Applied Machine Learning: what is ML?

NAWWAF KHARMA

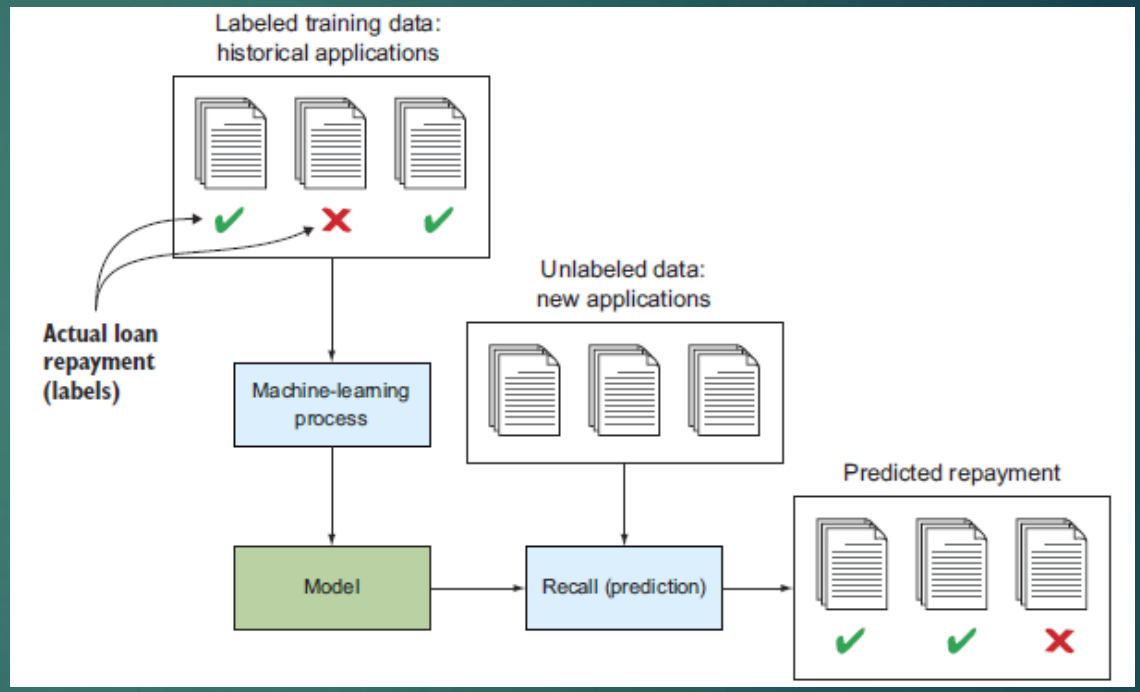
What is Machine Learning

- 1.2.2 The Machine Learning Approach

Machine learning vs. manual expertise vs. hard-coded rules

- Automated
- Data-driven
- Continuous improvement / adaptation

Train on historical data to build a model, which is then applied to new data



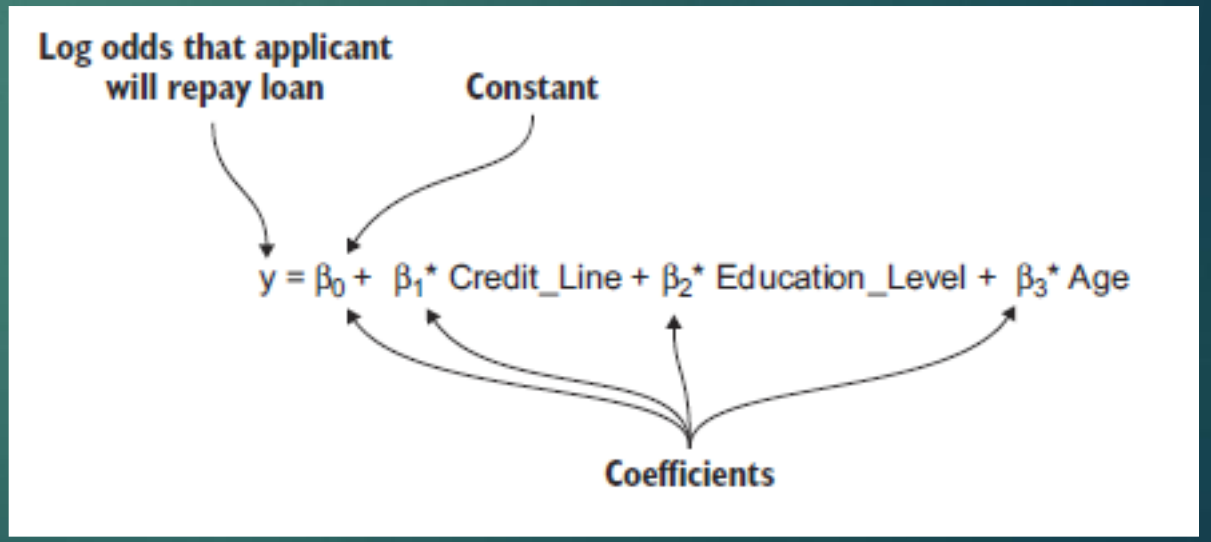
What is Machine Learning

- ML produces a model
- .. Sophisticated or simple (parametric)
- In this example, you could use logistic regression to model the loan-approval process. In logistic regression, the logarithm of the odds (the log odds) that each loan is repaid is modeled as a linear function of the input features.
- Side-note:

Odds(A) = P(A)/P(~A)

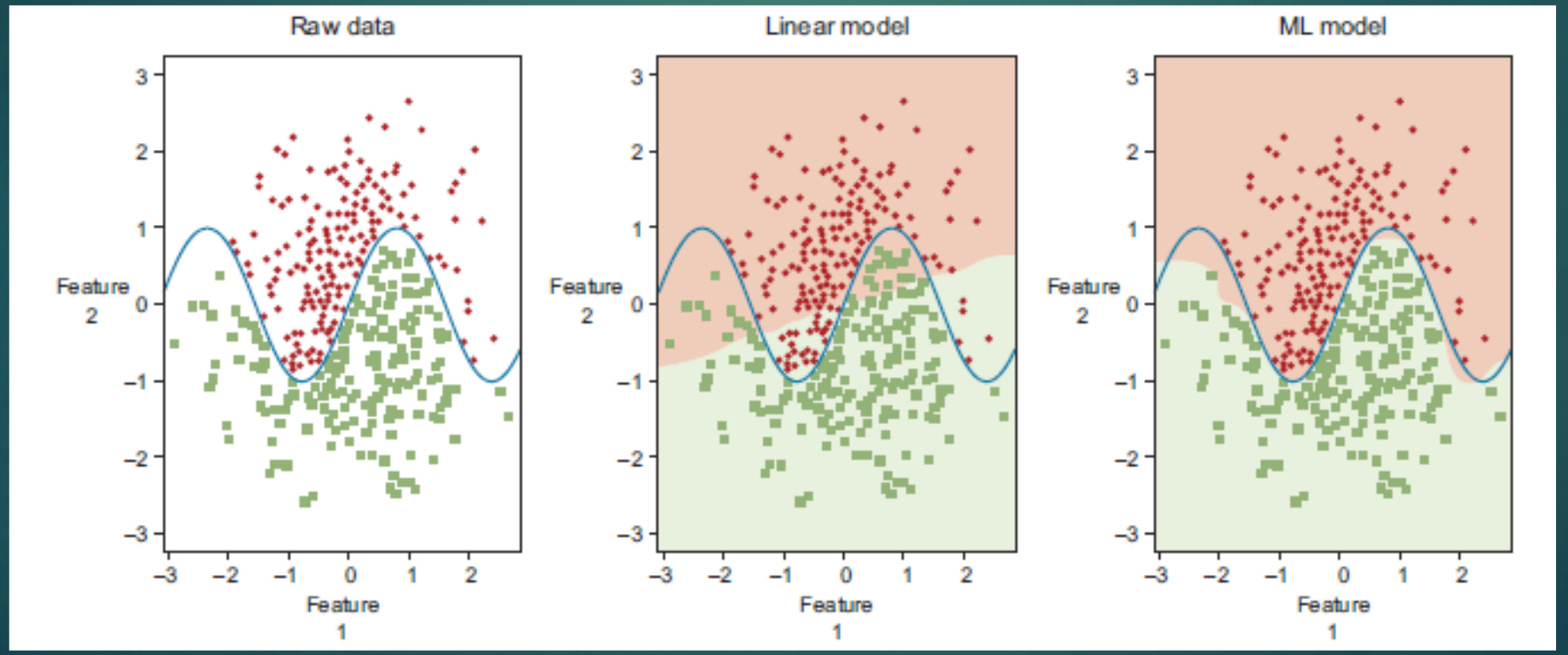
Log(Odds): Log(1/1)= 0;
Log(near-infinity)= large number

So, large numbers is near certainty, while 0 indicates a perfectly even random event.



What is Machine Learning

- The optimal value of each coefficient is learned from 1000 training data points
- more sophisticated data boundary ➔ more parameters ➔ more training data



Cont.

- parametric models work well when you have prior understanding of the relationship between your inputs and the response you're trying to predict.
- What you need are more flexible models that can automatically discover complex trends and structure in data, without being told what the patterns look like: nonparametric machine-learning algorithms come to the rescue.
- attain such high levels of accuracy on complicated, high-dimensional, real-world datasets, nonparametric ML models are the approach of choice for many data-driven problems.
- Examples of such models include: k-nearest neighbours, support vector machines, decision trees and ensemble methods

Cont.

- 1.2.3 Five Advantages to Machine Learning

Accurate—ML uses data to discover the optimal decision-making engine

Automated—the ML model can learn new patterns automatically

Fast—can generate answers in a matter of milliseconds

Customizable—Many data-driven problems can be addressed

Scalable—As your business grows, ML easily scales to handle increased data

Challenges:

(1) Requires data (sometime lots of data) in usable form

(2) Problem needs to be formulated in a way that allows the ML to produce results that are actionable and measurable

Cont.

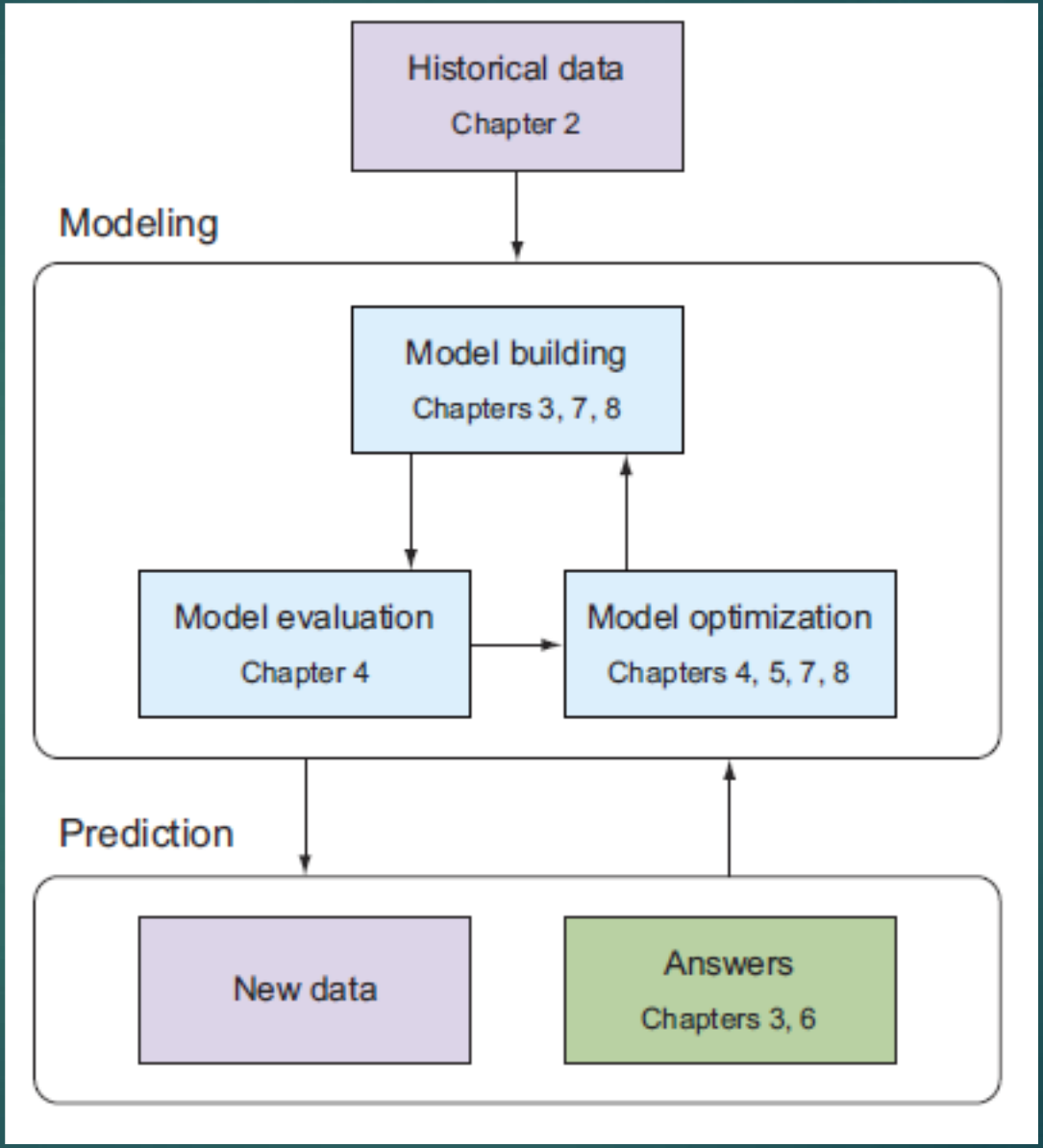
- Example: A more difficult example might go along these lines: find the optimum media mix and combination of advertising units to increase brand awareness for a new product line.
- Simply formulating the problem requires constructing a way of measuring brand awareness, an understanding of the alternative media options under consideration, and data that reflects pertinent experience with the alternatives and associated outcomes.

(3) Feature engineering: the process of transforming inputs into predictive features

(4) The problem of overfitting: building a model that fits the training data too well but fails to generalize to unseen (new) data.

Cont.

- 1.3 Following the ML Workflow: from data to deployment



Cont.

- 1.3.1 Data collection and preparation

Think of the tabular format as a spreadsheet in which data is distributed in rows and columns, with each row corresponding to an instance or example of interest, and each column representing a measurement on this instance. (this is a heterogonous data set: has integers, categories etc.)

Features in columns

Person	Name	Age	Income	Marital status
1	Jane Doe	24	81,200	Single
2	John Smith	41	121,000	Married

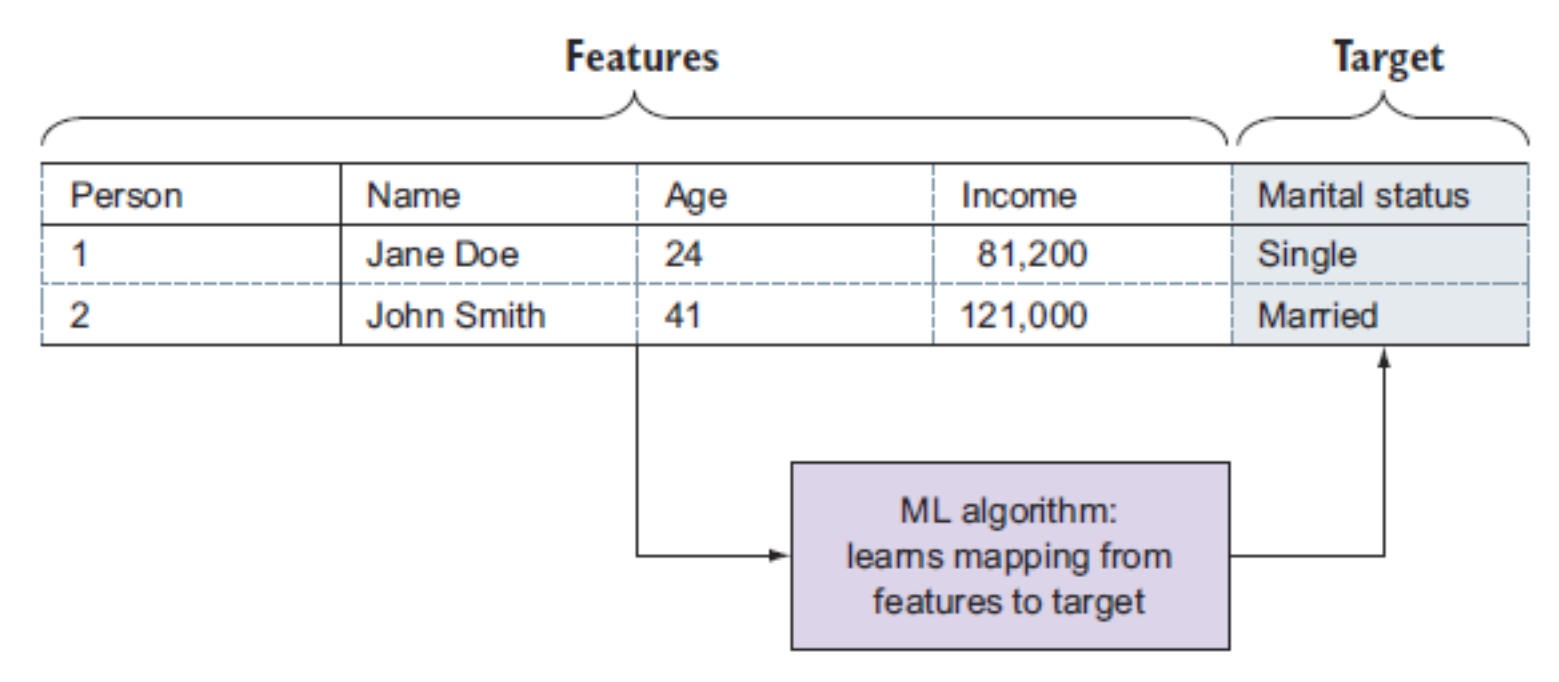
Examples in rows

> real-world data is messy: has missing values + errors

Cont.

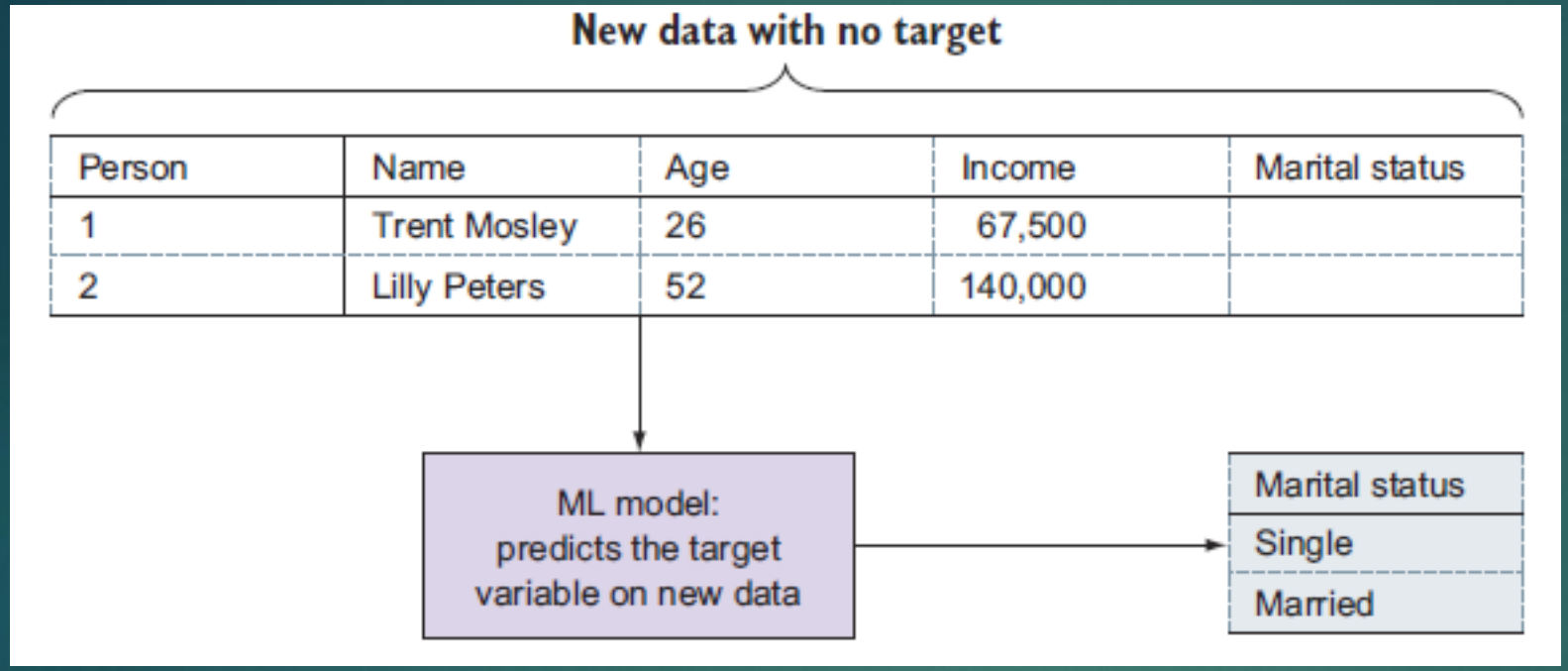
- 1.3.2 Learning a model from data

Predict the target: the machine learning process



> Some algorithms are insensitive to useless features, some features can be processed to become useful and some cannot and could ruin the model

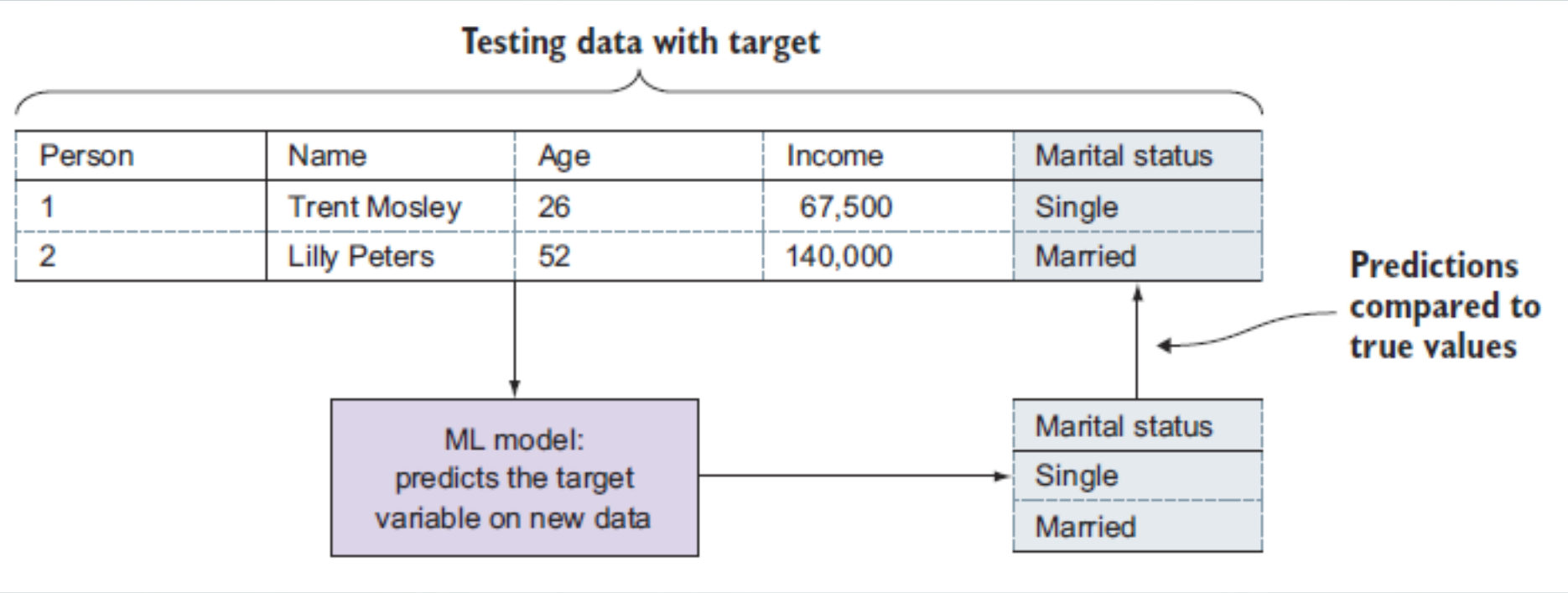
Cont. (learning a model cont.)




```
data = load_data("data/people.csv")
model = build_model(data, target="Marital status")
new_data = load_data("data/new_people.csv")
predictions = model.predict(new_data)
```

Cont.

- 1.3.3 Evaluating model performance




```
data = load_data(...)
training_data, testing_data = split_data(data)
model = build_model(training_data, target="Marital status")
true_values = testing_data.extract_column("Marital status")
predictions = model.predict(testing_data)
accuracy = compare_predictions(predictions, true_values)
```

Cont.

- 1.3.4 Optimizing model performance

(1) Tuning the model **parameters—ML** algorithms are configured with parameters specific to the underlying algorithm, and the optimal value of these parameters often depends on the type and structure of the data.

(2) Selecting a subset of **features—Many** ML problems include a large number of features, and the noise from those features can sometimes make it hard for the algorithm to find the real signal in the data

(3) Preprocessing the **data-** Most real-world datasets, however, aren't in such a clean state, and you'll have to perform cleaning and processing, a process widely referred to as data wrangling

Cont.

- **1.4 Boosting model performance with advanced techniques:**

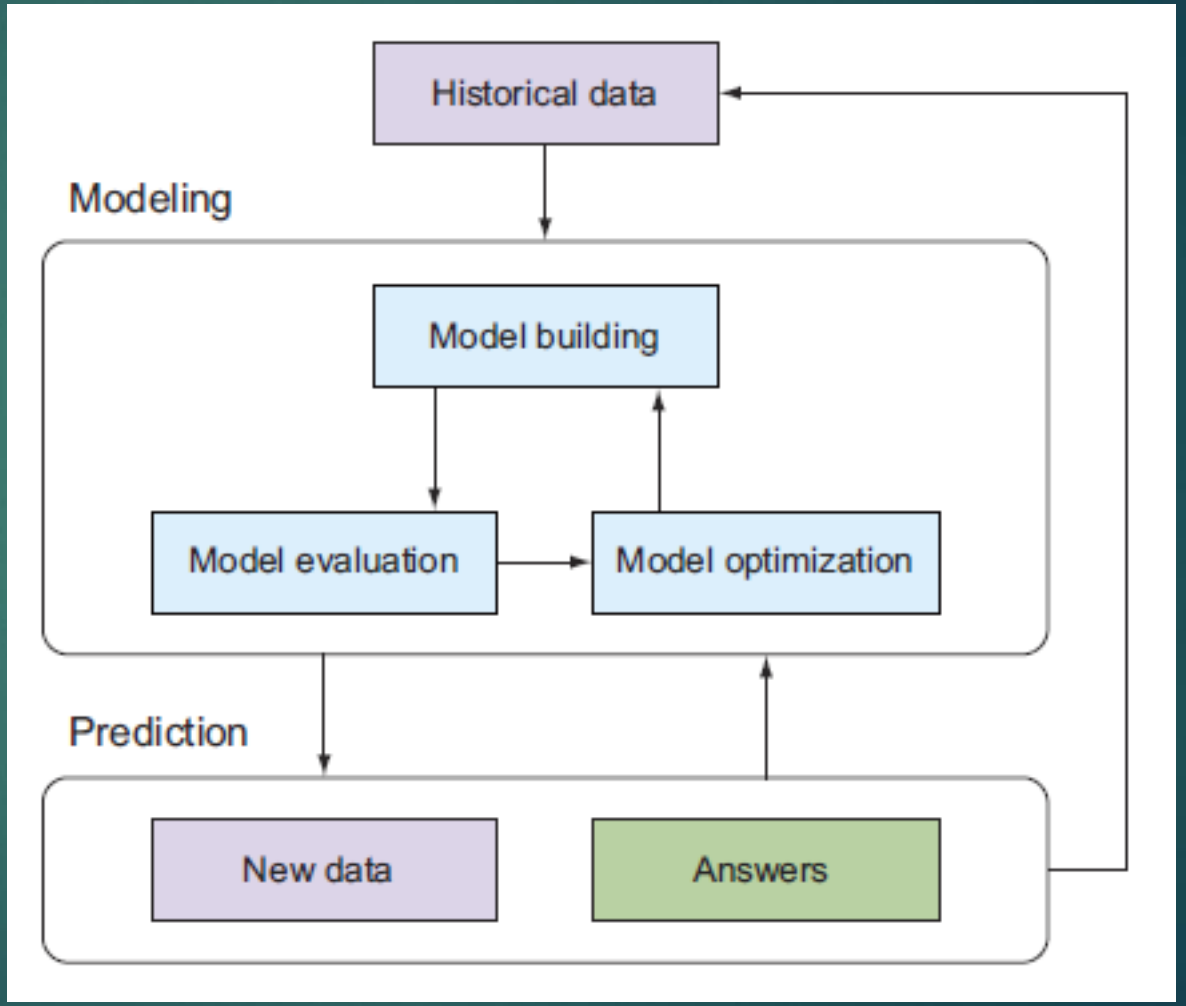
> 1.4.1 Data preprocessing and feature engineering

Dates and times: observations of repetitive activity

Location: extract additional information that's useful

Digital Media: Edges, shapes, and color spectra

> 1.4.2 Improving models continually with online methods



Summary

- Machine-learning algorithms are distinguished from rule-based systems in that they create their own models based on data.
- Machine learning is often more *accurate, automated, fast, customizable, and scalable* than manually constructed rule-based systems.
- Machine-learning challenges include identifying and formulating problems to which ML can be applied, acquiring and transforming data to make it usable, finding the right algorithms for the problem, feature engineering, and overfitting
- The basic machine-learning workflow consists of data preparation, model building, model evaluation, optimization, and predictions on new data.
- Online learning models continually relearn by using the results of their predictions to update themselves.