# What is machine learning?

UNDERSTANDING MACHINE LEARNING

**Lis Sulmont**
Curriculum Manager, DataCamp

datacamp

**AlphaZero beat humans at Chess and StarCraft, now it's working with quantum computers**

AlphaZero beat humans at Chess and StarCraft, now it's working with quantum computers

Machine Learning Gives Rise to Better Lung Disease Models from Stem Cells

AI Blood Test Predicts Neurodegenerative Disease Progression, Severity

AlphaZero beat humans at Chess and StarCraft, now it's working with quantum computers

Machine Learning Gives Rise to Better Lung Disease Models from Stem Cells

AI Blood Test Predicts Neurodegenerative Disease Progression, Severity

Amazon is injecting Alexa with more artificial intelligence than ever

How Google Maps uses machine learning to predict bus traffic delays in real time

# This A.I. Bot Writes Such Convincing Ads, Chase Just 'Hired' It to Write Marketing Copy

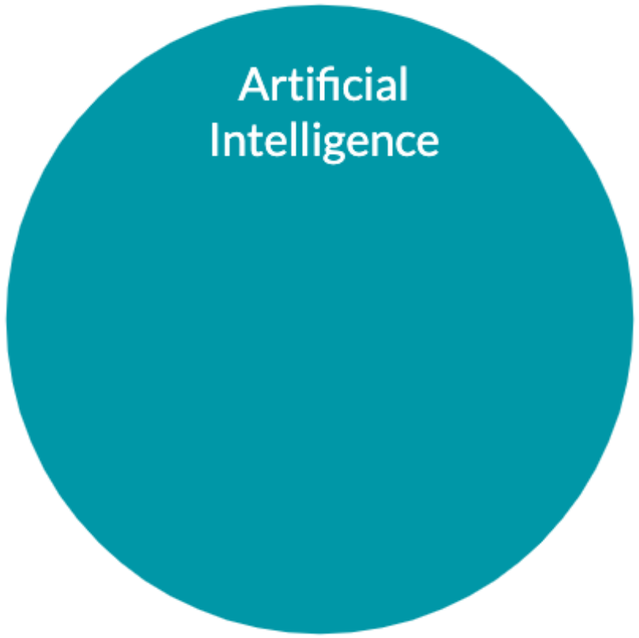
During a trial run, JPMorgan Chase found ads written by the machine learning platform got far more clicks than ads written by humans.

# A.I. is transforming the job interview—and everything after


Waymo's Self-Driving Technology Gets Smarter, Recognizes Billions of Objects Thanks To Content Search
New machine learning algorithm produces "near-perfect" fake human faces
The AI-Art Gold Rush Is Here
An artificial-intelligence "artist" got a solo show at a Chelsea gallery. Will it reinvent art, or destroy it?

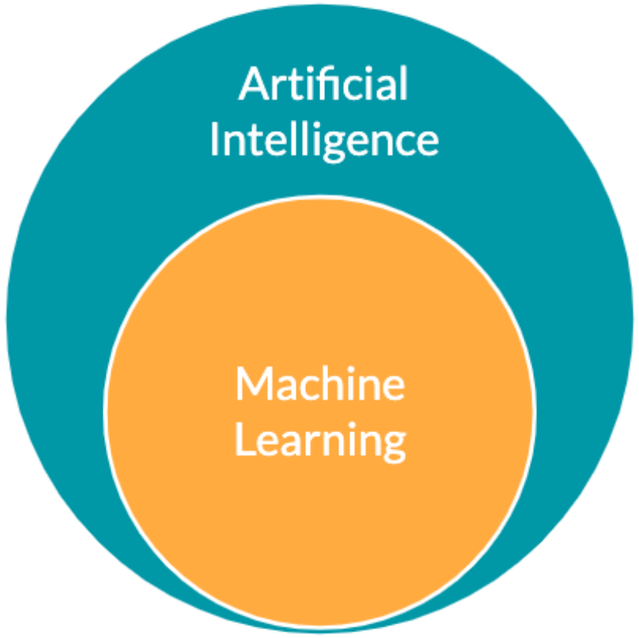

White House reportedly aims to double AI research budget to $2B
Finland seeks to teach 1% of all Europeans basics on AI

# Artificial intelligence (AI)




A huge set of tools for making computers behave intelligently

# Artificial intelligence (AI)

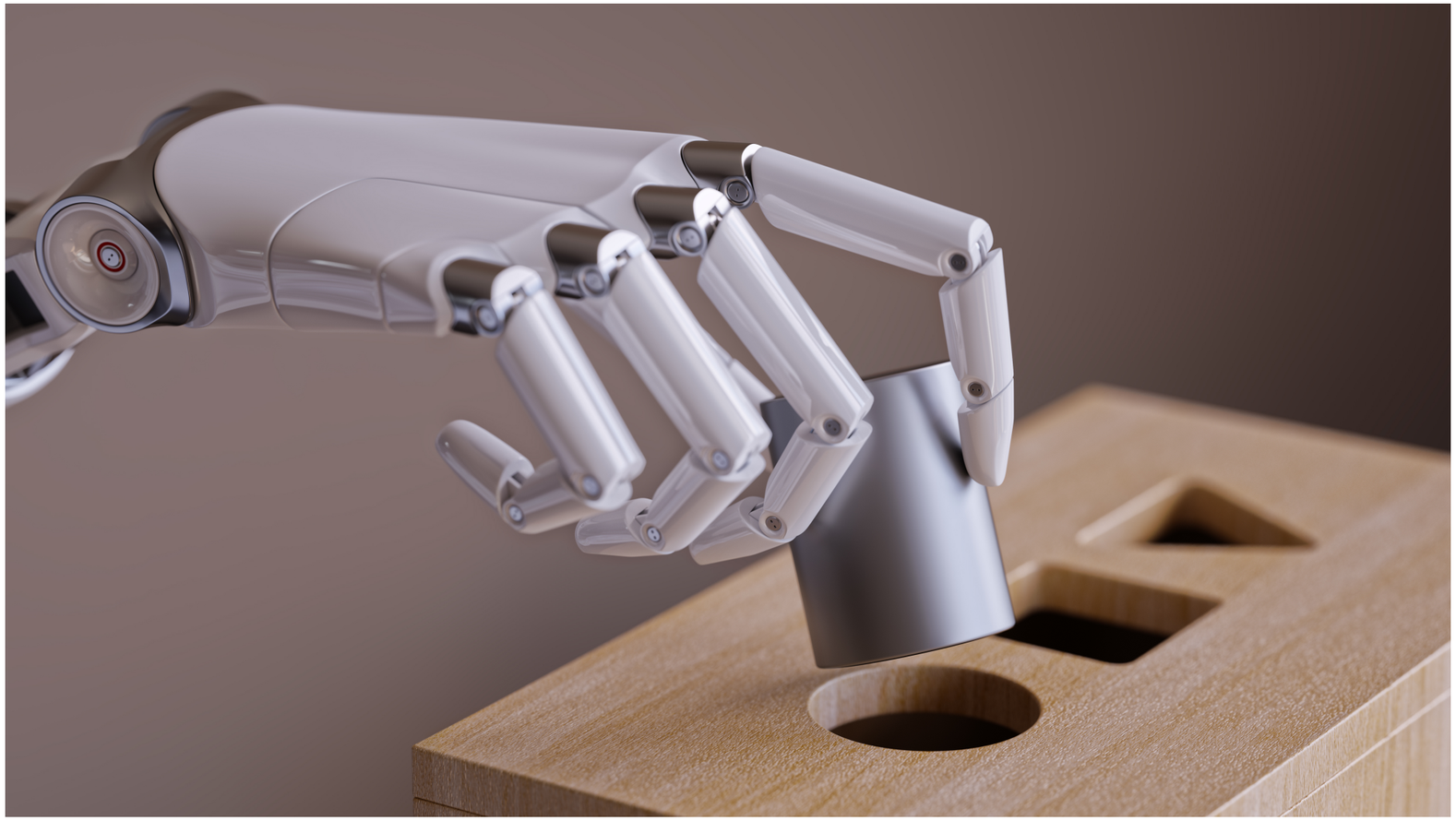



A huge set of tools for making computers behave intelligently

Machine learning is the most prevalent subset of AI

# Defining machine learning:

A set of tools for making inferences and predictions from data



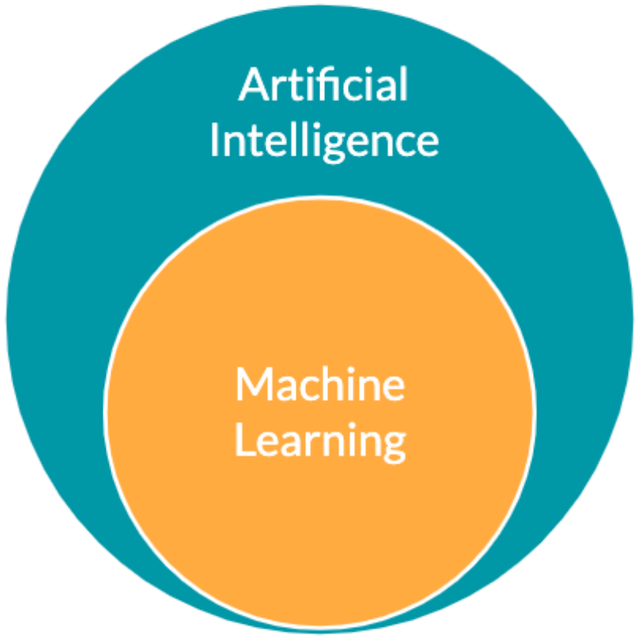
# Defining machine learning: what can it do?

- **Predict** future events
  - *Will it rain tomorrow?*
    - Yes (75% probability)
- **Infer** the causes of events and behaviors
  - *Why does it rain?*
    - Time of the year, humidity levels, temperature, location, etc
- **Infer** patterns
  - *What are the different types of weather conditions?*
    - Rain, sunny, overcast, fog, etc

# Defining machine learning: how does it work?

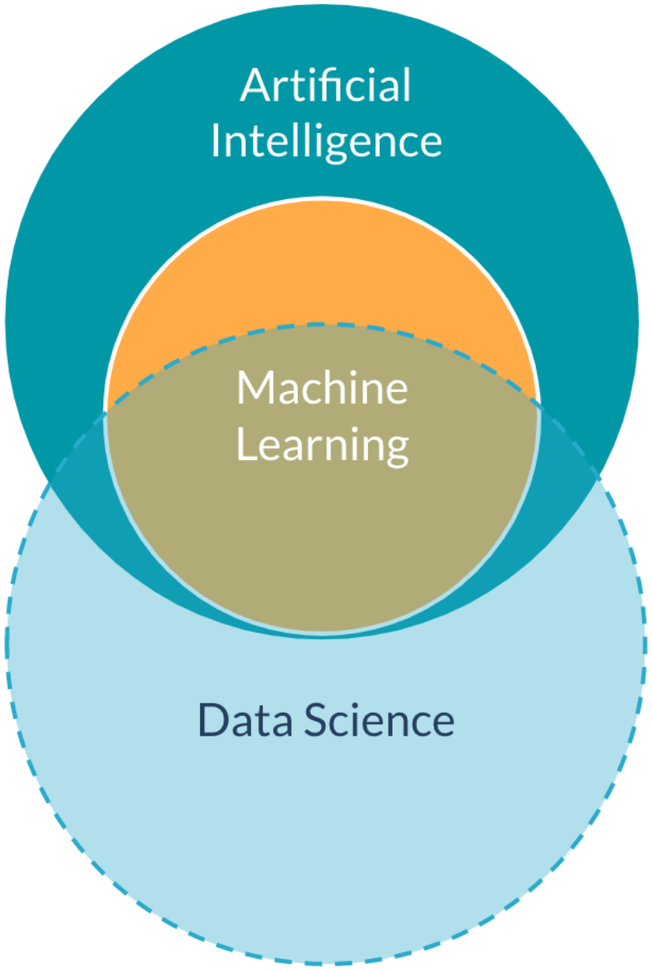
- Interdisciplinary mix of statistics and computer science
- Ability to learn without being explicitly programmed
- Learn patterns from existing data and applies it to new data
- Relies on high-quality data
- ... more to come throughout the course!

# Data science




Data science is about discovering and communicating insights from data

# Data science

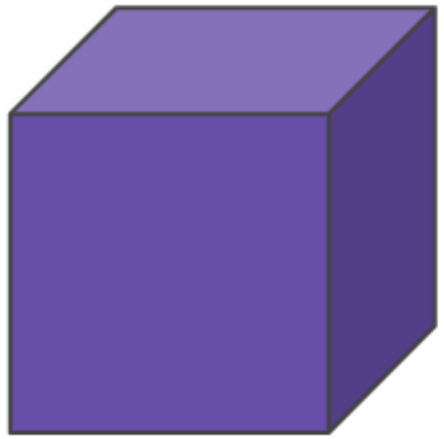



Data science is about making discoveries and creating insights from data

Machine learning is often an important tool for data science work

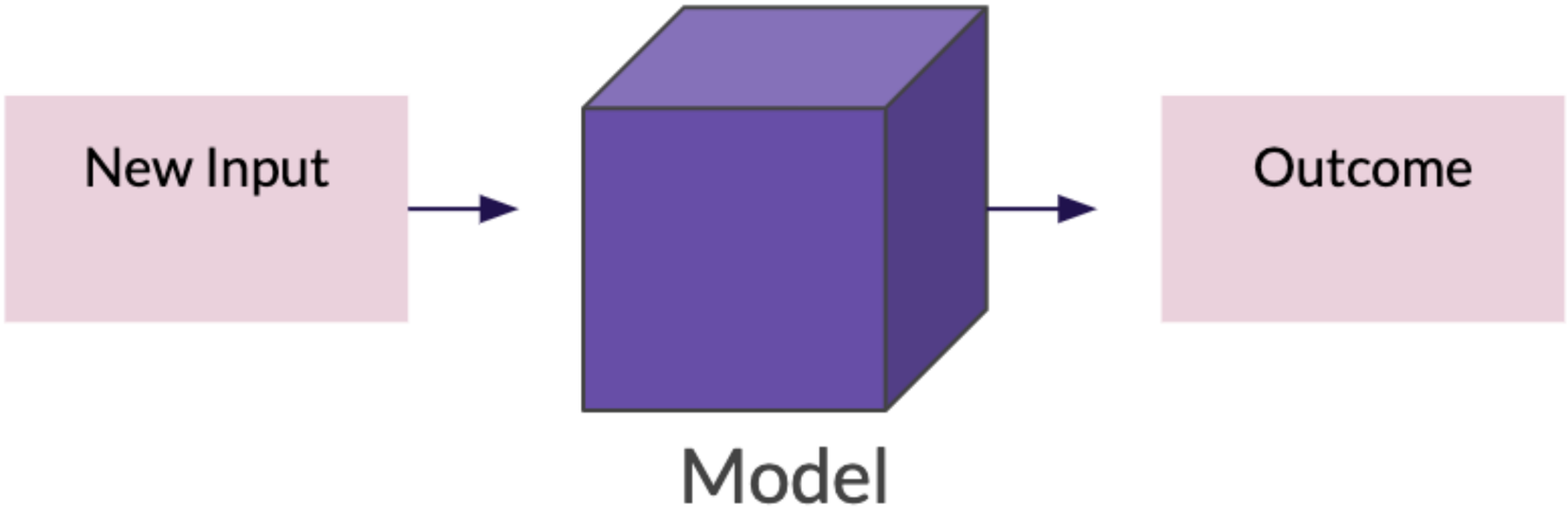
# Machine learning model

A statistical representation of a real-world process based on data



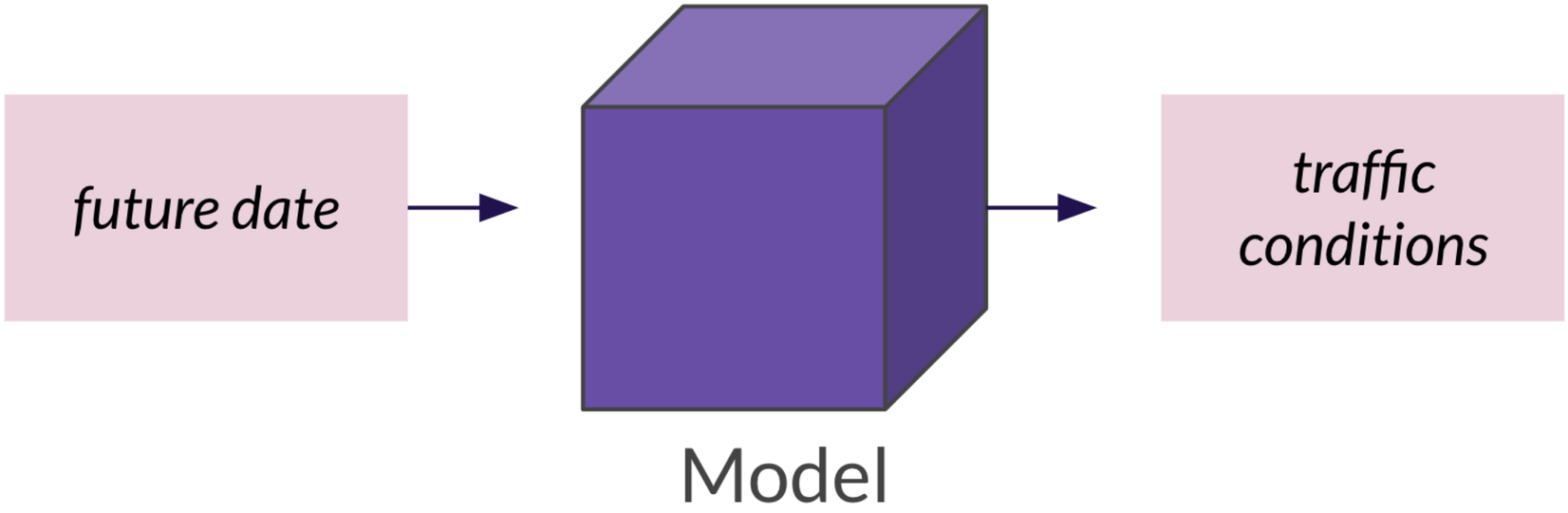
# Machine learning model

A statistical representation of a real-world process based on data



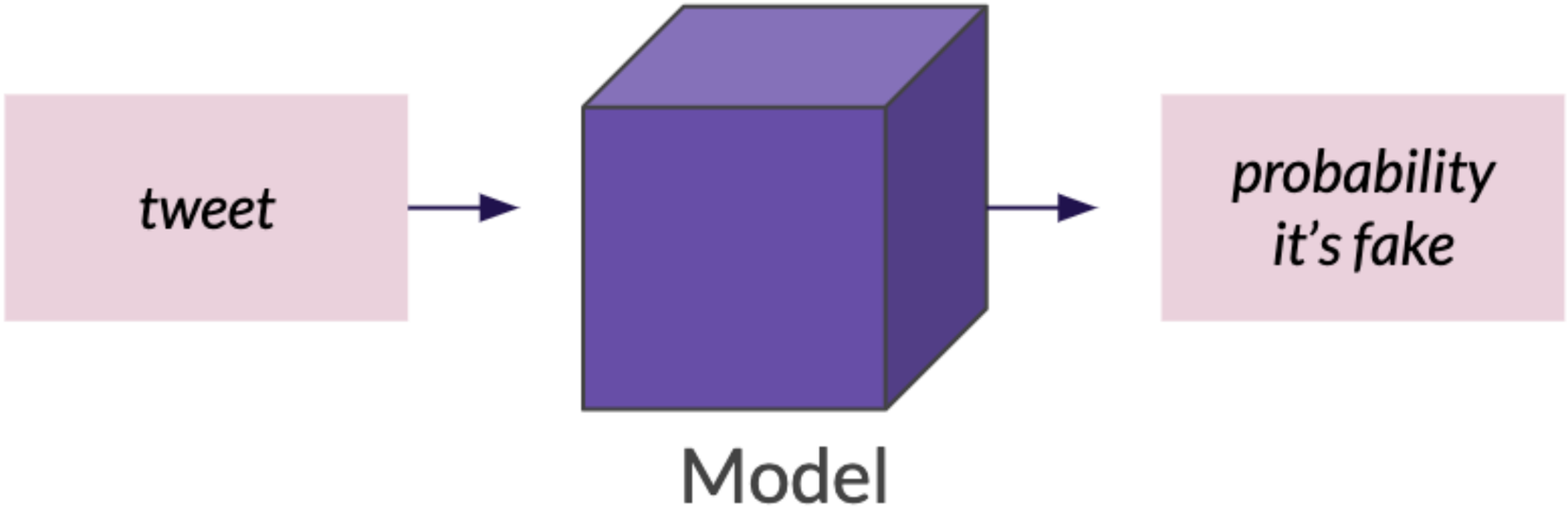
# Machine learning model

A statistical representation of a real-world process based on data



# Machine learning model

A statistical representation of a real-world process based on data



# Let's practice!

UNDERSTANDING MACHINE LEARNING

# Machine learning concepts

## UNDERSTANDING MACHINE LEARNING

**Lis Sulmont**
Curriculum Manager, DataCamp

datacamp

# Three types of machine learning

1) Reinforcement learning

2) Supervised learning

3) Unsupervised learning

# Training data

- **Training data:** existing data to learn from
- **Training a model:** when a model is being built from training data
  - Can take nanoseconds to weeks

# Supervised learning training data

| Age | Sex | Cholesterol | Cigarettes per day | Family history of heart disease | Chest pain type | Blood sugar | Heart disease |
|---|---|---|---|---|---|---|---|
| 55 | M | 221 | 5 | True | typical angina | 118 | True |
| 50 | F | 196 | 0 | False | non-anginal pain | 98 | False |
| 53 | F | 215 | 0 | True | asymptomatic | 110 | True |
| 62 | M | 245 | 3 | False | typical angina | 126 | True |
| 48 | M | 190 | 0 | True | non-anginal pain | 99 | False |
| 70 | M | 201 | 0 | True | typical angina | 105 | False |

# Supervised learning training data

Target Variable

| Age | Sex | Cholesterol | Cigarettes per day | Family history of heart disease | Chest pain type | Blood sugar | Heart disease |
|---|---|---|---|---|---|---|---|
| 55 | M | 221 | 5 | True | typical angina | 118 | True |
| 50 | F | 196 | 0 | False | non-anginal pain | 98 | False |
| 53 | F | 215 | 0 | True | asymptomatic | 110 | True |
| 62 | M | 245 | 3 | False | typical angina | 126 | True |
| 48 | M | 190 | 0 | True | non-anginal pain | 99 | False |
| 70 | M | 201 | 0 | True | typical angina | 105 | False |

# Supervised learning training data

Target Variable

| Age | Sex | Cholesterol | Cigarettes per day | Family history of heart disease | Chest pain type | Blood sugar | Heart disease |
|---|---|---|---|---|---|---|---|
| 55 | M | 221 | 5 | True | typical angina | 118 | True |
| 50 | F | 196 | 0 | False | non-anginal pain | 98 | False |
| 53 | F | 215 | 0 | True | asymptomatic | 110 | True |
| 62 | M | 245 | 3 | False | typical angina | 126 | True |
| 48 | M | 190 | 0 | True | non-anginal pain | 99 | False |
| 70 | M | 201 | 0 | True | typical angina | 105 | False |

Labels

# Supervised learning training data

Observations *or* Examples

| Age | Sex | Cholesterol | Cigarettes per day | Family history of heart disease | Chest pain type | Blood sugar | Heart disease |
|---|---|---|---|---|---|---|---|
| 55 | M | 221 | 5 | True | typical angina | 118 | True |
| 50 | F | 196 | 0 | False | non-anginal pain | 98 | False |
| 53 | F | 215 | 0 | True | asymptomatic | 110 | True |
| 62 | M | 245 | 3 | False | typical angina | 126 | True |
| 48 | M | 190 | 0 | True | non-anginal pain | 99 | False |
| 70 | M | 201 | 0 | True | typical angina | 105 | False |

# Supervised learning training data

Features

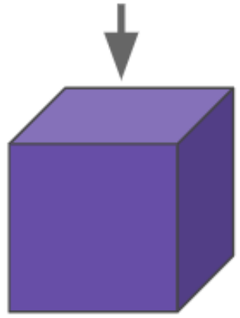
| Age | Sex | Cholesterol | Cigarettes per day | Family history of heart disease | Chest pain type | Blood sugar | Heart disease |
|---|---|---|---|---|---|---|---|
| 55 | M | 221 | 5 | True | typical angina | 118 | True |
| 50 | F | 196 | 0 | False | non-anginal pain | 98 | False |
| 53 | F | 215 | 0 | True | asymptomatic | 110 | True |
| 62 | M | 245 | 3 | False | typical angina | 126 | True |
| 48 | M | 190 | 0 | True | non-anginal pain | 99 | False |
| 70 | M | 201 | 0 | True | typical angina | 105 | False |

# After training (supervised learning)

| Age | Sex | Cholesterol | Cigarettes per day | Family history of heart disease | Chest pain type | Blood sugar | Heart disease |
|---|---|---|---|---|---|---|---|
| 65 | F | 208 | 2 | False | typical angina | 105 | ??? |

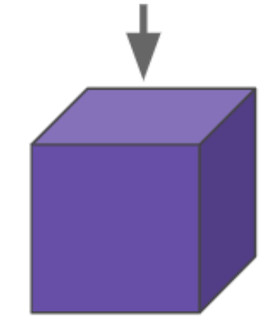
# After training (supervised learning)

| Age | Sex | Cholesterol | Cigarettes per day | Family history of heart disease | Chest pain type | Blood sugar | Heart disease |
|---|---|---|---|---|---|---|---|
| 65 | F | 208 | 2 | False | typical angina | 105 | ??? |



# After training (supervised learning)

| Age | Sex | Cholesterol | Cigarettes per day | Family history of heart disease | Chest pain type | Blood sugar | Heart disease |
|---|---|---|---|---|---|---|---|
| 65 | F | 208 | 2 | False | typical angina | 105 | ??? |



| Heart disease |
|---|
| False |

# Supervised vs unsupervised learning

- **Supervised learning**
  - Training data is "labeled"
- **Unsupervised learning**
  - Training data only has features
  - Useful for:
    - Anomaly detection
    - Clustering, e.g., *dividing data into groups*

Target Variable

| Blood sugar | Heart disease |
|---|---|
| 118 | True |
| 98 | False |
| 110 | True |
| 126 | True |
| 99 | False |
| 105 | False |

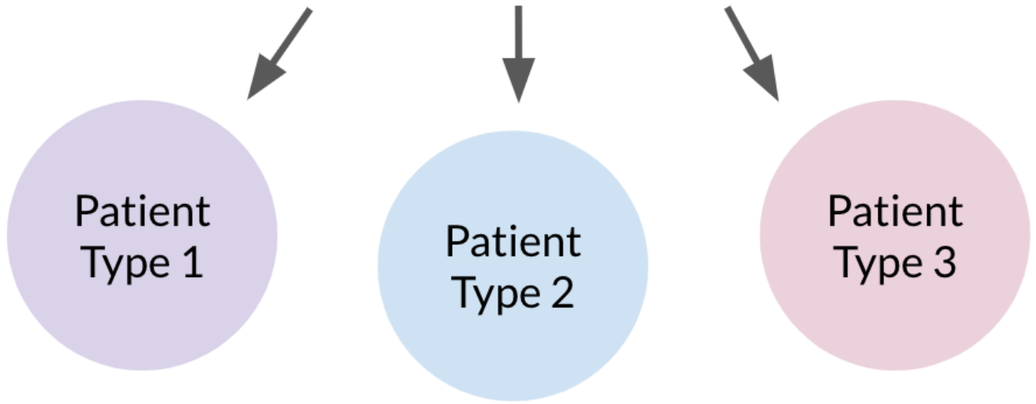
Labels

# Unsupervised learning training data

| Age | Sex | Cholesterol | Cigarettes per day | Family history of heart disease | Chest pain type | Blood sugar | Heart disease |
|---|---|---|---|---|---|---|---|
| 55 | M | 221 | 5 | True | typical angina | 118 | True |
| 53 | F | 199 | 0 | True | non-anginal pain | 98 | True |
| 53 | F | 215 | 0 | True | asymptomatic | 110 | True |
| 62 | M | 245 | 3 | False | typical angina | 126 | True |
| ... | ... | ... | ... | ... | ... | ... | ... |

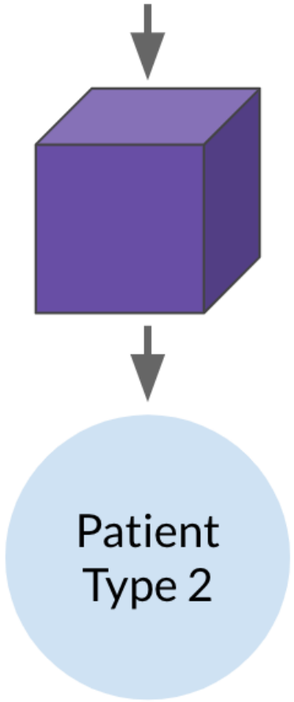
# Unsupervised learning training data

| Age | Sex | Cholesterol | Cigarettes per day | Family history of heart disease | Chest pain type | Blood sugar | Heart disease |
|---|---|---|---|---|---|---|---|
| 55 | M | 221 | 5 | True | typical angina | 118 | True |
| 53 | F | 199 | 0 | True | non-anginal pain | 98 | True |
| 53 | F | 215 | 0 | True | asymptomatic | 110 | True |
| 62 | M | 245 | 3 | False | typical angina | 126 | True |
| ... | ... | ... | ... | ... | ... | ... | ... |



# After training (unsupervised learning)

| Age | Sex | Cholesterol | Cigarettes per day | Family history of heart disease | Chest pain type | Blood sugar | Heart disease |
|---|---|---|---|---|---|---|---|
| 65 | F | 208 | 2 | False | typical angina | 105 | True |



# Unsupervised Learning

- In reality, data doesn't always come with labels
  - Requires manual labor to label
  - Labels are unknown
- No labels: model is unsupervised and finds its own patterns

# Let's practice!

UNDERSTANDING MACHINE LEARNING

# Machine learning workflow

UNDERSTANDING MACHINE LEARNING

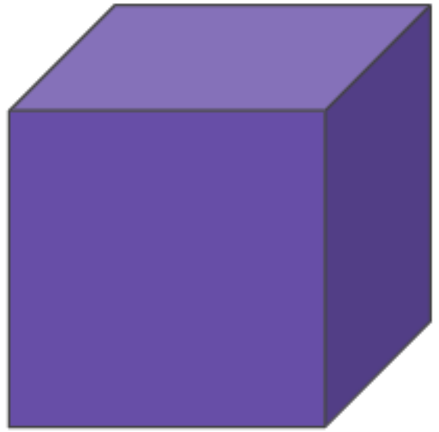
**Lis Sulmont**
Curriculum Manager, DataCamp

datacamp

# Machine learning workflow

Historical data

?



Machine learning model

# Our scenario

**Our dataset:** NYC property sales from 2015-2019

Includes:

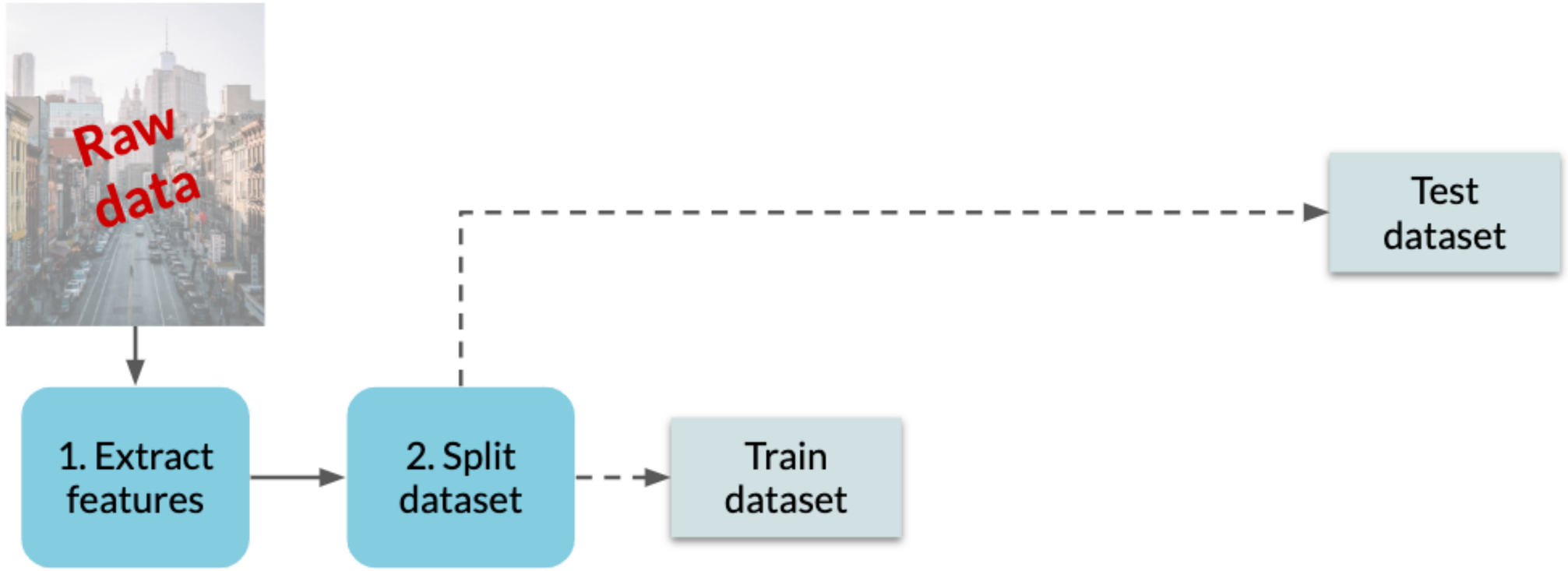
- Square feet
- Neighborhood
- Year built
- Sale price
- And more!

**Our target:** Sale price

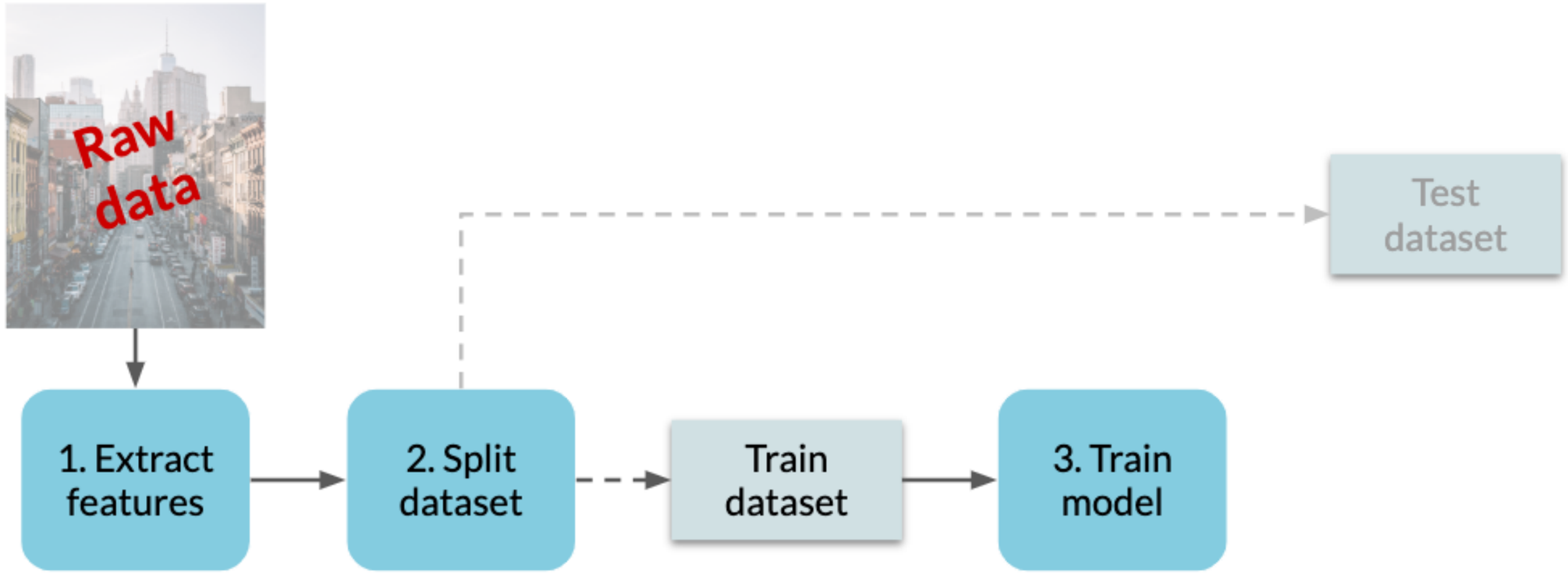
# Step 1: Extract features


1. Extract features

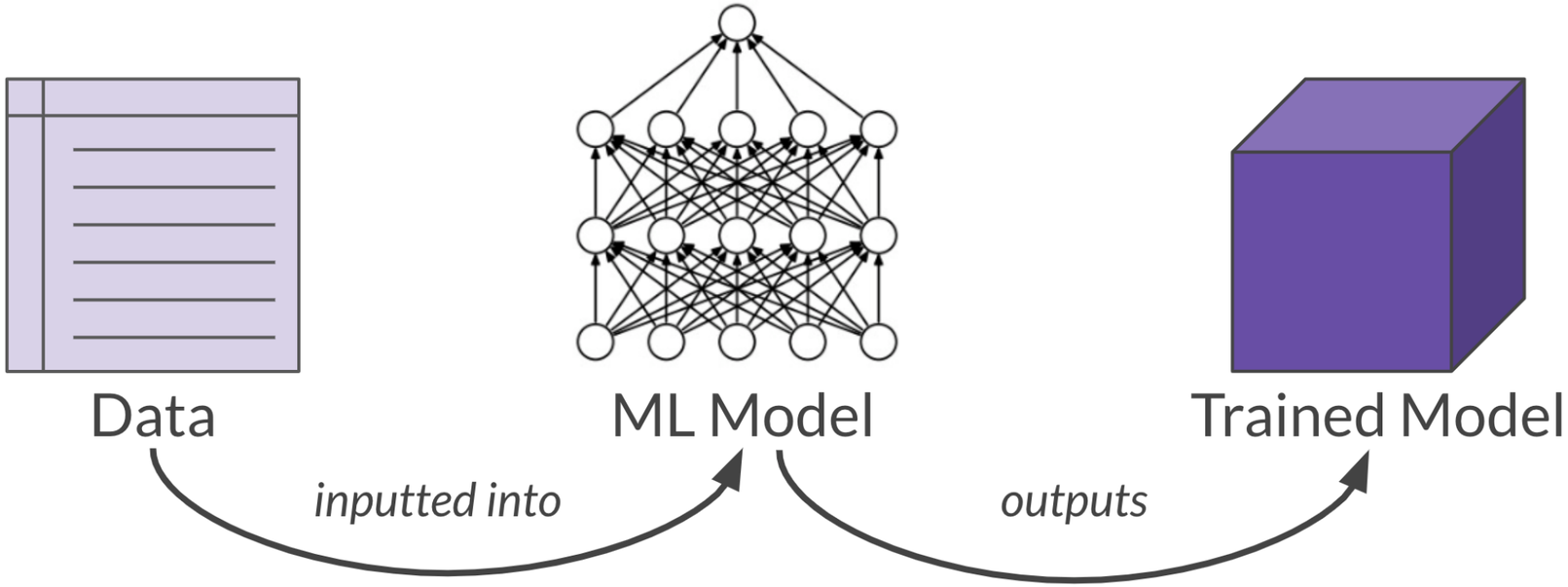
# Step 2: Split dataset



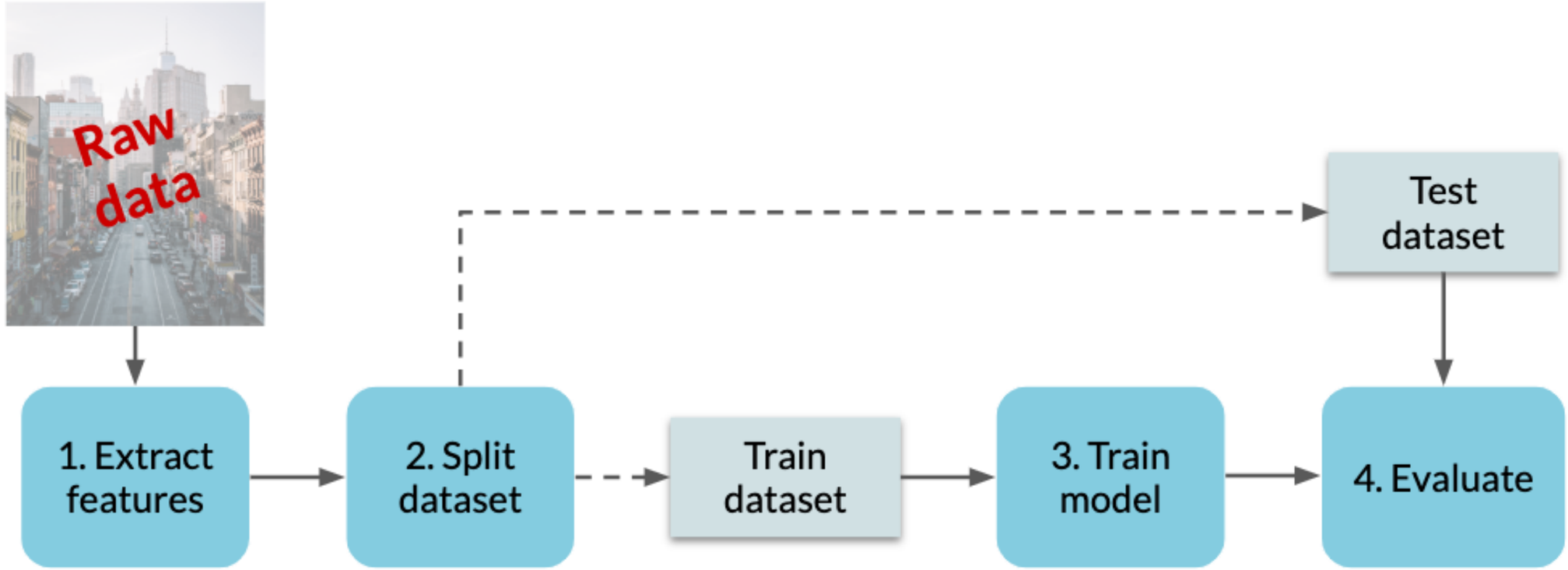
# Step 3: Train model



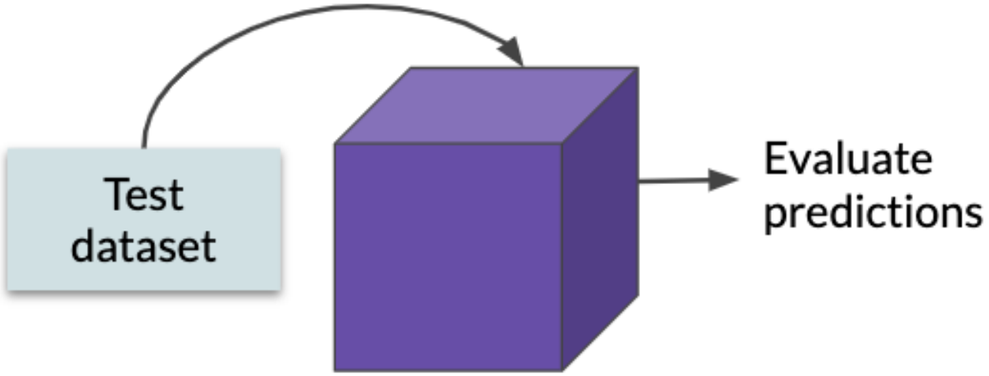
# Step 3: Train model



# Step 4: Evaluate

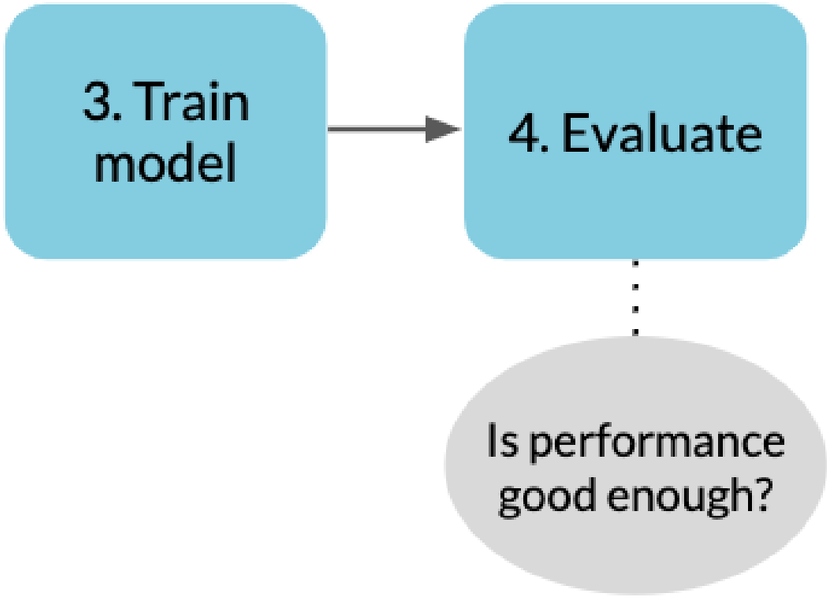


# Step 4: Evaluate

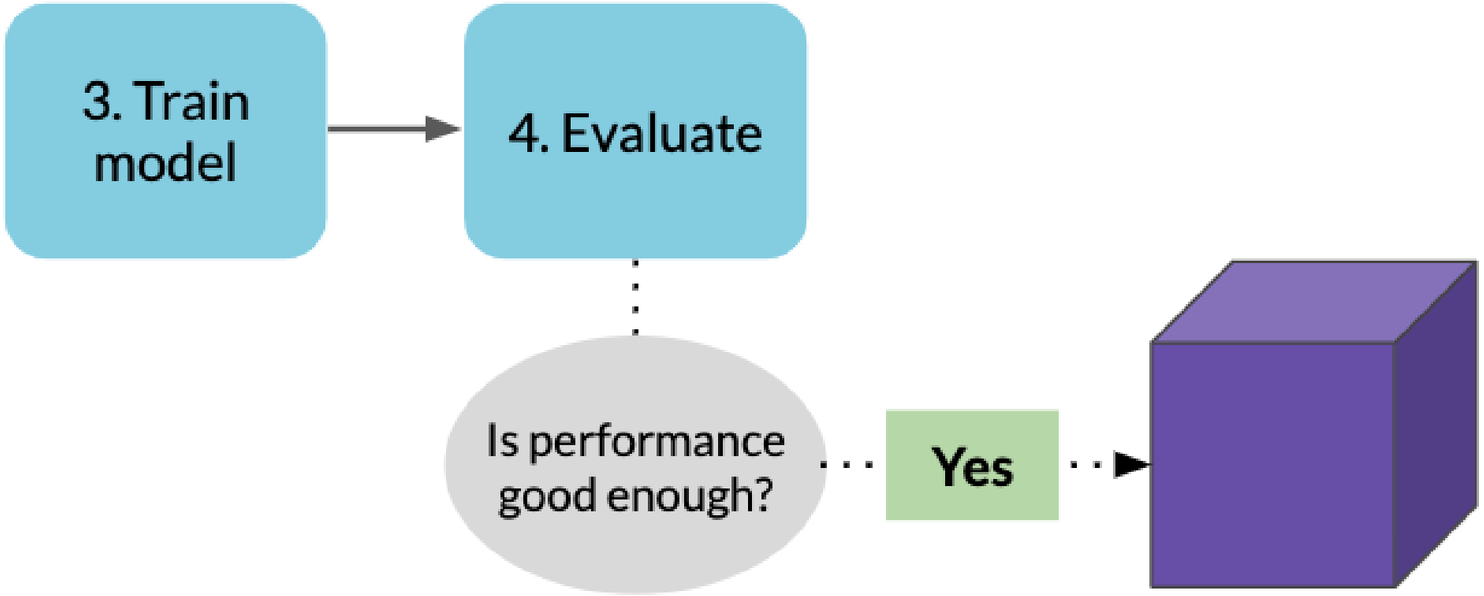



- Test dataset: "unseen" data
- Many ways to evaluate:
  - What is the average error of the predictions?
  - What percent of apartments did the model accurately predict within a 10% margin?

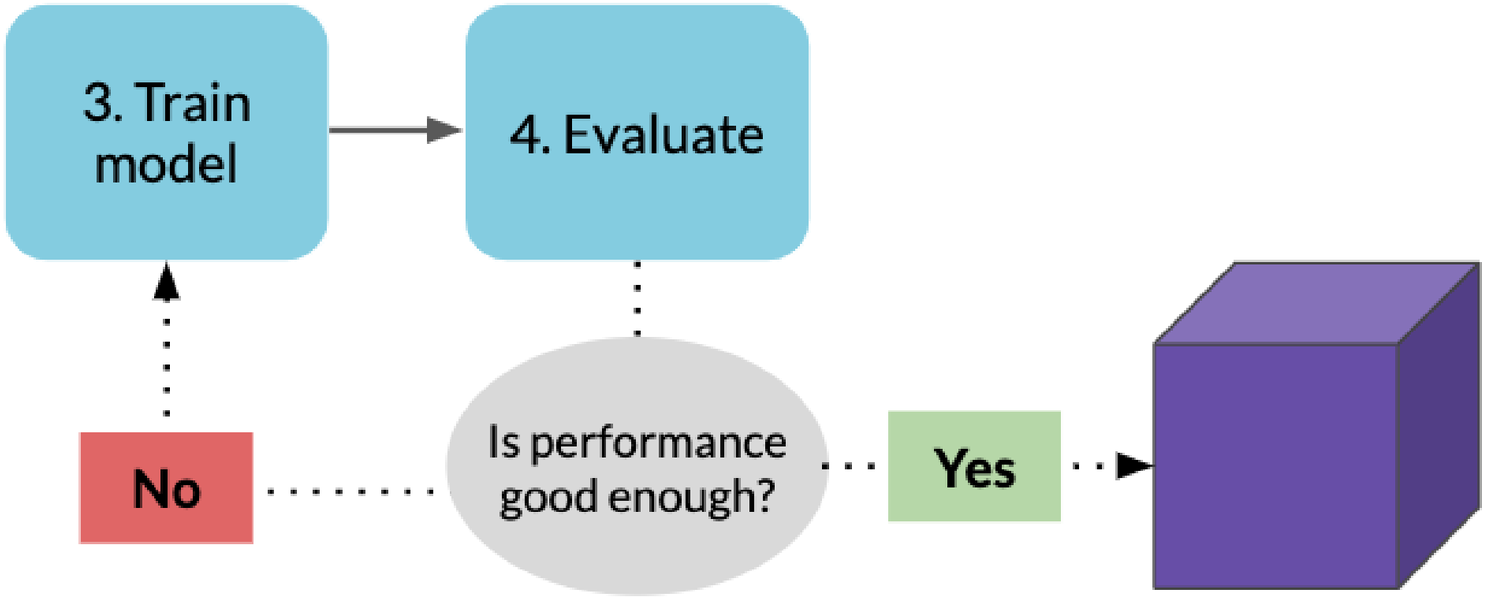
# Step 4: Evaluate



# Step 4: Evaluate

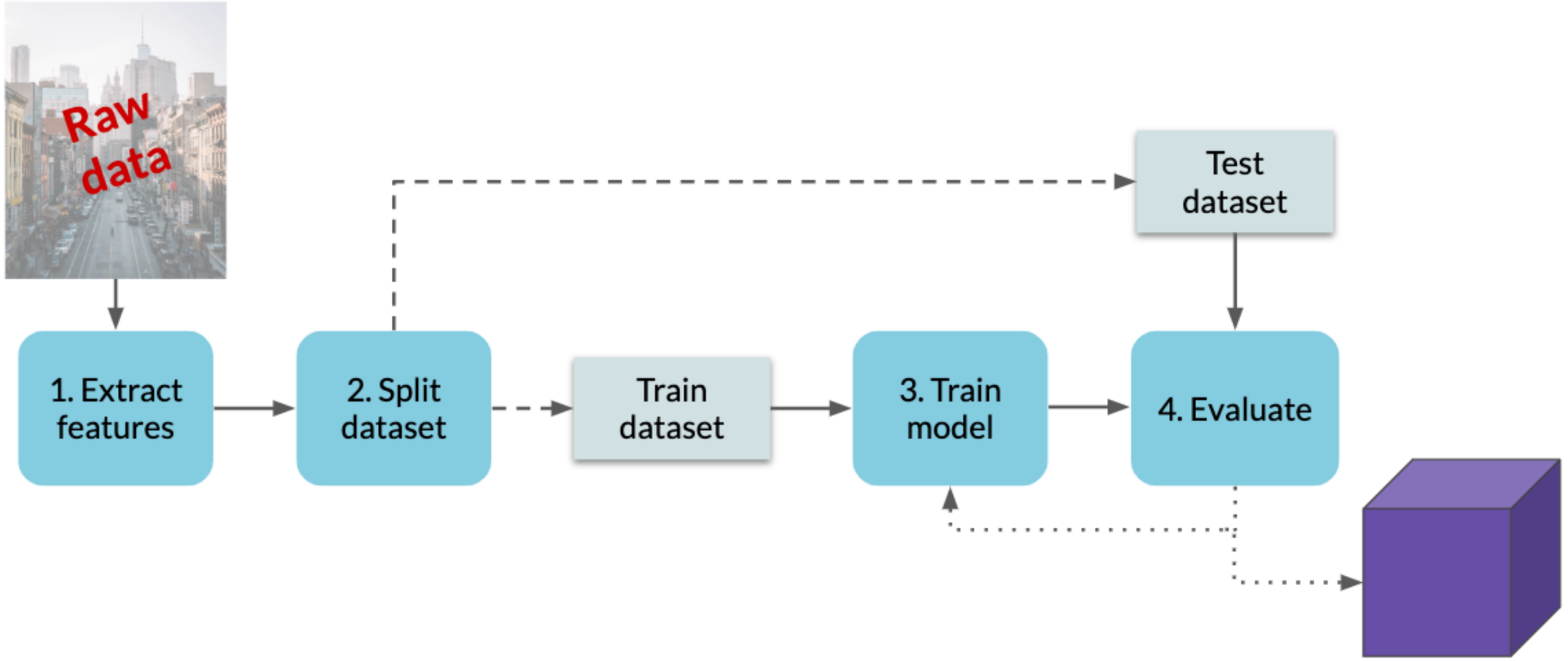


# Step 4: Evaluate




- If not, tune the model and re-train it:
  - e.g., change the model's options, add/remove features

# Machine learning workflow



# Summary of steps

1. **Extract features**
   - Choosing features and manipulating the dataset
2. **Split dataset**
   - Train and test dataset
3. **Train model**
   - Input train dataset into a machine learning model
4. **Evaluate**
   - *If desired performance isn't reached:* tune the model and repeat Step 3

# Let's practice!

UNDERSTANDING MACHINE LEARNING