Defining Machine Learning Problems

Amber Israelsen
AUTHOR | DEVELOPER | TRAINER

www.amberisraelsen.com

Course Overview

Course Introduction	Identifying Opportunities for Machine Learning	Defining Machine Learning Problems	Fetching and Preparing Data
Training and Evaluating the Model	Deploying and Monitoring the Model	The AWS Machine Learning Stack	Next Steps

Module Overview

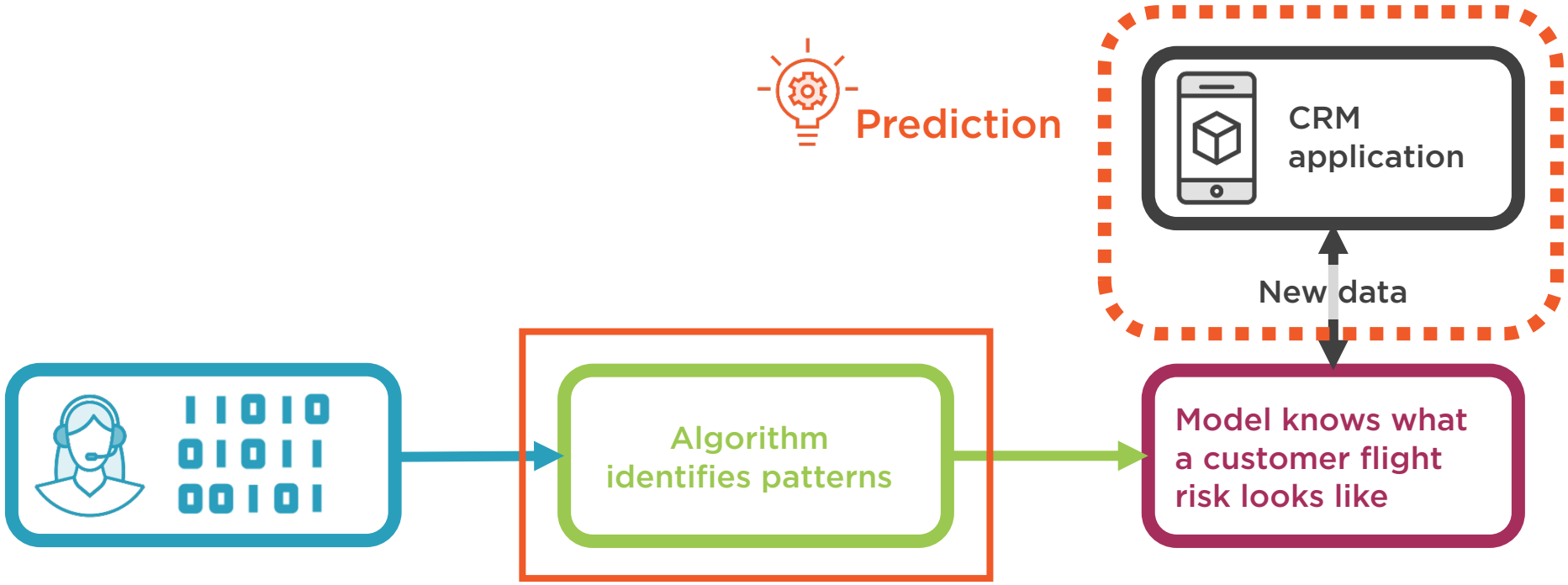
Types of machine learning problems

Defining the machine learning problem

- Concepts
- Application

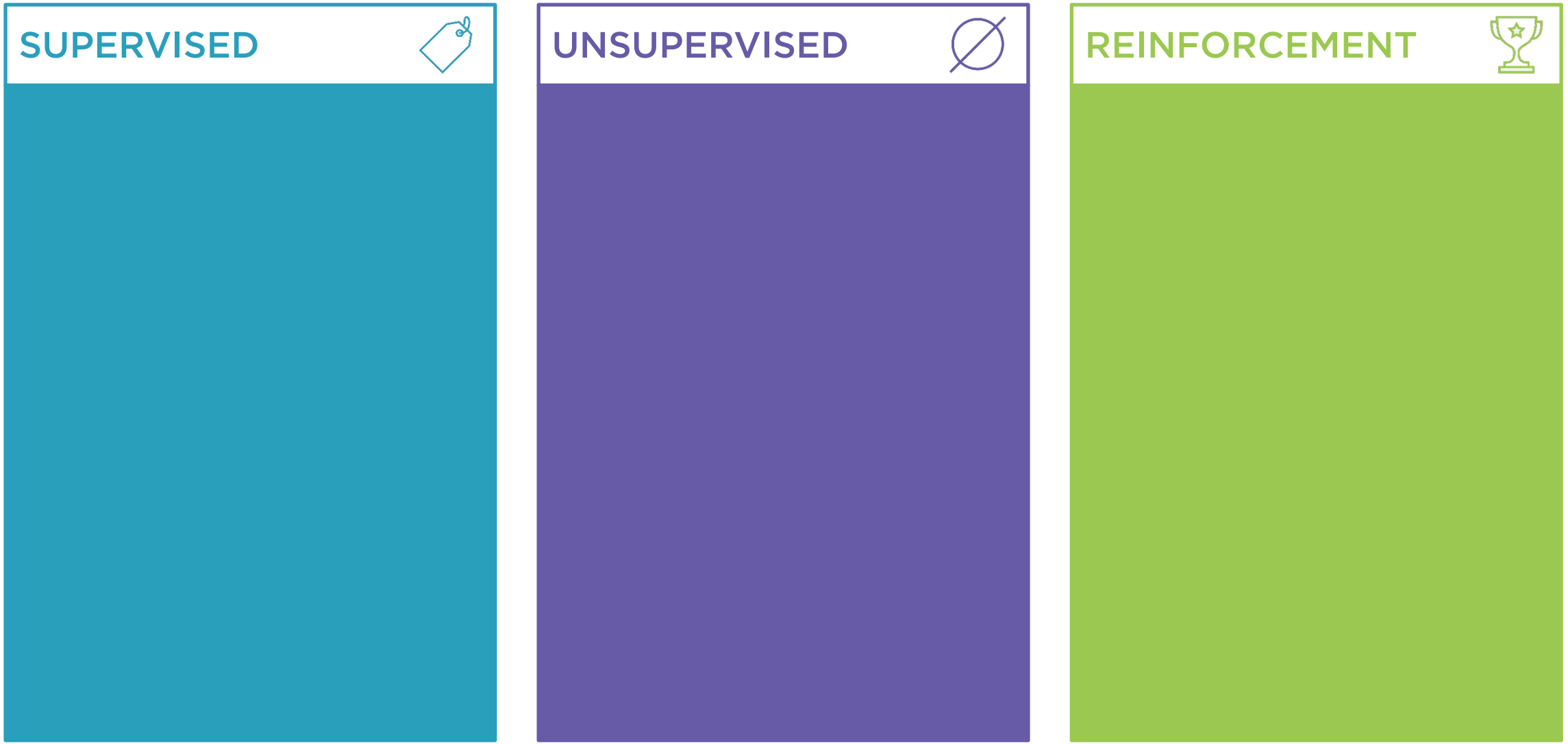
How can machine learning help us solve the problem of customer churn?

Example: Customer Churn



Types of Machine Learning Problems

Types of Machine Learning Problems



Is This Customer Likely to Leave?

Yes

No

Is This a Picture of a Cat?

Yes

No

Is This a Fraudulent Transaction?

Yes

No

Is This a Cat, Dog, or Bird?

Cat

Dog

Bird

Types of Machine Learning Problems

SUPERVISED

Learn from labeled data

Classification

BINARY

[yes/no]

[true/false]

[fraud, not fraud]

MULTICLASS

[cat, dog, bird]

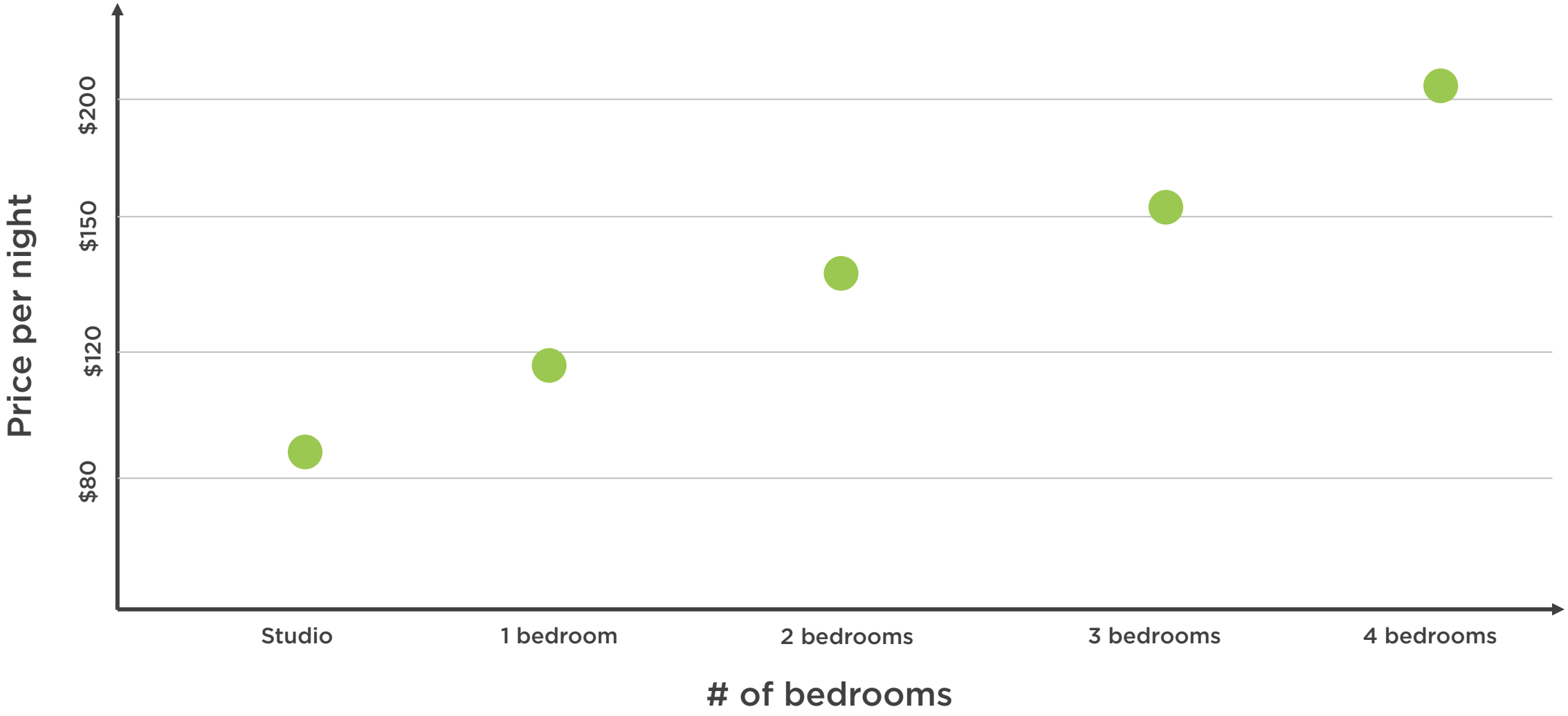
[house, condo, townhome, apartment]

UNSUPERVISED

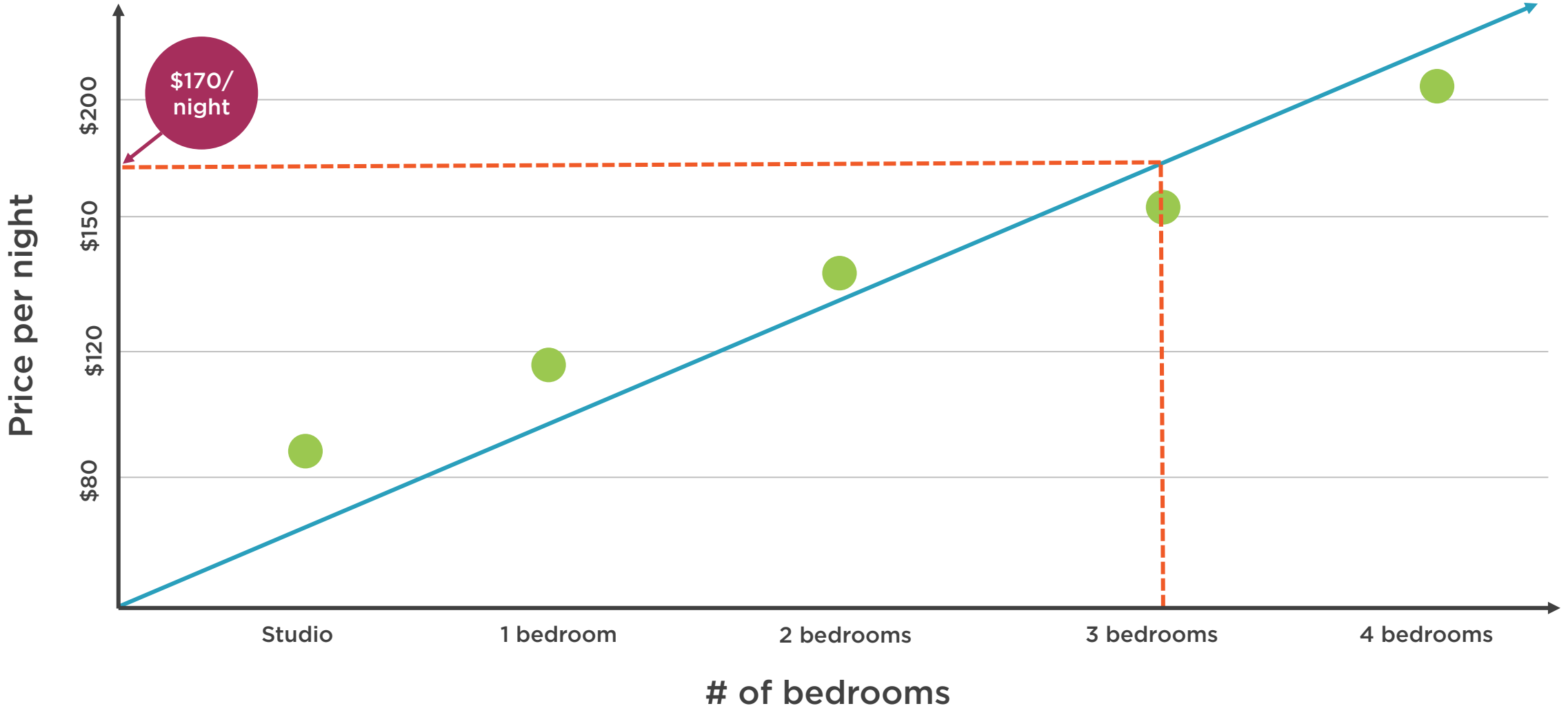
REINFORCEMENT

How Much Rent Should I Charge for My 3-bedroom Vacation Rental?

How Much Rent Should I Charge for My 3-bedroom Vacation Rental?



How Much Rent Should I Charge for My 3-bedroom Vacation Rental?



Types of Machine Learning Problems

SUPERVISED

Learn from labeled data

Classification

BINARY

[yes/no]

[true/false]

[fraud, not fraud]

MULTICLASS

[cat, dog, horse]

[house, condo, townhome, apartment]

Regression

Uses continuous values

[predicting a stock price]

[predicting sales of a product]

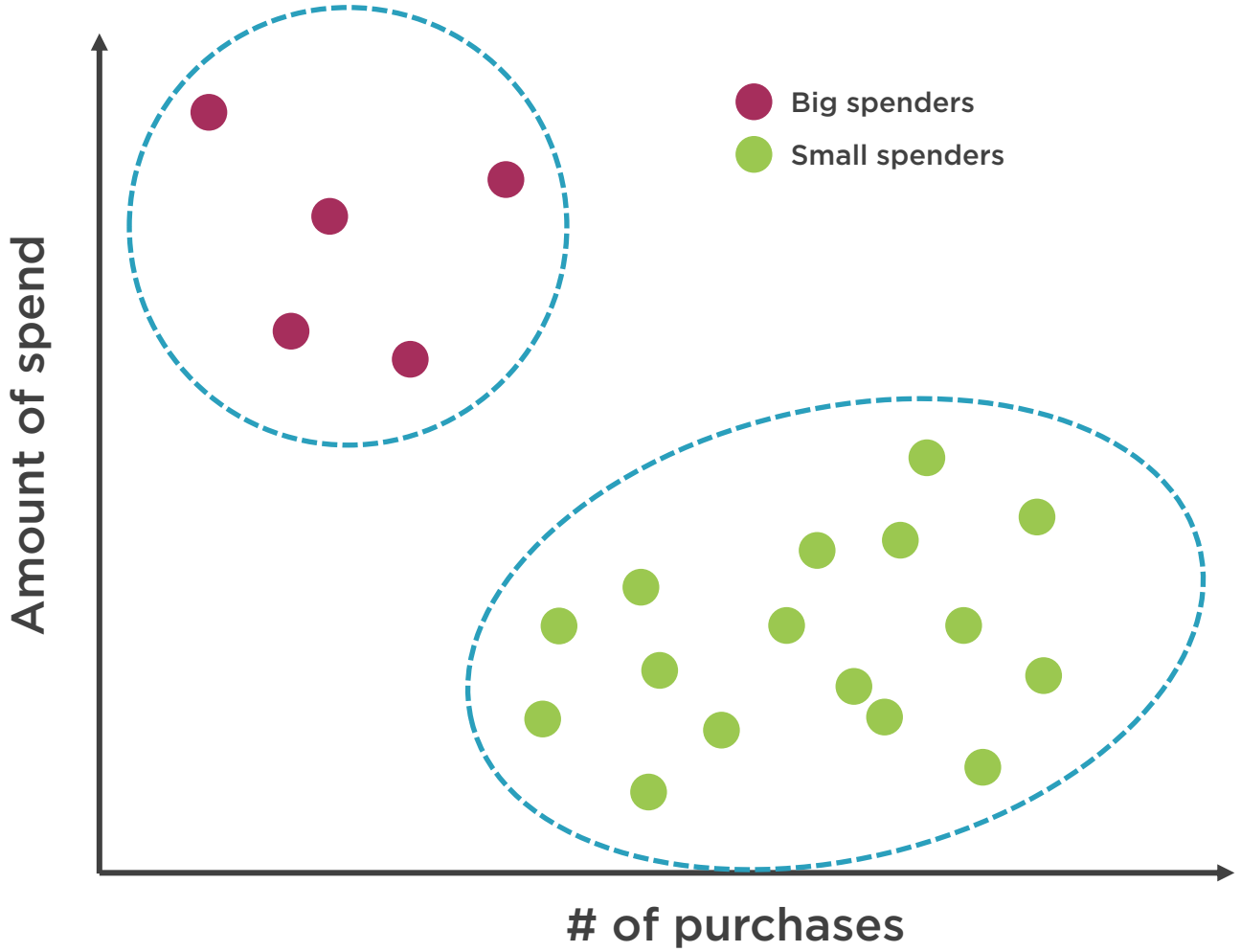
UNSUPERVISED

REINFORCEMENT

What Are My Customer Segments?

What Are My Customer Segments?

What Are My Customer Segments?



Types of Machine Learning Problems

SUPERVISED

Learn from labeled data

Classification

BINARY

[yes/no]

[true/false]

[fraud, not fraud]

MULTICLASS

[cat, dog, horse]

[house, condo, townhome, apartment]

Regression

Uses continuous values

[predicting a stock price]

[predicting sales of a product]

UNSUPERVISED

Learn from finding hidden patterns in unlabeled data

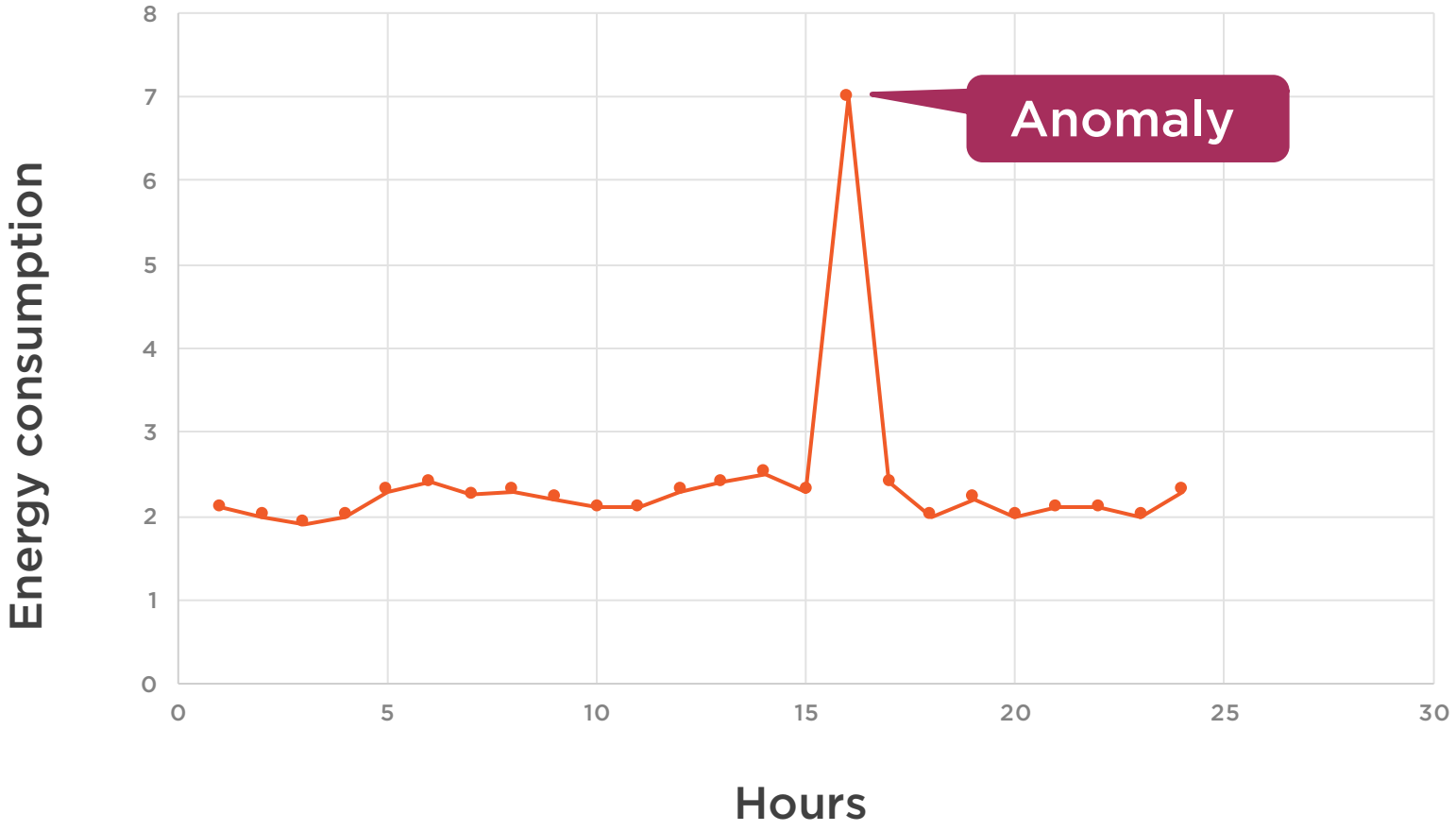
Clustering

Groups data into clusters based on similar features

[after analyzing patient data, you find that certain groups respond better to treatment]

REINFORCEMENT

Are There Anomalies in Electricity Usage?



Types of Machine Learning Problems

SUPERVISED

Learn from labeled data

Classification

BINARY

[yes/no]

[true/false]

[fraud, not fraud]

MULTICLASS

[cat, dog, horse]

[house, condo, townhome, apartment]

Regression

Uses continuous values

[predicting a stock price]

[predicting sales of a product]

UNSUPERVISED

Learn from finding hidden patterns in unlabeled data

Clustering

Groups data into clusters based on similar features

[after analyzing patient data, you find that certain groups respond better to treatment]

Anomaly Detection

Finds outliers in data

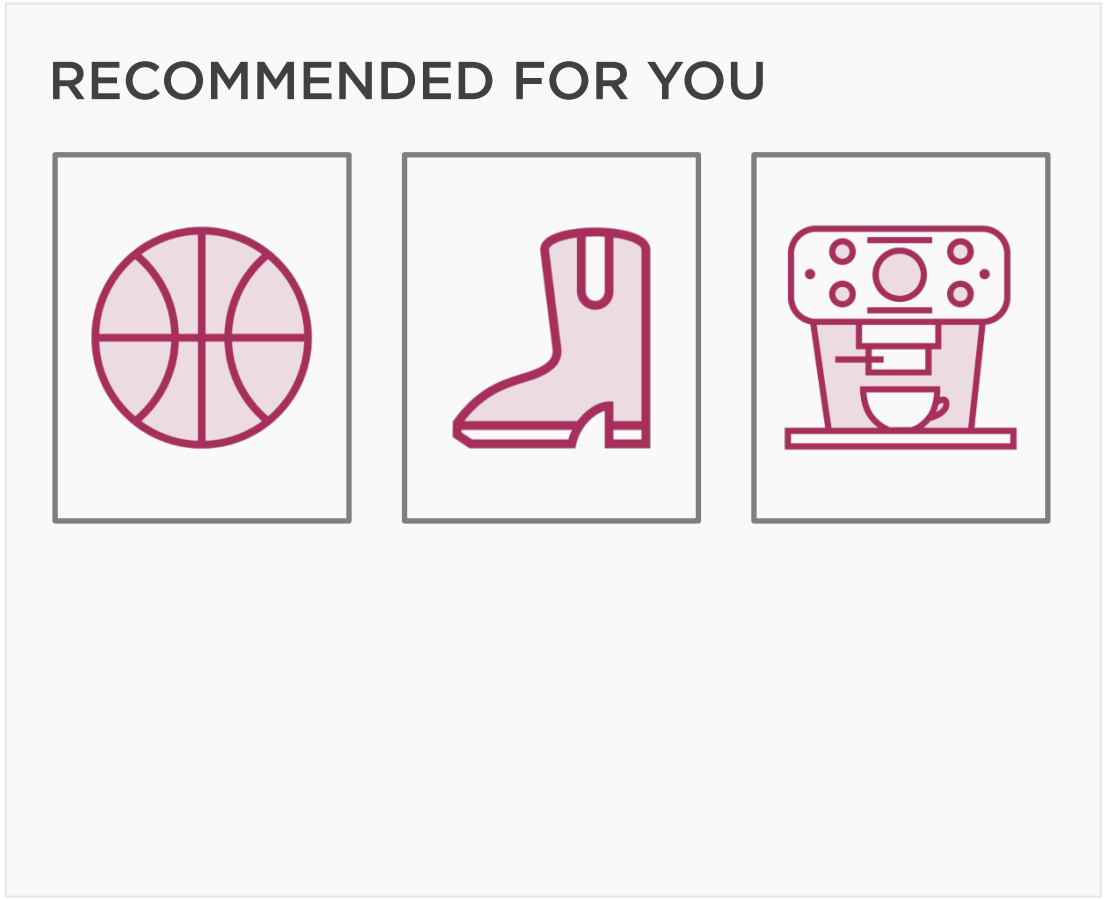
[suspicious network traffic]

[abnormal heart beat]

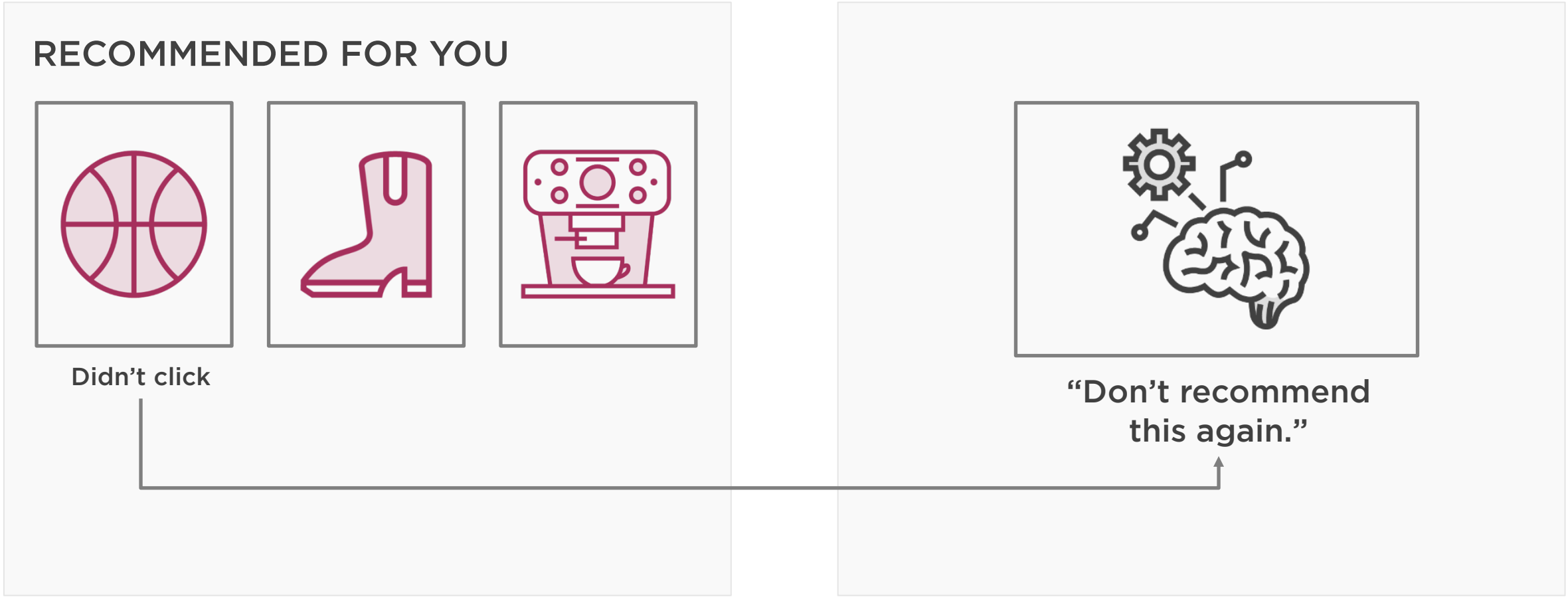
REINFORCEMENT

How Can I Cross-sell or Up-sell My Customers?

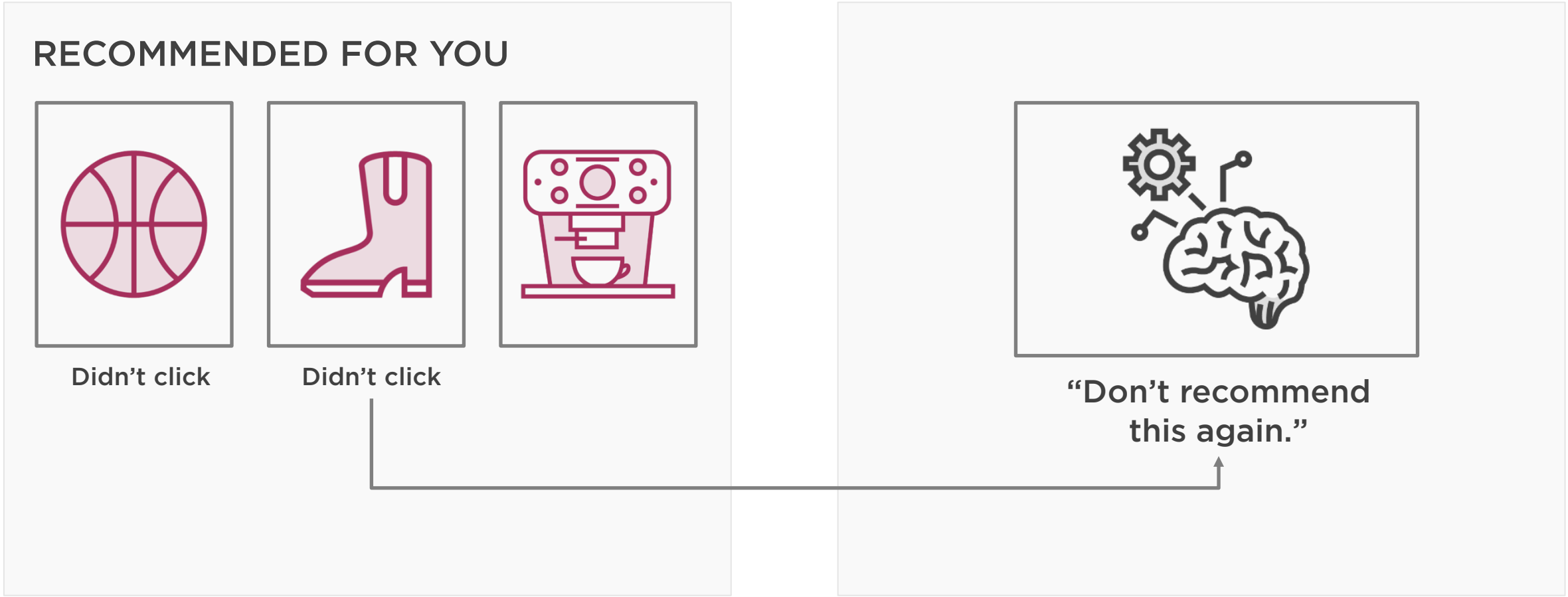
How Can I Cross-sell or Up-sell My Customers?



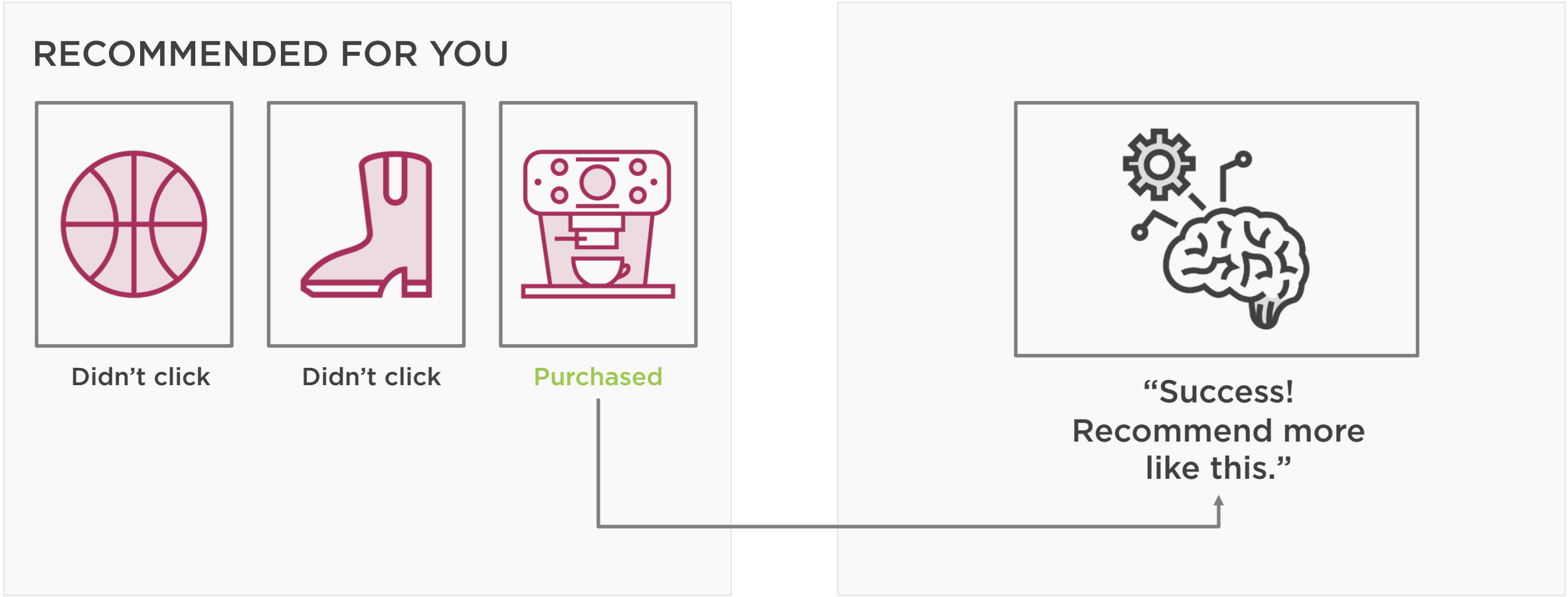
How Can I Cross-sell or Up-sell My Customers?



How Can I Cross-sell or Up-sell My Customers?



How Can I Cross-sell or Up-sell My Customers?



Types of Machine Learning Problems

SUPERVISED

Learn from labeled data

Classification

BINARY

[yes/no]

[true/false]

[fraud, not fraud]

MULTICLASS

[cat, dog, horse]

[house, condo, townhome, apartment]

Regression

Uses continuous values

[predicting a stock price]

[predicting sales of a product]

UNSUPERVISED

Learn from finding hidden patterns in unlabeled data

Clustering

Groups data into clusters based on similar features

[after analyzing patient data, you find that certain groups respond better to treatment]

Anomaly Detection

Finds outliers in data

[suspicious network traffic]

[abnormal heart beat]

REINFORCEMENT

Learn from interacting with the environment to maximize reward

The algorithm receives feedback when it does something right or wrong

[gaming]

[recommendation systems]

[robot navigation]

[self-driving cars, like Amazon's Deep Racer]

Defining the Machine Learning Problem

What is the problem you're trying to solve?

Business Problem

The company needs to predict sales of products in order to inform manufacturing decisions

Machine Learning Problem Example 1

“Predict which products will sell more than 100 units”

TARGET:

TYPE:

Target

Final output you’re trying to predict

Features (the columns)
(a.k.a. “attributes”)

Label/
Target

Product Name	Price	Free Shipping	Release Date	Discount Eligible	Part of Bundle	Sold More Than 100 Units
3D globe puzzle	$29.95	Yes	12/2017	Yes	Yes	True
Passport holder	$20.45	Yes	2/2018	No	No	True
Women’s travel pants	$47.88	No	4/2017	Yes	Yes	False
Scratch-it maps	$19.99	No	12/2018	No	Yes	False
Luggage tags	$12.99	Yes	3/2017	Yes	No	True
Monogrammed luggage	$129.79	No	2/2019	Yes	Yes	False
Men’s travel vest	$39.99	No	11/2017	Yes	No	False

Observations (the rows)

Machine Learning Problem Example 1

"Predict which products will sell more than 100 units"

TARGET: Binary (true/false)

TYPE: Binary classification (supervised)

Machine Learning Problem Example 2

“Predict the number of sales for each product”

TARGET:

TYPE:

Machine Learning Problem Example 2

"Predict the number of sales for each product"

TARGET: Numerical value (continuous)

TYPE: Regression (supervised)

Questions to Ask

What is the current pain point?

How are we currently solving this problem?

What data do we have?

Is the data labeled?

How will we define success?

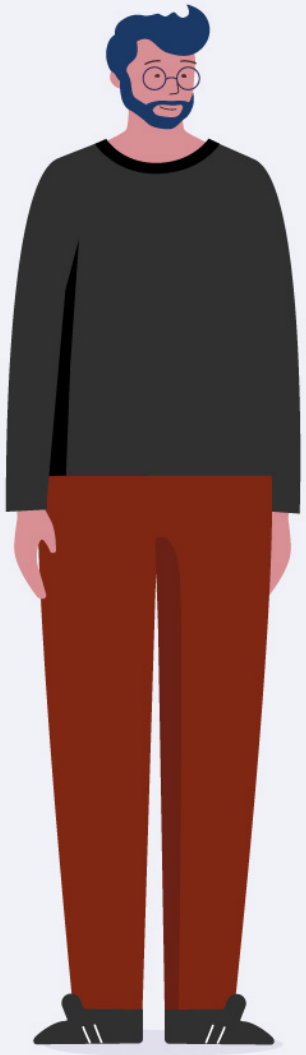
What are the trade-offs?

What is out of scope?

Scenario: Applying the Concepts

GLOBOMANTICS

Wireless Carrier



Ray

- VP of Customer Experience
- Concerned about customer churn

Questions to Ask

What is the current pain point?

- Customer churn rate last year was 14.5%
- We have no way to predict who will leave so we can incentivize them to stay

How are we currently solving this problem?

What data do we have?

Is the data labeled?

How will we define success?

What are the trade-offs?

What is out of scope?

Questions to Ask

What is the current pain point?

How are we currently solving this problem?

- If a customer gets angry during a service call and threatens to leave, a customer service agent can offer them an incentive

What data do we have?

Is the data labeled?

How will we define success?

What are the trade-offs?

What is out of scope?

Questions to Ask

What is the current pain point?

How are we currently solving this problem?

What data do we have?

Is the data labeled?

How will we define success?

What are the trade-offs?

What is out of scope?

Features (the columns)

Label/Target ("Churn?")

State	Account Length	Area Code	Phone	Int'l Plan	VMail Plan	VMail Message	Day Mins	Day Calls	Day Charge	Eve Mins	Eve Calls	Eve Charge	Night Mins	Night Calls	Night Charge	Intl Mins	Intl Calls	Intl Charge	CustServ Calls	Churn?
KS	128	415	382-4657	no	yes	25	265.1	110	45.07	197.4	99	16.78	244.7	91	11.01	10	3	2.7	1	False.
OH	107	415	371-7191	no	yes	26	161.6	123	27.47	195.5	103	16.62	254.4	103	11.45	13.7	3	3.7	1	False.
NJ	137	415	358-1921	no	no	0	243.4	114	41.38	121.2	110	10.3	162.6	104	7.32	12.2	5	3.29	0	False.
OH	84	408	375-9999	yes	no	0	299.4	71	50.9	61.9	88	5.26	196.9	89	8.86	6.6	7	1.78	2	False.
OK	75	415	330-6626	yes	no	0	166.7	113	28.34	148.3	122	12.61	186.9	121	8.41	10.1	3	2.73	3	False.
AL	118	510	391-8027	yes	no	0	223.4	98	37.98	220.6	101	18.75	203.9	118	9.18	6.3	6	1.7	0	False.
MA	121	510	355-9993	no	yes	24	218.2	88	37.09	348.5	108	29.62	212.6	118	9.57	7.5	7	2.03	3	False.
MO	147	415	329-9001	yes	no	0	157	79	26.69	103.1	94	8.76	211.8	96	9.53	7.1	6	1.92	0	False.
LA	117	408	335-4719	no	no	0	184.5	97	31.37	351.6	80	29.89	215.8	90	9.71	8.7	4	2.35	1	False.
WV	141	415	330-8173	yes	yes	37	258.6	84	43.96	222	111	18.87	326.4	97	14.69	11.2	5	3.02	0	False.
IN	65	415	329-6603	no	no	0	129.1	137	21.95	228.5	83	19.42	208.8	111	9.4	12.7	6	3.43	4	True.
RI	74	415	344-9403	no	no	0	187.7	127	31.91	163.4	148	13.89	196	94	8.82	9.1	5	2.46	0	False.
IA	168	408	363-1107	no	no	0	128.8	96	21.9	104.9	71	8.92	141.1	128	6.35	11.2	2	3.02	1	False.
MT	95	510	394-8006	no	no	0	156.6	88	26.62	247.6	75	21.05	192.3	115	8.65	12.3	5	3.32	3	False.
VA	76	510	356-2992	no	yes	33	189.7	66	32.25	212.8	65	18.09	165.7	108	7.46	10	5	2.7	1	False.
TX	73	415	373-2782	no	no	0	224.4	90	38.15	159.5	88	13.56	192.8	74	8.68	13	2	3.51	1	False.
FL	147	415	396-5800	no	no	0	155.1	117	26.37	239.7	93	20.37	208.8	133	9.4	10.6	4	2.86	0	False.
CO	77	408	393-7984	no	no	0	62.4	89	10.61	169.9	121	14.44	209.6	64	9.43	5.7	6	1.54	5	True.
AZ	130	415	358-1958	no	no	0	183	112	31.11	72.9	99	6.2	181.8	78	8.18	9.5	19	2.57	0	False.
SC	111	415	350-2565	no	no	0	110.4	103	18.77	137.3	102	11.67	189.6	105	8.53	7.7	6	2.08	2	False.
VA	132	510	343-4696	no	no	0	81.1	86	13.79	245.2	72	20.84	237	115	10.67	10.3	2	2.78	0	False.

Observations (the rows)

Machine Learning Problem

"Predict if a customer will leave"

TARGET: Binary (true/false)

TYPE: Binary classification (supervised)

Questions to Ask

What is the current pain point?

How are we currently solving this problem?

What data do we have?

Is the data labeled?

How will we define success?

- Customer churn rate of 4%
- Customer retention incentives not to exceed $750,000

What are the trade-offs?

What is out of scope?

Questions to Ask

What is the current pain point?

How are we currently solving this problem?

What data do we have?

Is the data labeled?

How will we define success?

What are the trade-offs?

- Some customers will be given incentives even though they were not churn risks
- Some customers who are given incentives will leave anyway

What is out of scope?

Questions to Ask

What is the current pain point?

How are we currently solving this problem?

What data do we have?

Is the data labeled?

How will we define success?

What are the trade-offs?

What is out of scope?

- Deactivation or downgrade of service or features
- Sales experience and performance of the customer service agent

Key Points to Remember

Supervised learning

- Learn from labeled data
- Binary/multiclass classification and regression

Unsupervised learning

- Learn from finding hidden patterns in unlabeled data
- Clustering and anomaly detection

Reinforcement learning

- Learn from interacting with the environment to maximize reward
- e.g., recommendation systems, self-driving cars

Features

- The “columns” of data supplied as input

Target

- The final output you’re trying to predict (a “label” on the input)

Observations

- The “rows” of data supplied as input

Spend ample time defining the problem, inputs, outputs and success metrics

Up Next:
Fetching and Preparing Data