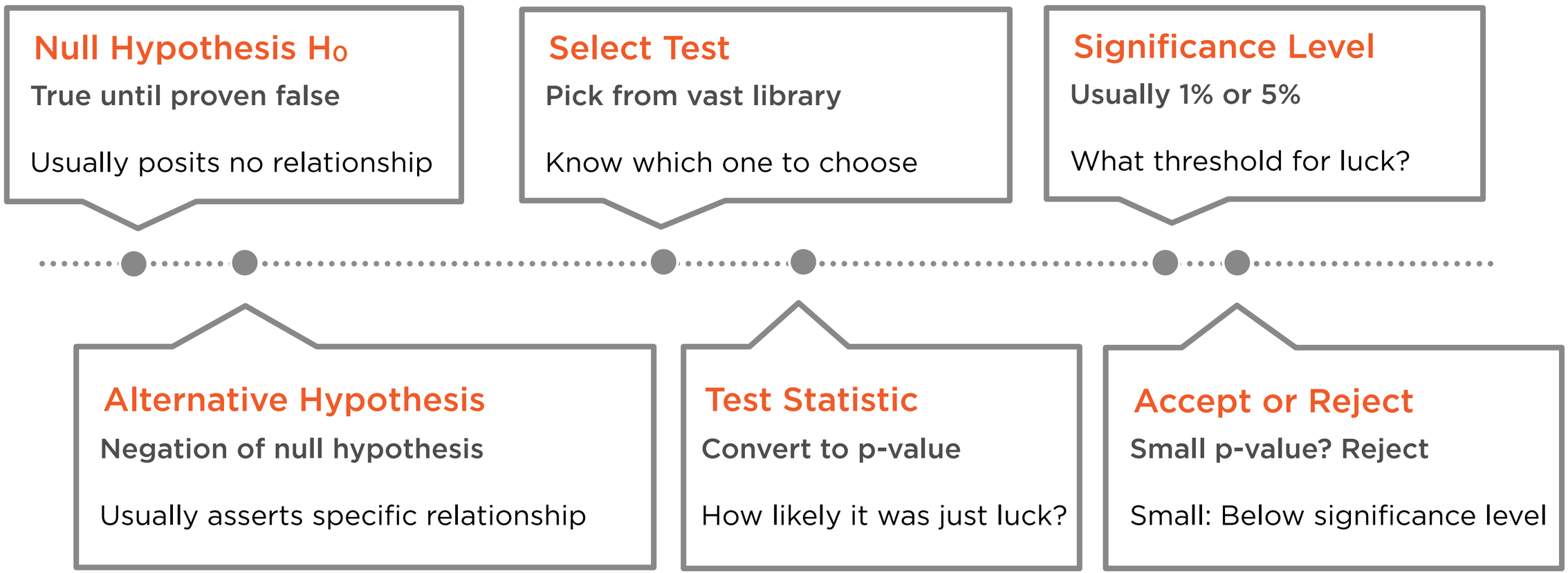
Errors in Hypothesis Testing

		Decision about Null Hypothesis	
		REJECT	DON'T REJECT
Null Hypothesis is actually	TRUE	Type I error	Correct Inference
	FALSE	Correct Inference	Type II error

Fail to realize that the test for the alternative hypothesis was statistically significant

The T-test

Hypothesis Testing



Hypothesis Testing

Null Hypothesis H_0

True until proven false

Usually posits no relationship

Select Test

Pick from vast library

Know which one to choose

Significance Level

Usually 1% or 5%

What threshold for luck?

Alternative Hypothesis

Negation of null hypothesis

Usually asserts specific relationship

Test Statistic

Convert to p-value

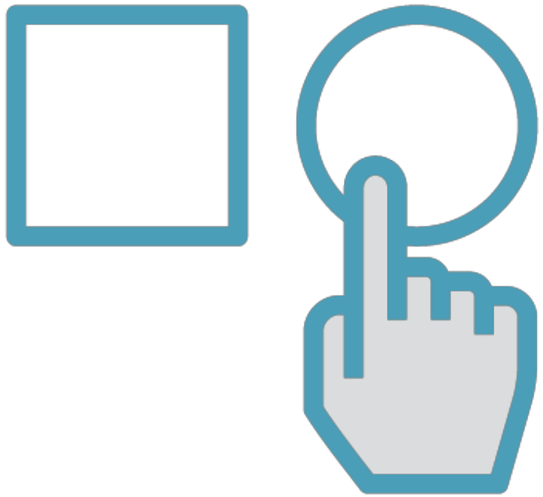
How likely it was just luck?

Accept or Reject

Small p-value? Reject

Small: Below significance level

Statistical Test Selection



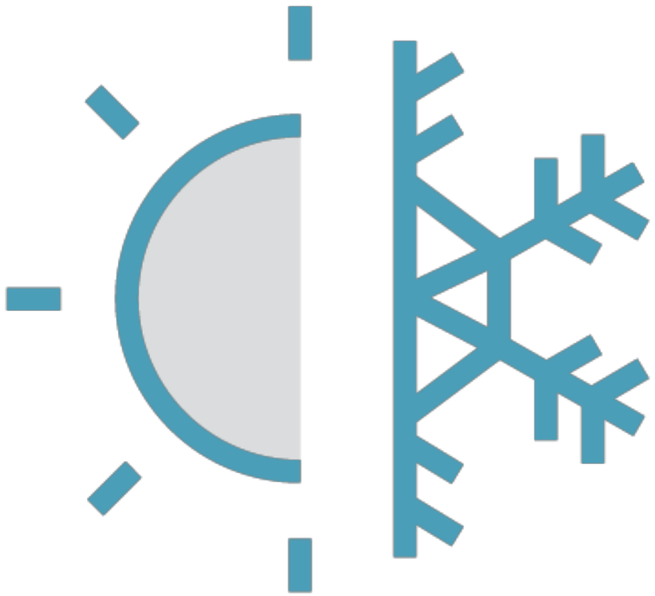
There are tests for pretty much everything

Developed by statisticians to be sound

Knowing which one to use is hard

Actually using them is relatively easy

T-tests



Most common, simple statistical tests out there

Used to learn about averages across two categories

Also tells whether the differences are significant

T-tests

Average male baby birth weight =
Average female baby birth weight?

Is the difference statistically significant?

T-tests

T-statistic

- Score which indicates the difference in means

P-value

- Whether the T-statistic is significant
- Low p-values of <5% mean the result cannot be due to chance

Types of T-tests

One sample location test

Two sample location test

Paired difference test

Regression coefficient test

One sample location test

One-sample location test

- What is the average weight of babies born in a certain town?
- Is it different from the average of the general population?

Two sample location test

Two-sample location test (independent samples t-test)

- Is the average weight of babies in Town A different from Town B?

Paired difference test

Paired difference test

- Is the average weight of babies born in winter different from babies born in summer?

Regression coefficient test

Regression coefficient test

- Is the coefficient of any of the independent variables > 0?

Mean and Variance

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

These statistics only apply to the sample of data, and so are known as sample statistics

The corresponding figures for all possible data points out there are called population statistics

From Sample to Population

Population

All the data out there in the universe

Sample

A subset - hopefully representative - of the population

From Sample to Population

Population

Representative Sample

Biased Sample

From Sample to Population

Sample Mean

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Population Mean

$$\mu = ?$$

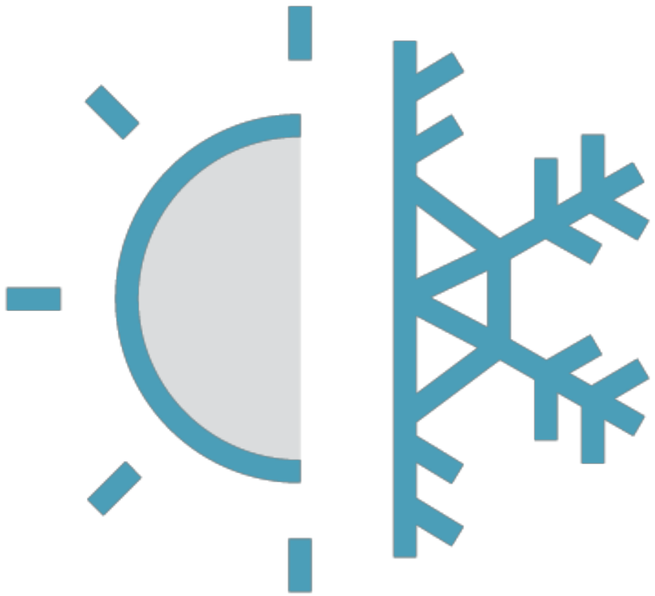
From Sample to Population

Sample Mean

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Population Mean

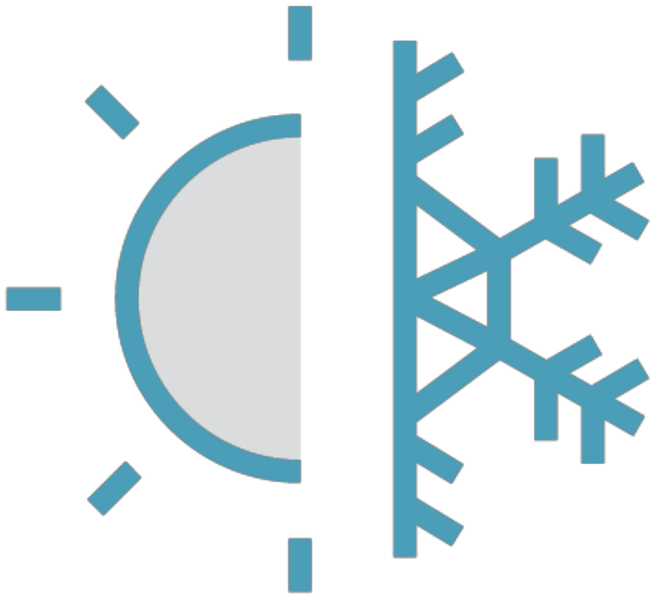
T-tests Assumptions



Notably, that

- populations are normal
- samples are representative
- samples are randomly drawn

T-tests



Work best for two group comparisons

Comparing multiple groups gets tricky

- need many pairwise tests
- increases likelihood of Type 1 error (alpha inflation)

For multiple groups, just use ANOVA

ANOVA

T-tests are useful to compare differences between **two** groups

Running **multiple** significance tests to compare across many groups is **risky**

ANOVA

ANalysis ***O***f ***VA***riance

ANOVA

Looks across multiple groups of populations, compares their means to produce one score and one significance value

ANOVA

Looks across multiple groups of populations, compares their means to produce one score and one significance value

Diabetes Risk

In order to compare across 3 groups the we'll need to perform multiple T-tests

Diabetes Risk

Perform a single ANOVA test to know whether the risk of diabetes is significantly different between these groups

ANOVA Hypotheses

Null Hypothesis

(H_0)

H_0: All groups of patients are at an equal risk of diabetes

Alternate Hypothesis

(H_1)

H_0: All groups of patients are NOT at an equal risk of diabetes

ANOVA

Looks across multiple groups of populations, compares their means to produce one score and one significance value

F-statistic

$$F = \frac{\text{Variance between groups}}{\text{Variance within a group}}$$

F-statistic

If the groups are similar, F ~ 1

If the groups are different, F will be large

P-value

Significance of the F-statistic

Smaller p-values indicate that the results are not due to chance

Large F-statistic and small p-value - means the null hypothesis can be rejected

ANOVA Hypotheses

Large F-statistic and small p-values < 0.05 significance level

Accept the alternative hypothesis and reject the null hypothesis

Alternate Hypothesis

(H_1)

H_0: All groups of patients are NOT at an equal risk of diabetes

ANOVA Hypotheses

Null Hypothesis

(H_0)

H_0: All groups of patients are at an equal risk of diabetes

Small F-statistic and large p-values > 0.05 significance level

Accept the null hypothesis and reject the alternative hypothesis

One-way ANOVA helps compare means across two or more groups

A **single** categorical variable is used to split the population into these groups

One-way ANOVA Assumptions

Notably, that

- populations are normal
- samples are representative
- samples are randomly drawn
- variances of the population are constant

Two-way ANOVA

Examines the influence of two different independent variables on one continuous dependent variable

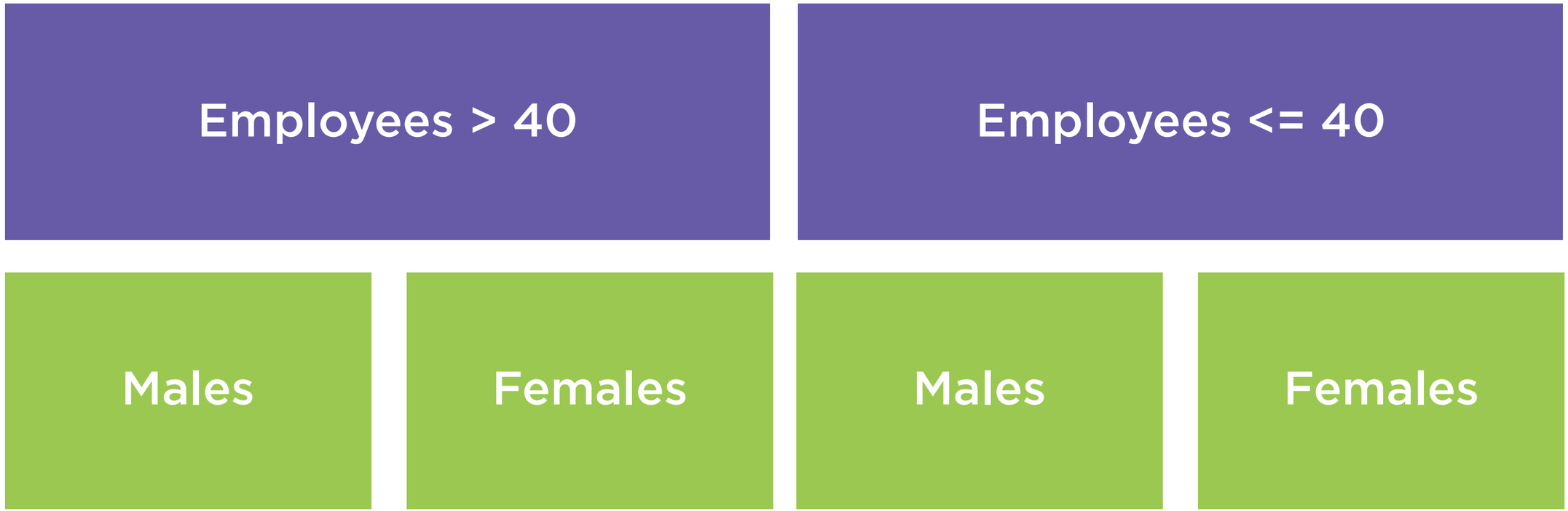
Two-way ANOVA

Examines the influence of two different independent variables on one continuous dependent variable

Two-way ANOVA

Employees > 40	Employees <= 40
Males	Females

Two-way ANOVA



Two-way ANOVA Hypotheses

Null Hypothesis (H_{01})

Null Hypothesis (H_{02})

Null Hypothesis (H_{03})

H_{01}: All groups have equal levels of stress

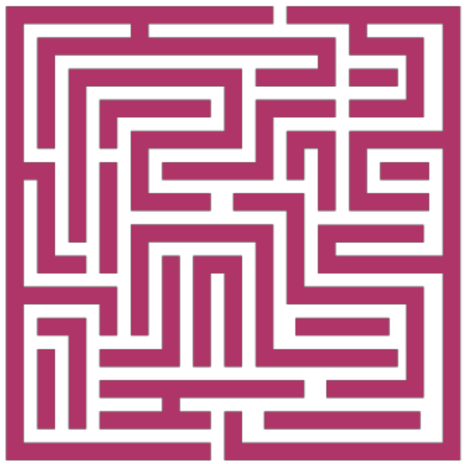
H_{02}: All ages have equal levels of stress

H_{03}: There is no interaction between age and gender

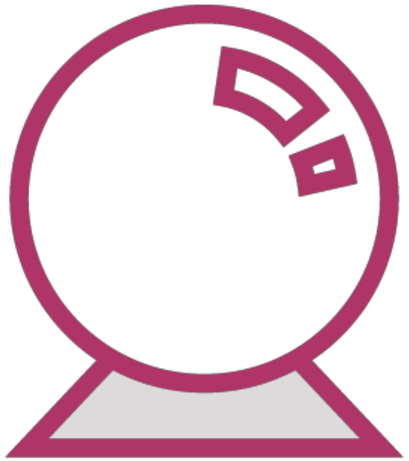
Common Machine Learning Workflows

A machine learning algorithm is an algorithm that is able to learn from data

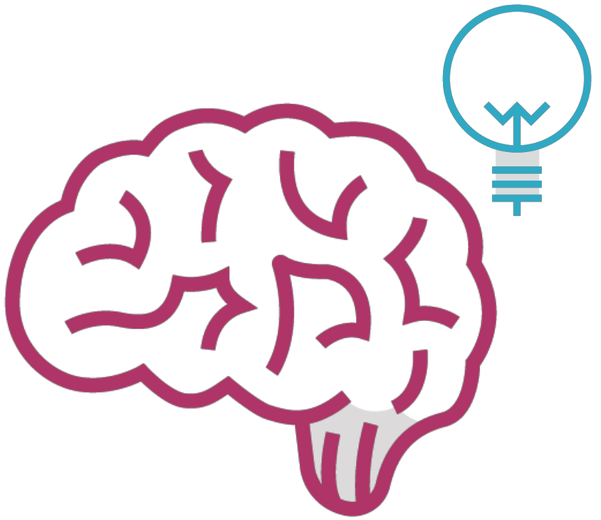
Machine Learning



Work with a huge maze of data

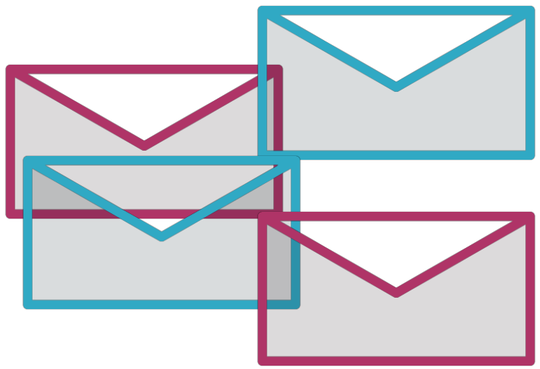


Find patterns

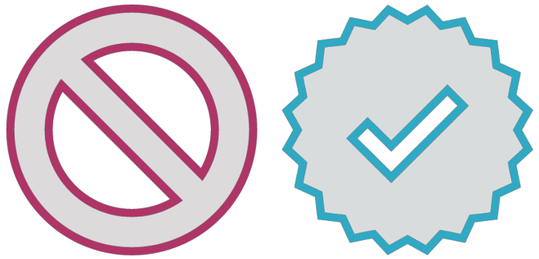


Make intelligent decisions

Machine Learning



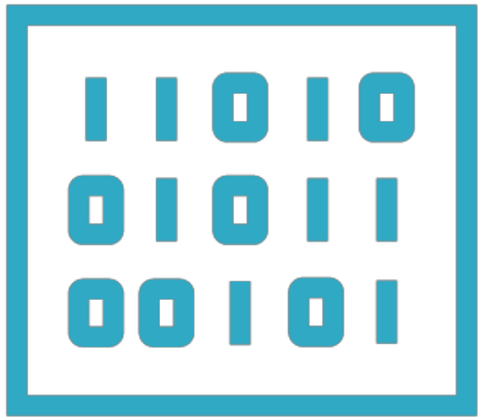
Emails on a server



Spam or Ham?

Trash or Inbox

Machine Learning



Images represented as pixels

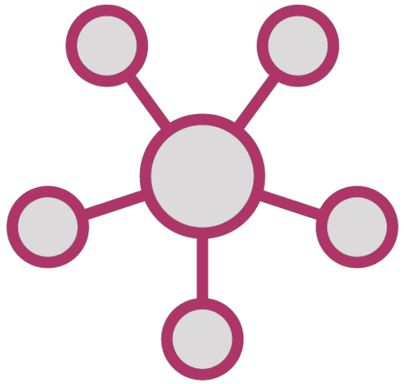
Identify edges, colors, shapes

A photo of a little girl

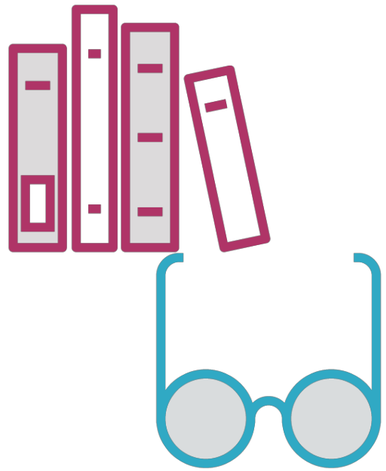
Types of Machine Learning Problems

Classification

Regression

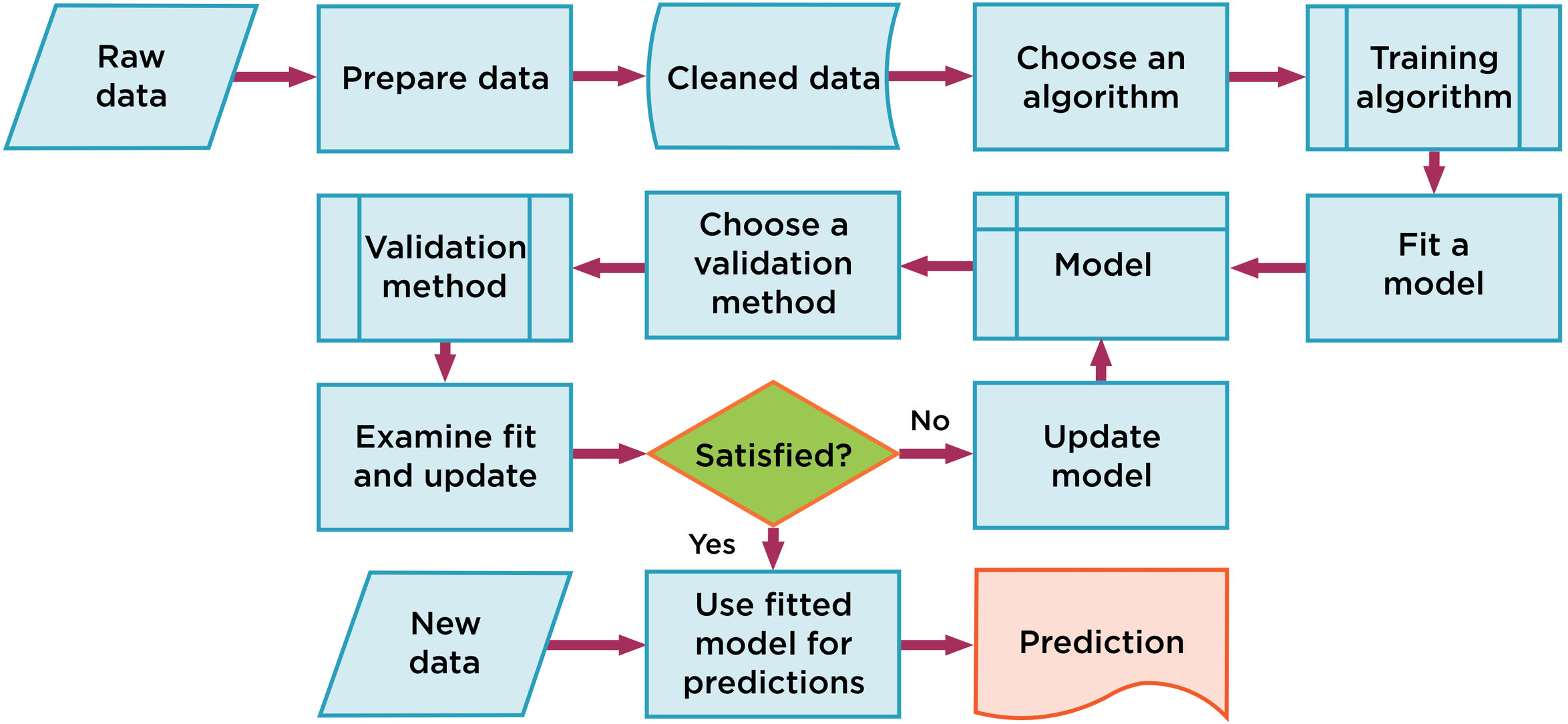


Clustering

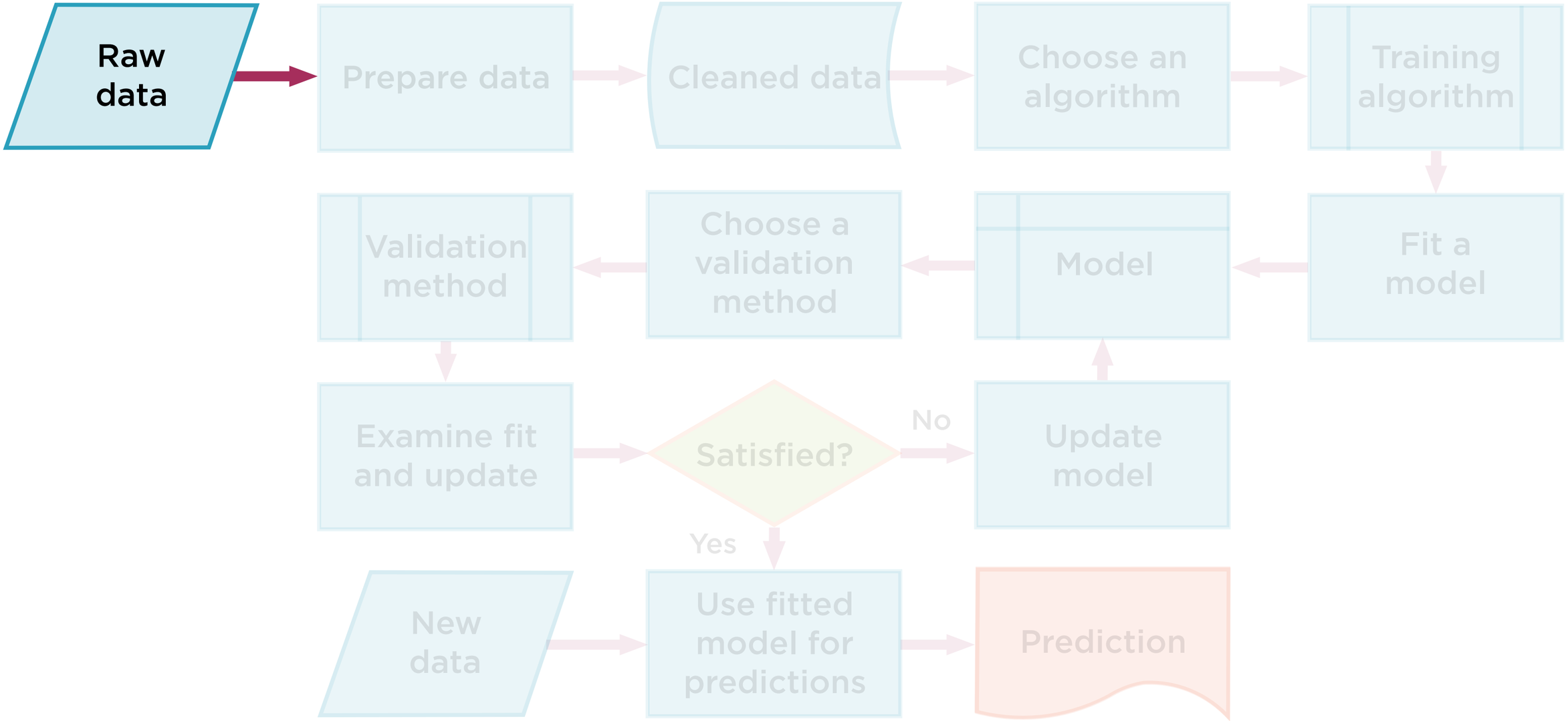


Dimensionality Reduction

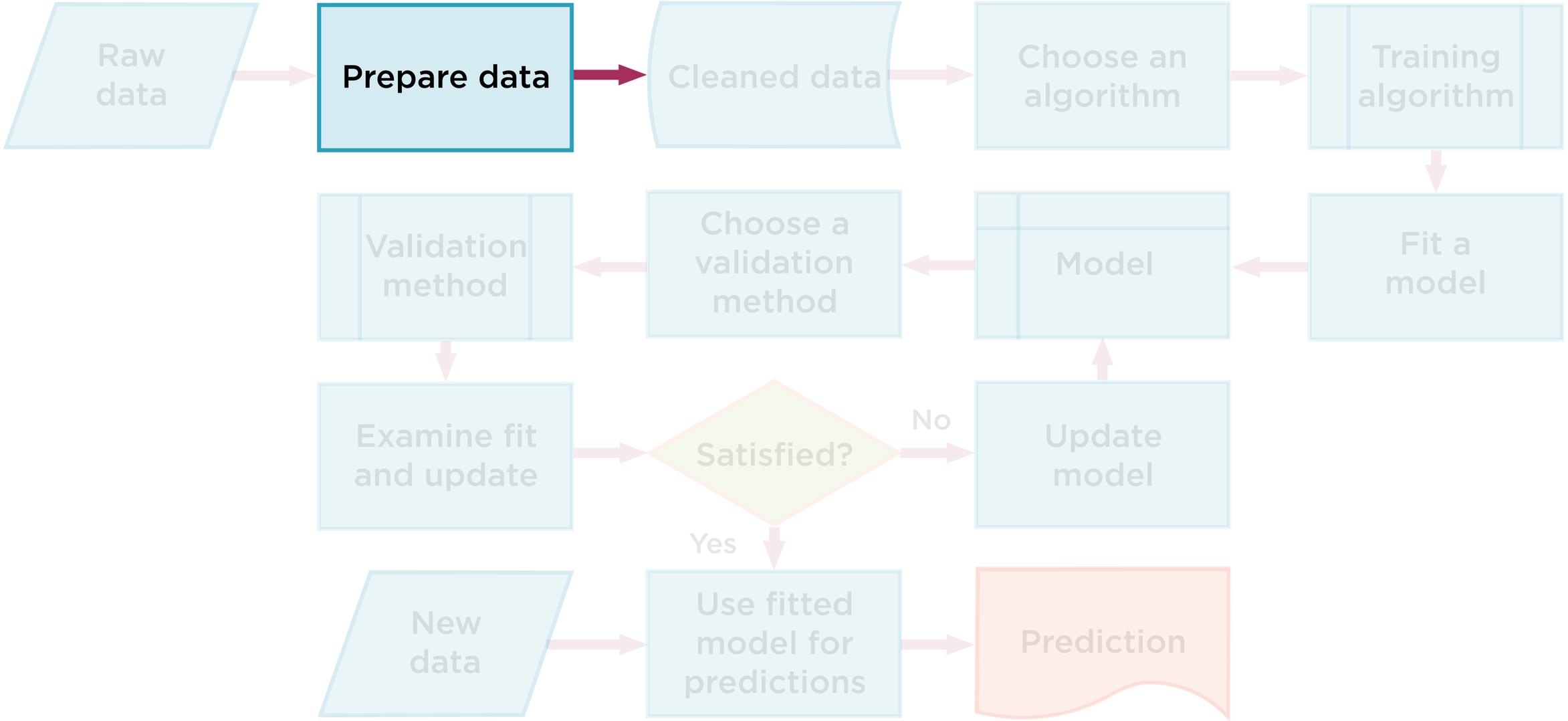
Basic Machine Learning Workflow



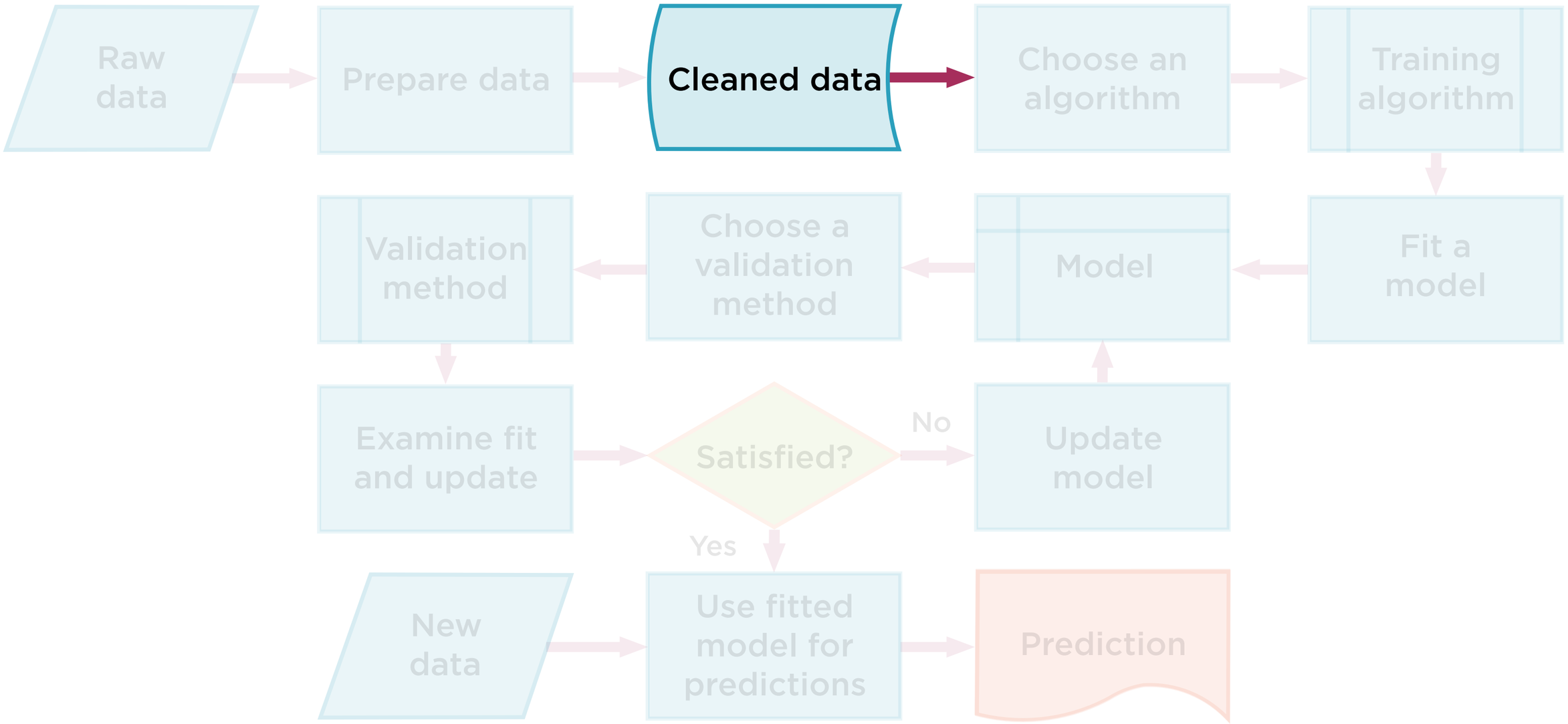
What Data Do You Have to Work With?



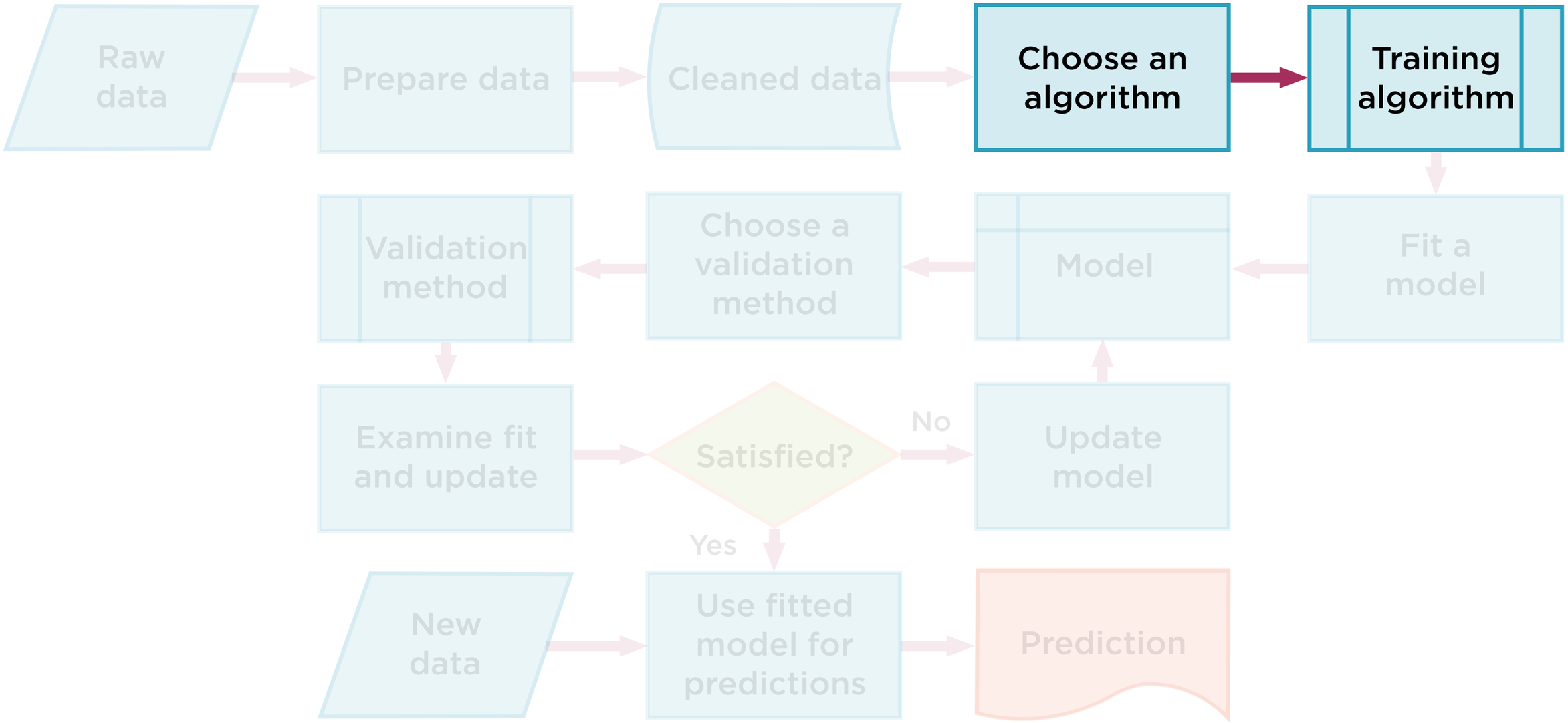
Load and Store Data



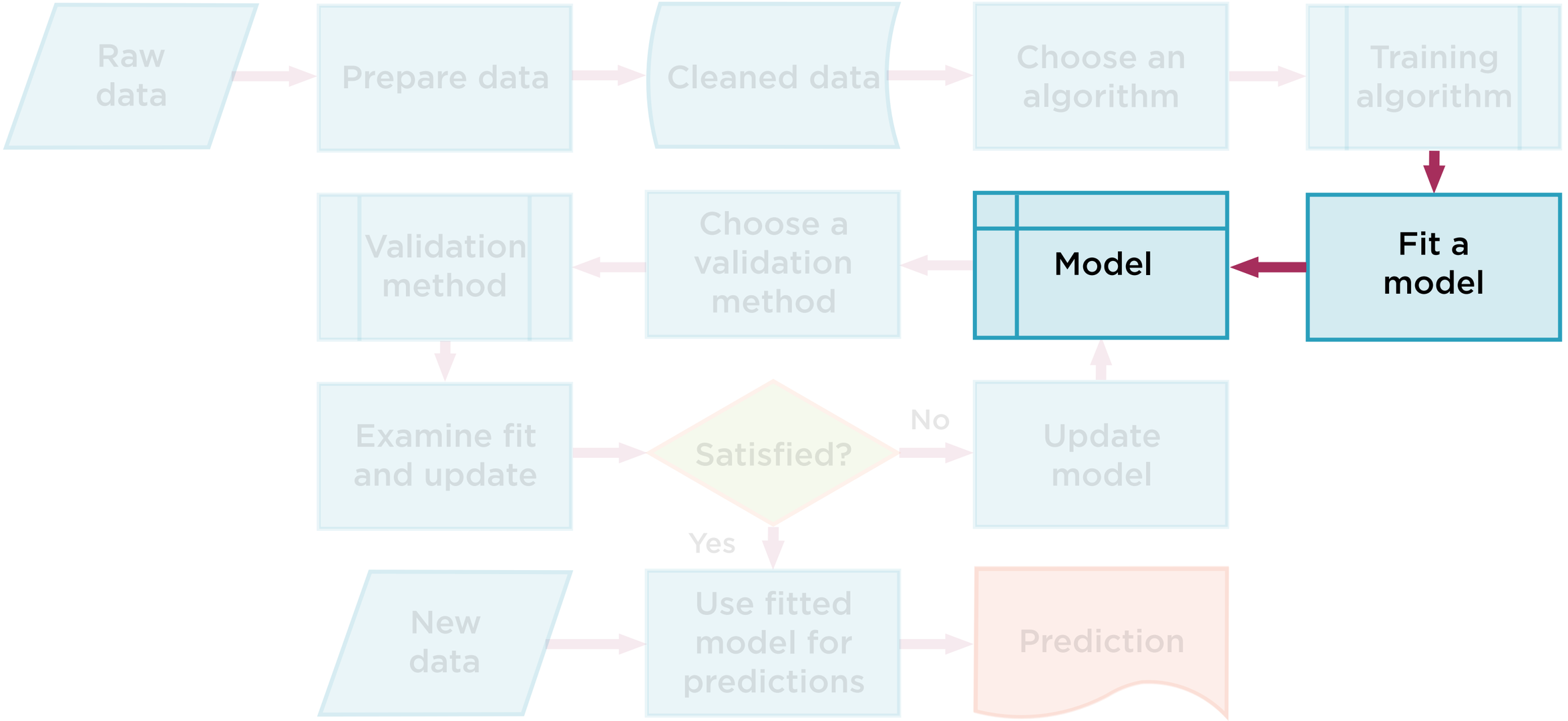
Data Preprocessing



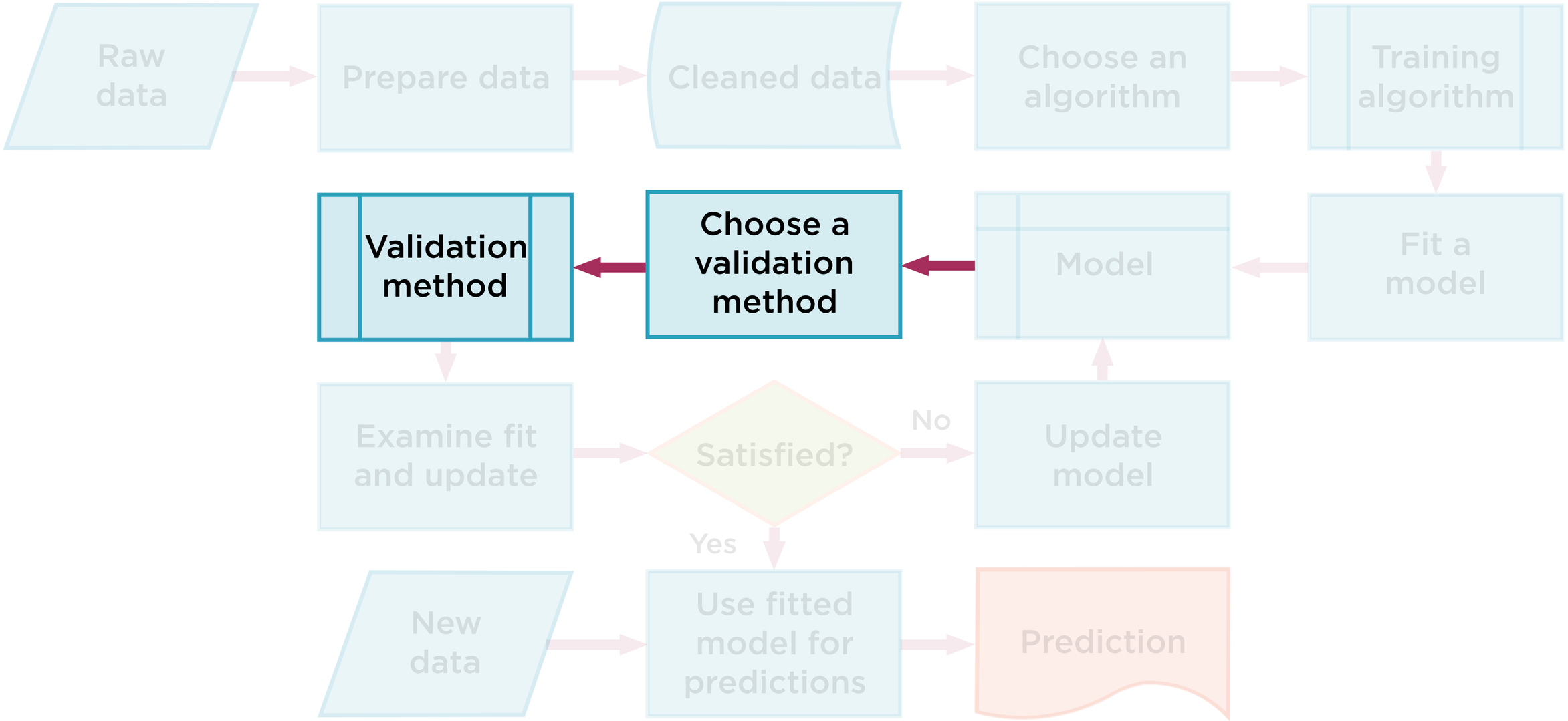
Decision Trees, Support Vector Machines?



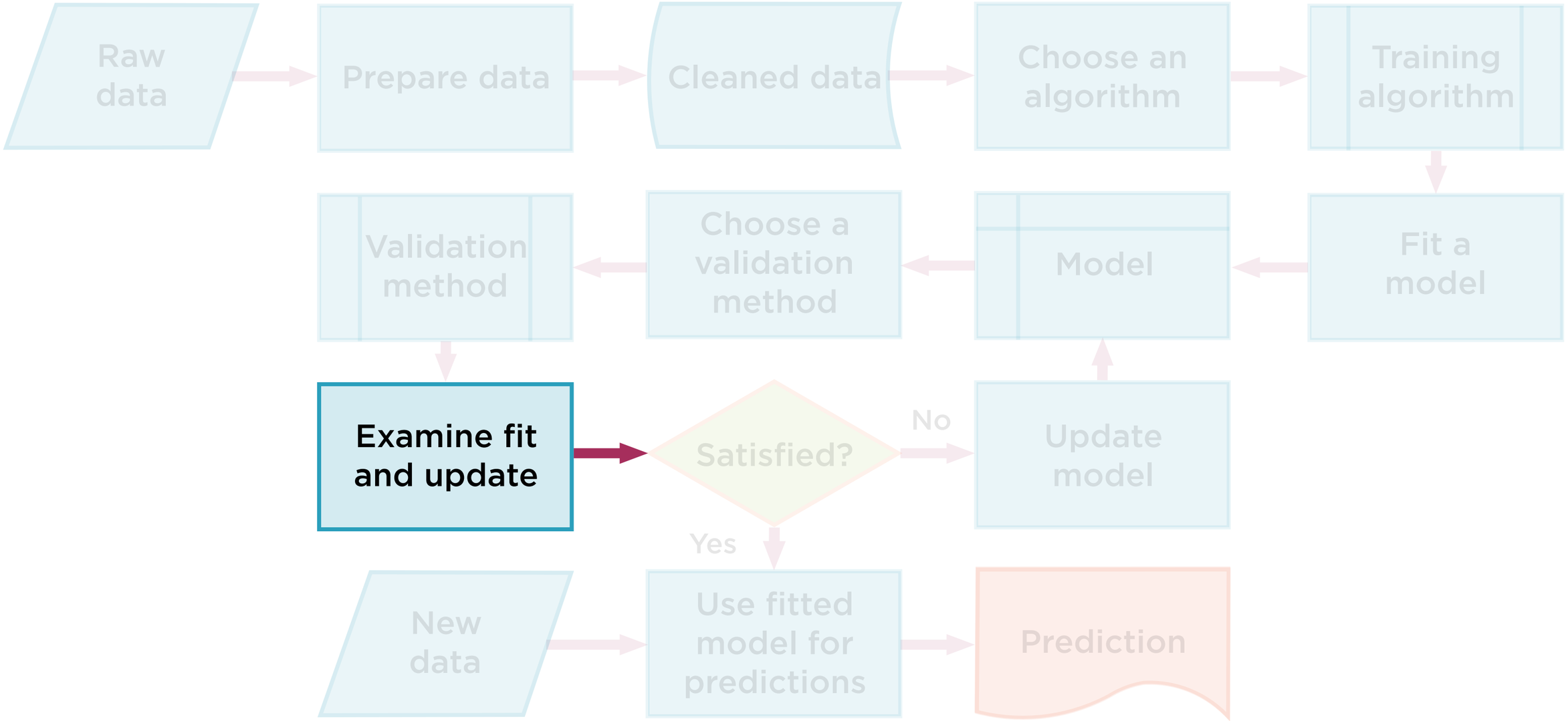
Training to Find Model Parameters



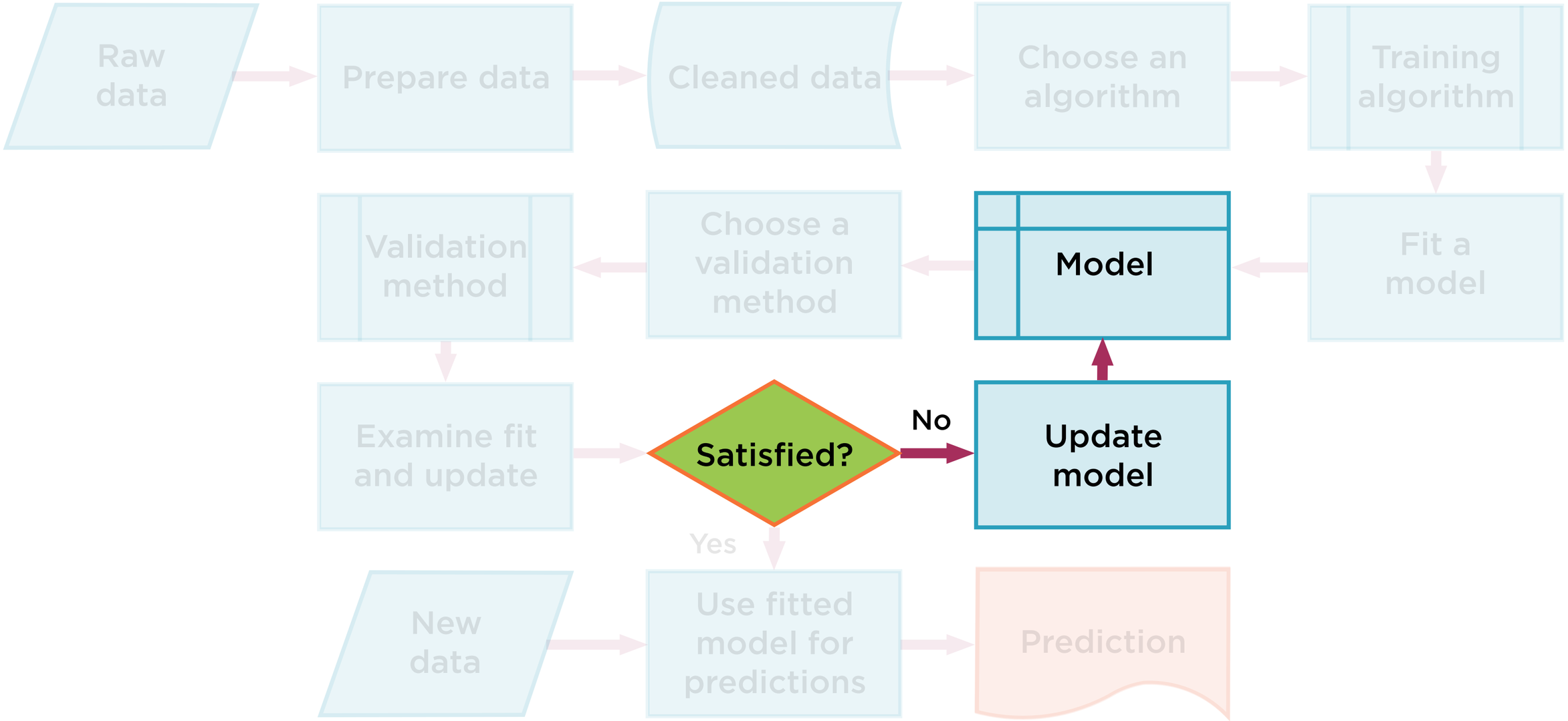
Evaluate the Model



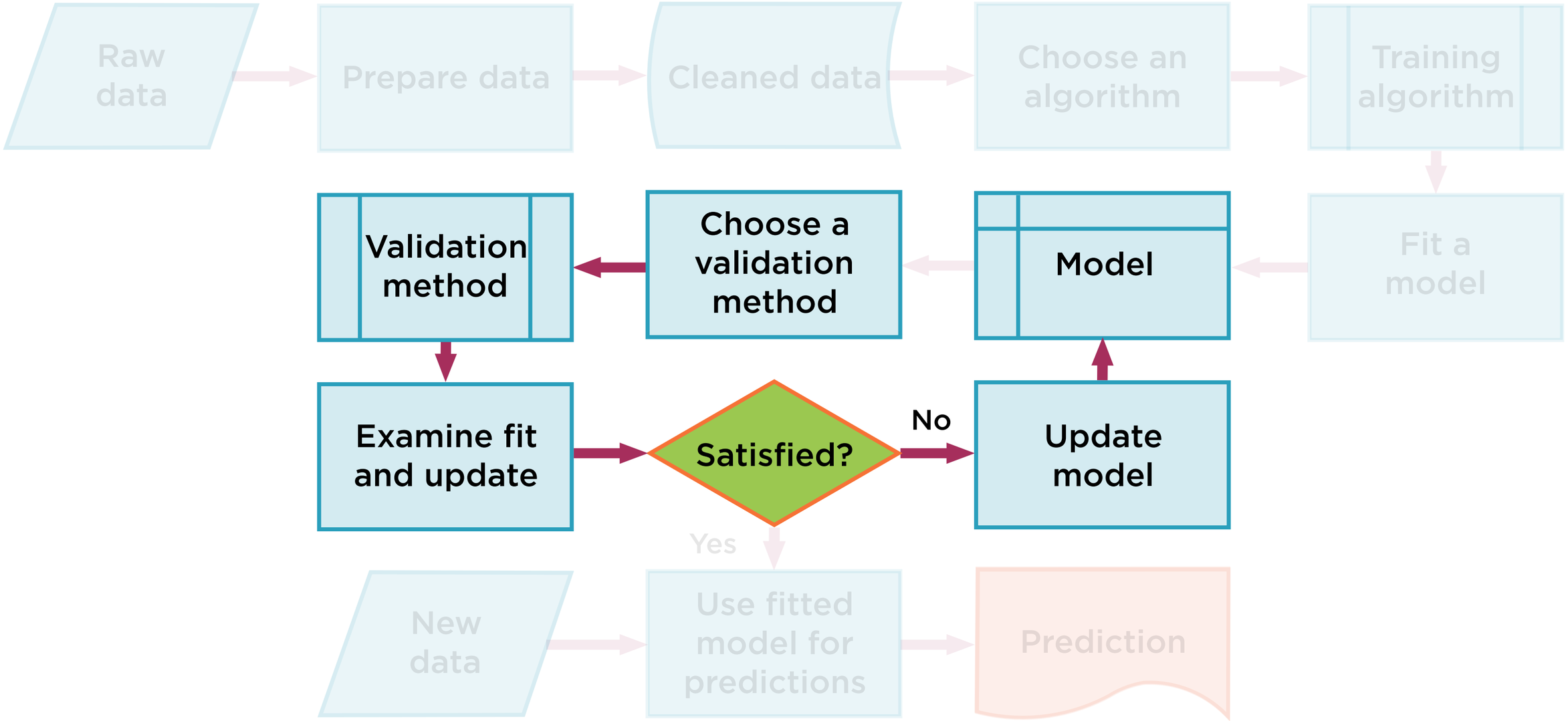
Score the Model



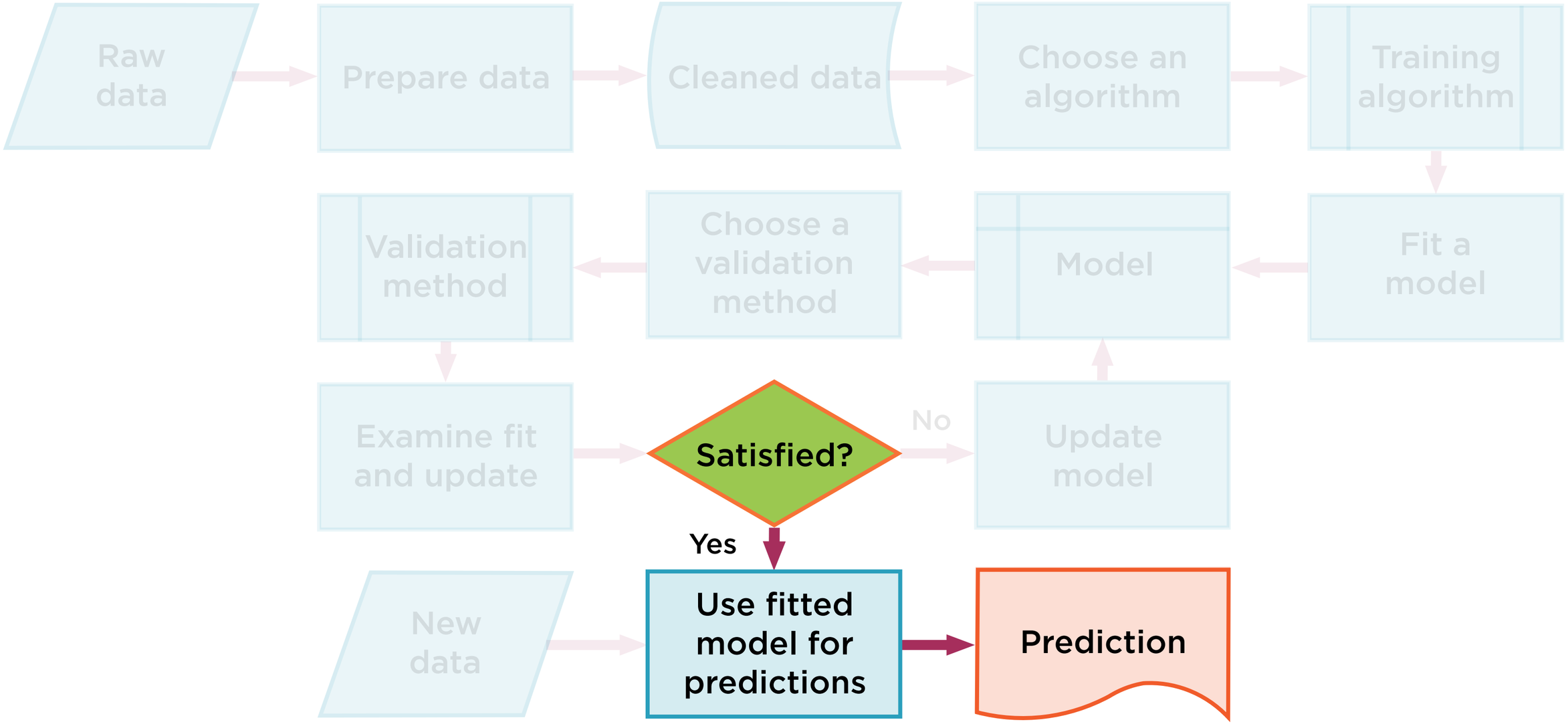
Different Algorithm, More Data, More Training?



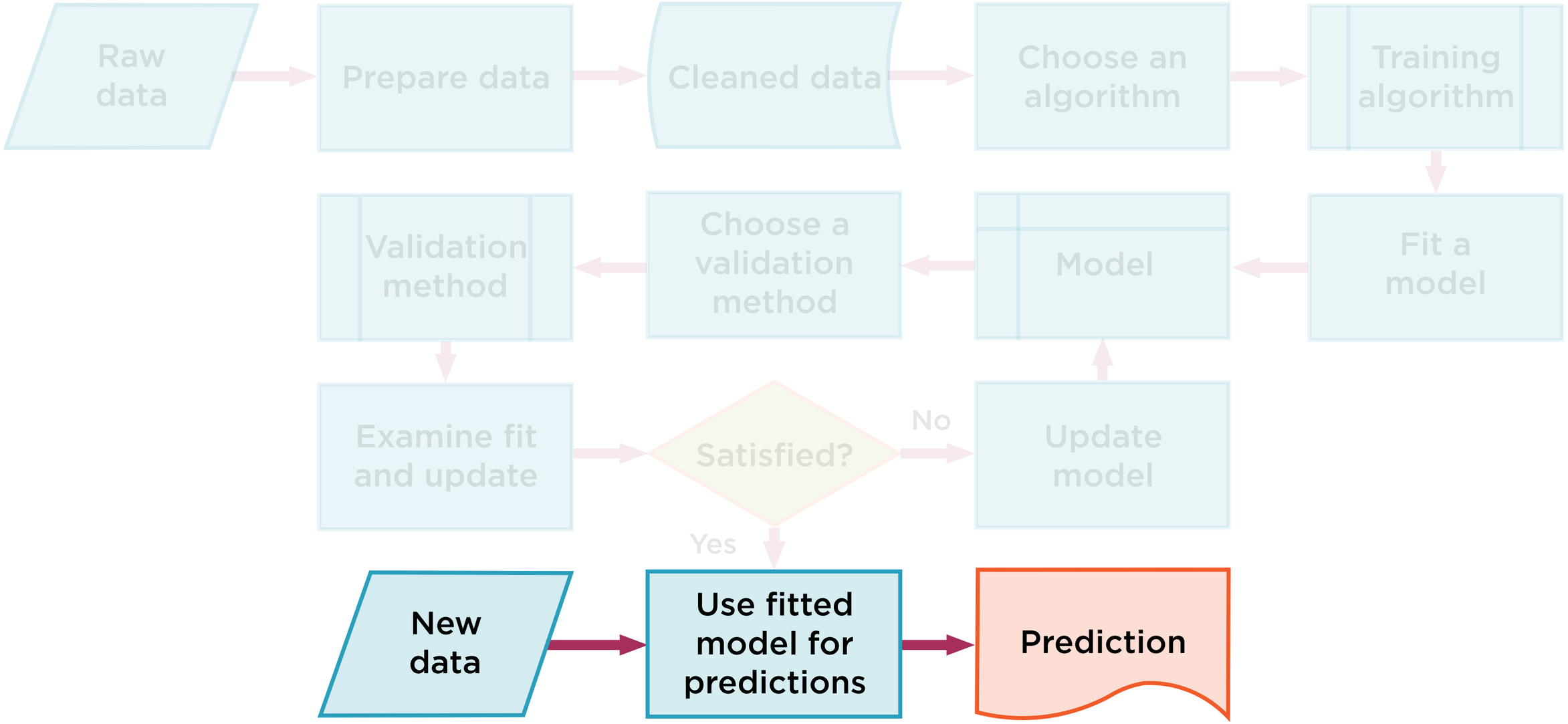
Iterate Till Model Finalized



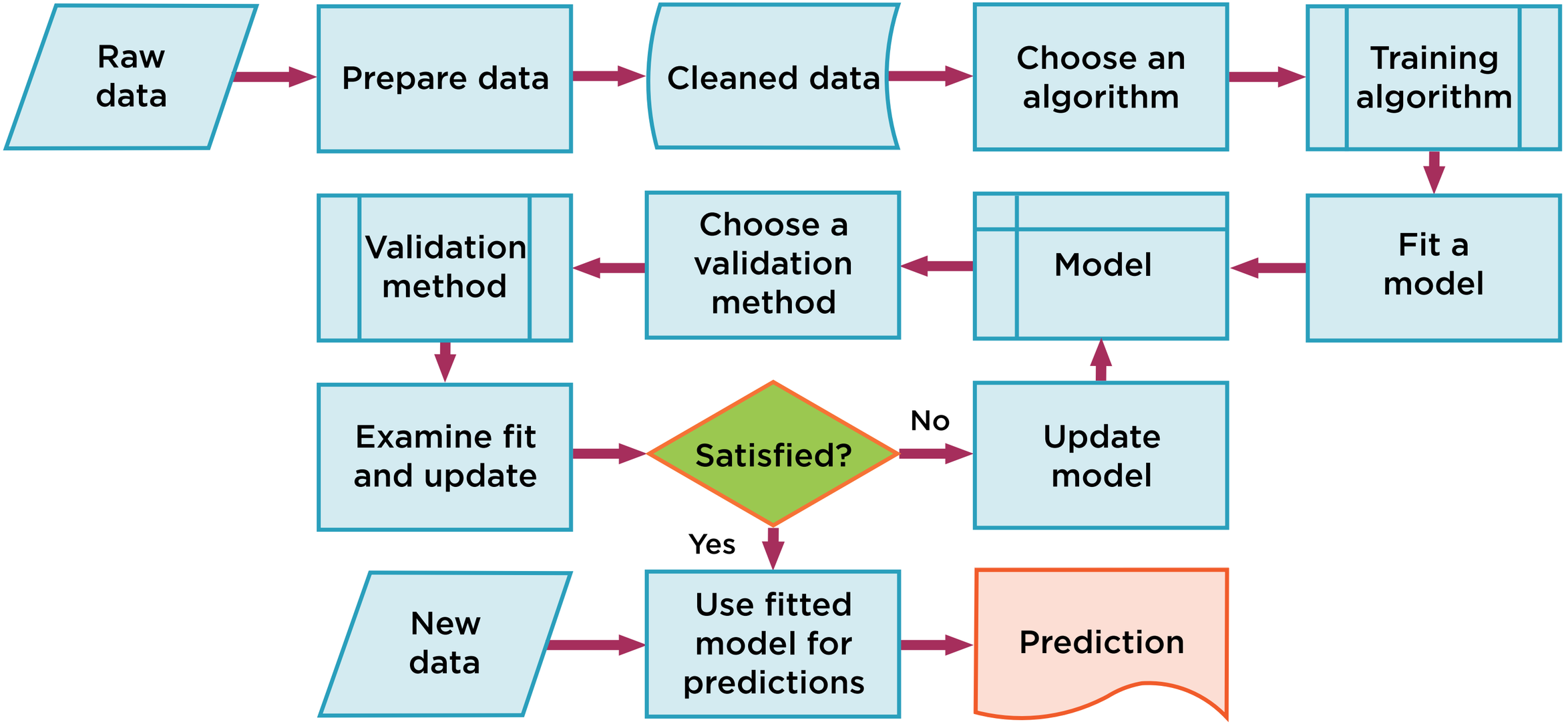
Model Used for Predictions



Retrained Using New Data



Basic Machine Learning Workflow



Summary

Hypothesis testing to evaluate proposed explanations for phenomenon

Understanding the T-test to test for differences between categories

Using ANOVA to test differences across multiple groups

Choosing an algorithm based on prediction target

Understanding the steps involved in building a model