:: From business problem to Machine learning problem: a Recipe

Problem formulation

2

Can you formulate your problem clearly?

- What do you want to predict given which input?
- Pattern: “given X, predict Y”
  - What is the input?
  - What is the output?

Example: sentiment analysis

- Given a customer review, predict its sentiment
  - Input: customer review text
  - Output: positive, negative, neutral

Source: SAP, https://open.sap.com/courses/ml1

:: From business problem to Machine learning problem: a Recipe

Collecting Data

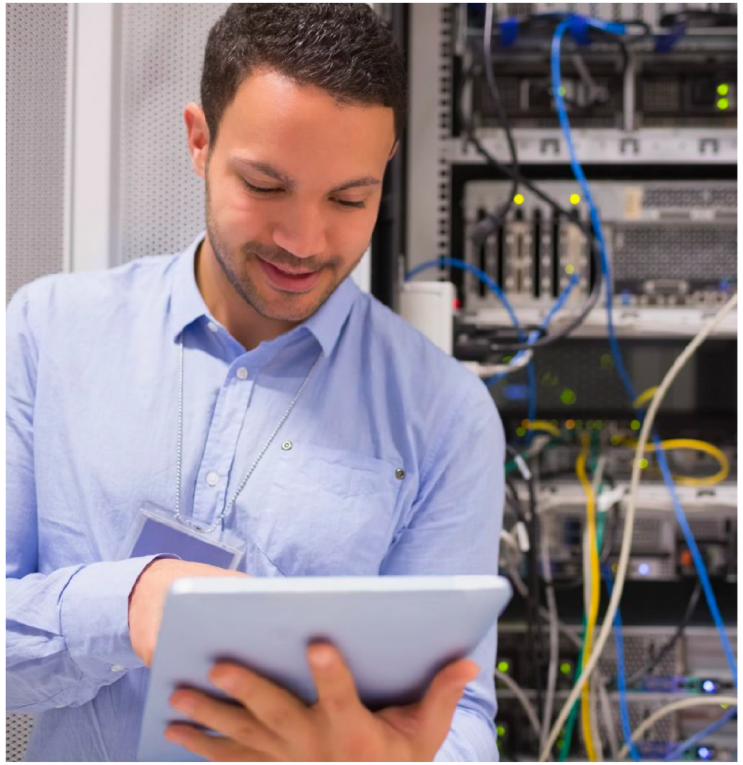
3

Do you have sufficient examples?

- Machine learning always requires data!
- Generally, the more data, the better
- Each example must contain two parts (supervised learning)
    - Features: attributes of the example
    - Label: the answer you want to predict

Example: sentiment analysis

- Thousands of customer reviews and ratings from the Web



Source: SAP, https://open.sap.com/courses/ml1

:: From business problem to Machine learning problem: a Recipe

Regularities in the data

4

Does your problem have a regular pattern?

- Machine learning learns regularities and patterns
- Hard to learn patterns that are rare or irregular

Example: sentiment analysis

- Positive words like *good*, *awesom*e, or *love* it appear more often in highly-rated reviews
- Negative words like *ba*d, *lous*y, or *disappointed* appear more often in poorly-rated reviews

Source: SAP, https://open.sap.com/courses/ml1

:: From business problem to Machine learning problem: a Recipe

Representations and features

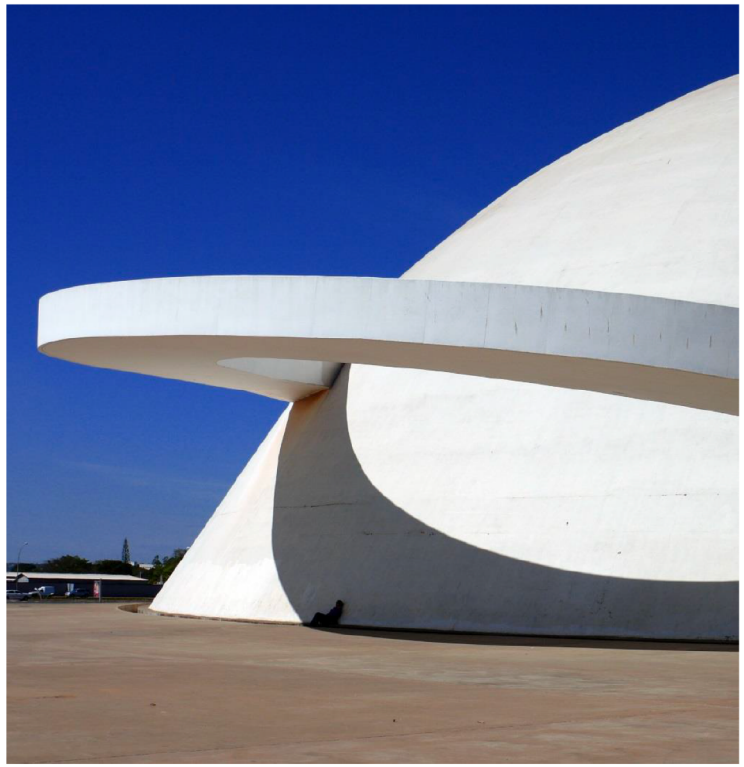
5

Can you find meaningful representations of your data?

- Machine learning algorithms ultimately operate on numbers
- Generally, examples are represented as feature vectors
- Good features often determine the success of machine learning

Example: sentiment analysis

- Represent customer review as vector of word frequencies
- Label is positive (4-5 stars), negative (1-2 stars), neutral (3 stars)
- Picture of data or a mathematical vector



Source: SAP, https://open.sap.com/courses/ml1

:: From business problem to Machine learning problem: a Recipe

Evaluating success

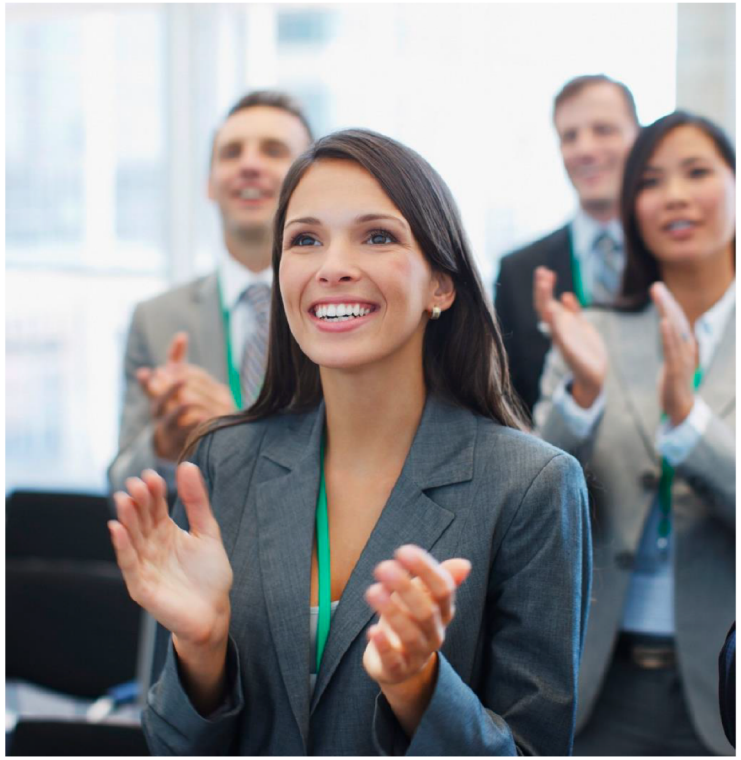
6

How do you define success?

- Machine learning optimizes a training criteria
- The evaluation function has to support the business goals

Example: sentiment analysis

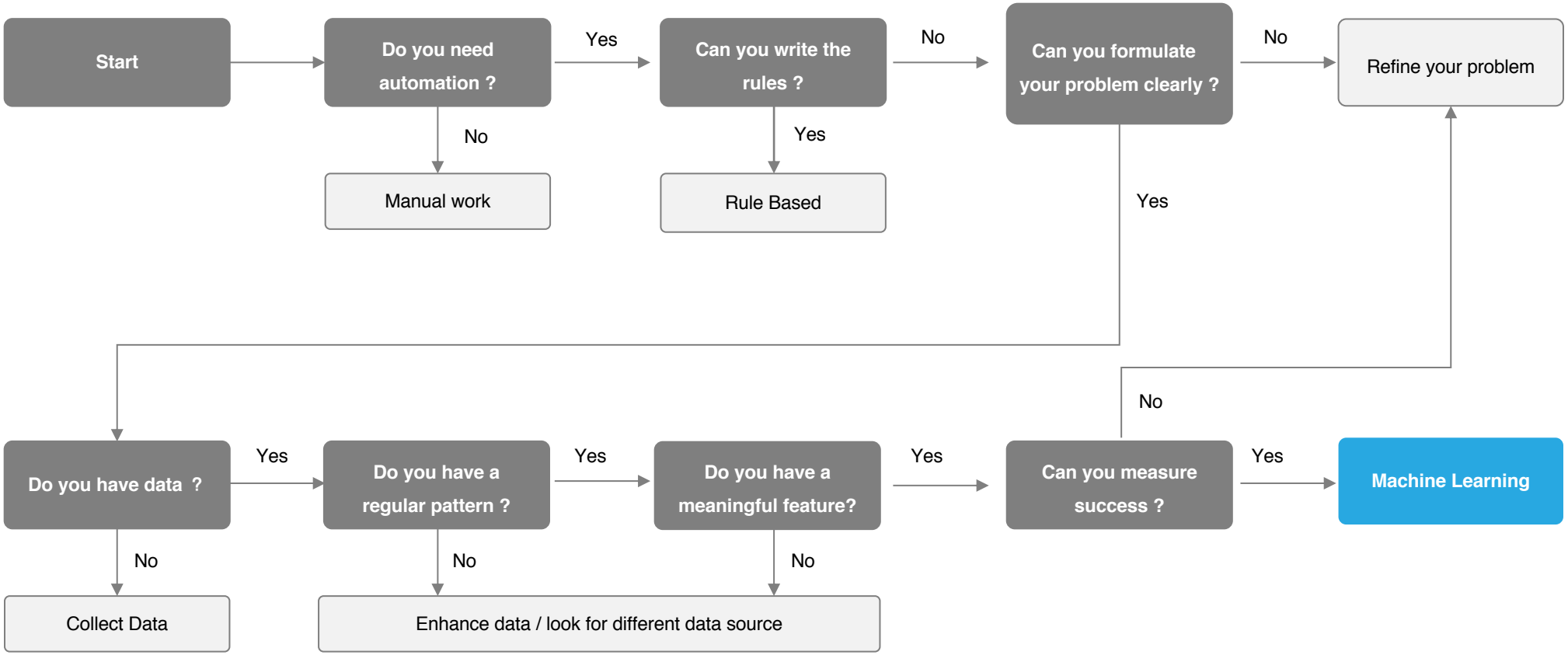
- Accuracy: percentage of correctly predicted labels



Source: SAP, https://open.sap.com/courses/ml1

:: From business problem to Machine learning problem: a Recipe

The “cheat sheet”



Source: SAP, https://open.sap.com/courses/ml1

:: Application Example: Natural Language Processing

Example 1: Support ticket classification

Classify support tickets into categories so that they can be routed to corresponding agents

1. **Do you need machine learning?**
   - High volume of support tickets
   - Human language is complex and ambiguous

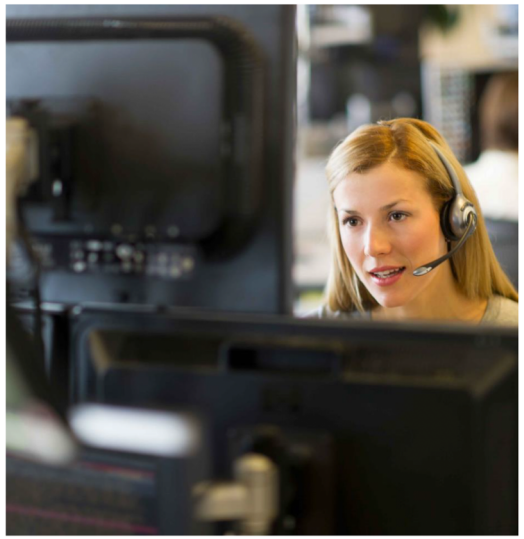
2. **Can you formulate your problem clearly?**
   - Given a customer support ticket, predict its service category
   - Input: customer support ticket; output: service category

3. **Do you have sufficient examples?**
   - Large volume of customer support tickets with respective service category from ticket support systems

4. **Does your problem have a regular pattern?**
   - Common customer issues will have many tickets
   - Issues will correlate with common keywords, e.g., bill or payment will appear more often in support tickets with category payments

5. **Can you find meaningful representations of your data?**
   - Represent customer support tickets as vector of word frequencies
   - Label is the service category of the customer support ticket

6. **How do you define success?**
   - Measure percentage of correctly predicted service categories



Source: SAP, Enterprise Machine Learning in a Nutshell

:: Application Example: Natural Language Processing

Example 2: Retail shelf analytics

Given a picture of a retail shelf, detect all products in the picture and compare with planned layout

1. **Do you need machine learning?**
   - High manual effort to monitor store shelves every day
   - Detecting products in images not possible with simple rules

2. **Can you formulate your problem clearly?**
   - Given a shelf image, first detect products and then compare their positions with the planned shelf layout
   - Input: photo; output: bounding boxes of products

3. **Do you have sufficient examples?**
   - Large volume of collected retail shelf pictures, manually labelled bounding boxes of products

4. **Does your problem have a regular pattern?**
   - Product packaging has regular shape, colors, and logos

5. **Can you find meaningful representations of your data?**
   - Represent photos as array of pixel values
   - Image patches with products are positive examples; random patches are negative examples

6. **How do you define success?**
   - Measure precision and recall of predicted bounding
   - boxes and similarity of the detected layout to the true layout

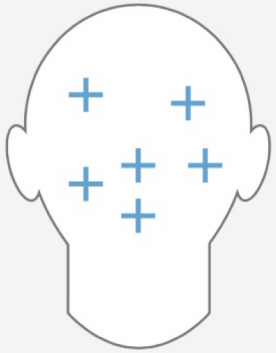
Source: SAP, Enterprise Machine Learning in a Nutshell

Wrap-up: When should you use machine learning ?

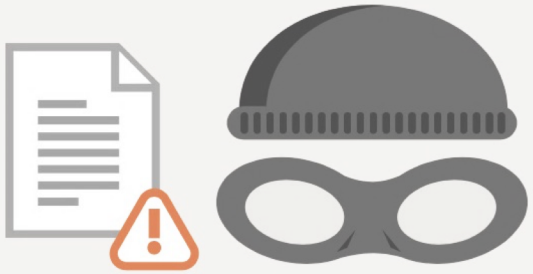
Consider using machine learning when you have a complex task or problem involving a large amount of data and lots of variables, but no existing formula or equation.

For example, machine learning is a good option if you need to handle situations like these:

Hand-written rules and equations are too complex—as in face recognition and speech recognition.



The rules of a task are constantly changing—as in fraud detection from transaction records.

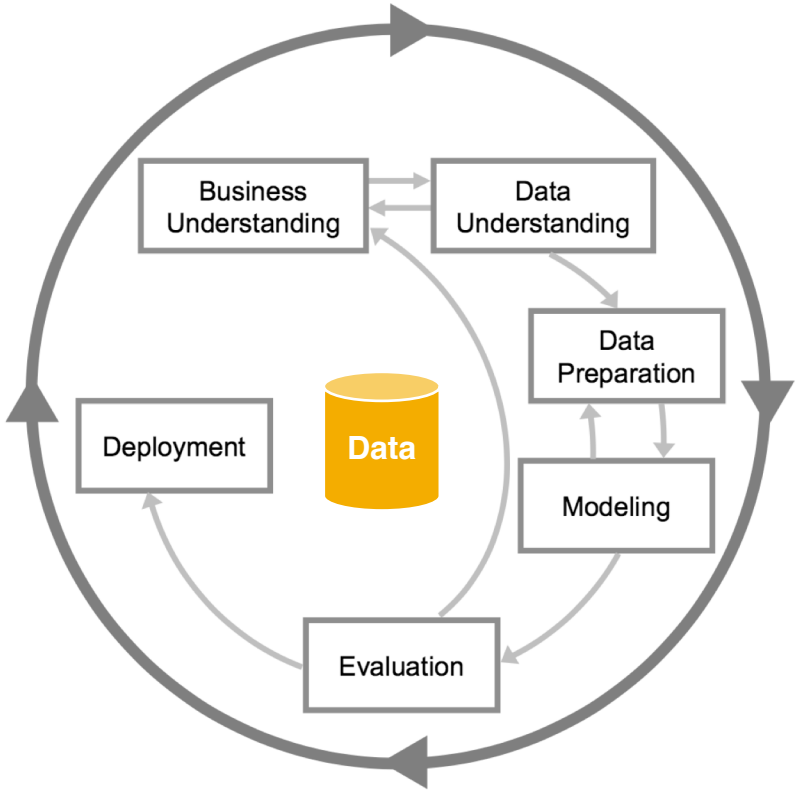


The nature of the data keeps changing, and the program needs to adapt—as in automated trading, energy demand forecasting, and predicting shopping trends.

Source: Mathworks

MathWorks

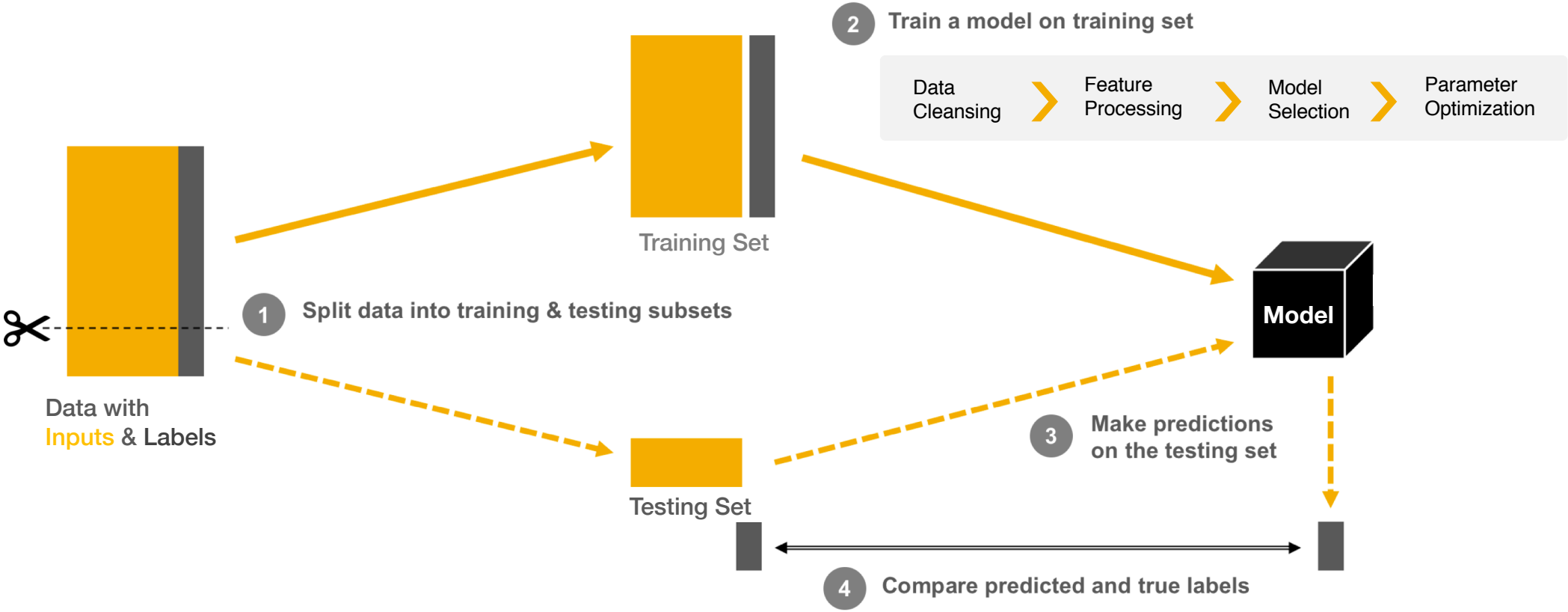
Process of Machine Learning



Cross-industry standard process for data mining (CRISP-DM)

Diagram: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining
Image: Designed by Asierromero / Freepik

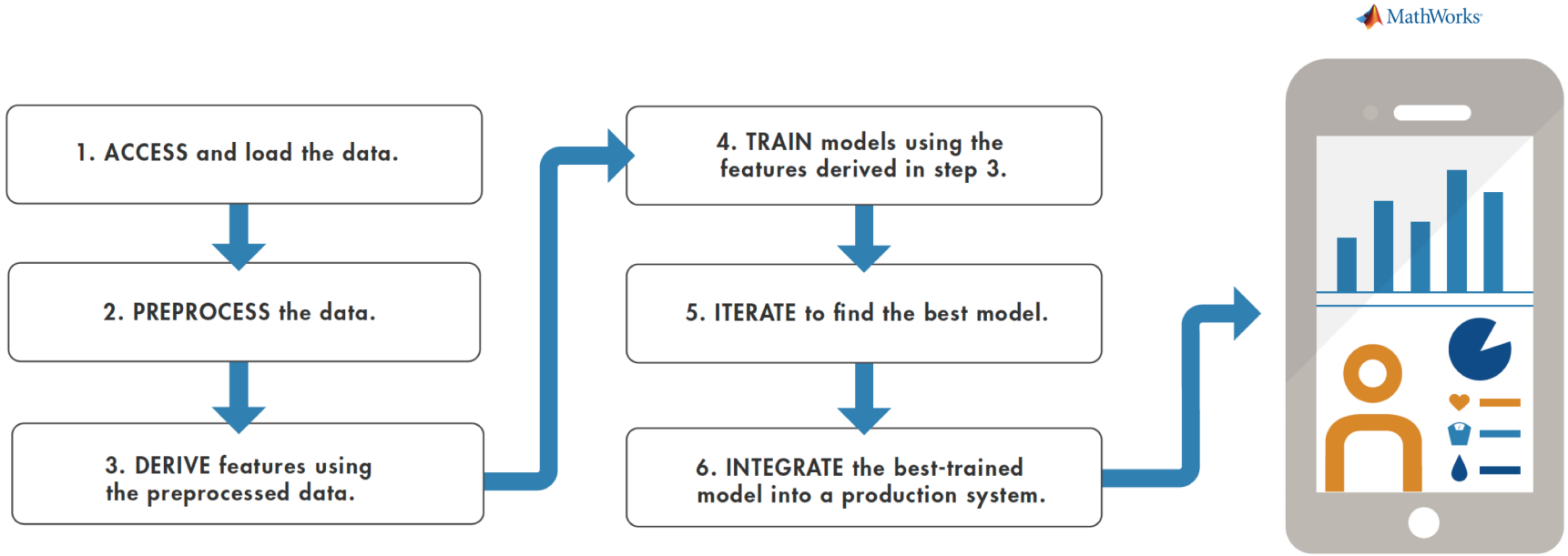
:: How to create machine learning models ?



Source : SAP

:: Machine Learning Workflow at a Glance

Workflow by MathWorks



Note:

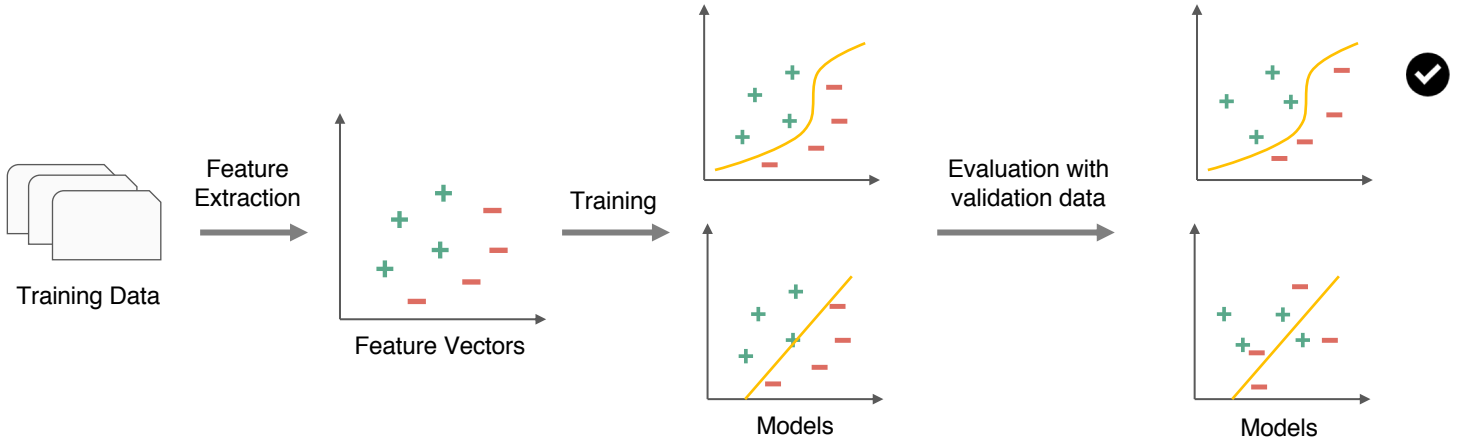
There are many ways to divide the tasks that make up the machine learning workflow into phases. No matter how you group them, the steps are the same.

Image Source: Mathworks

:: Typical machine learning tasks

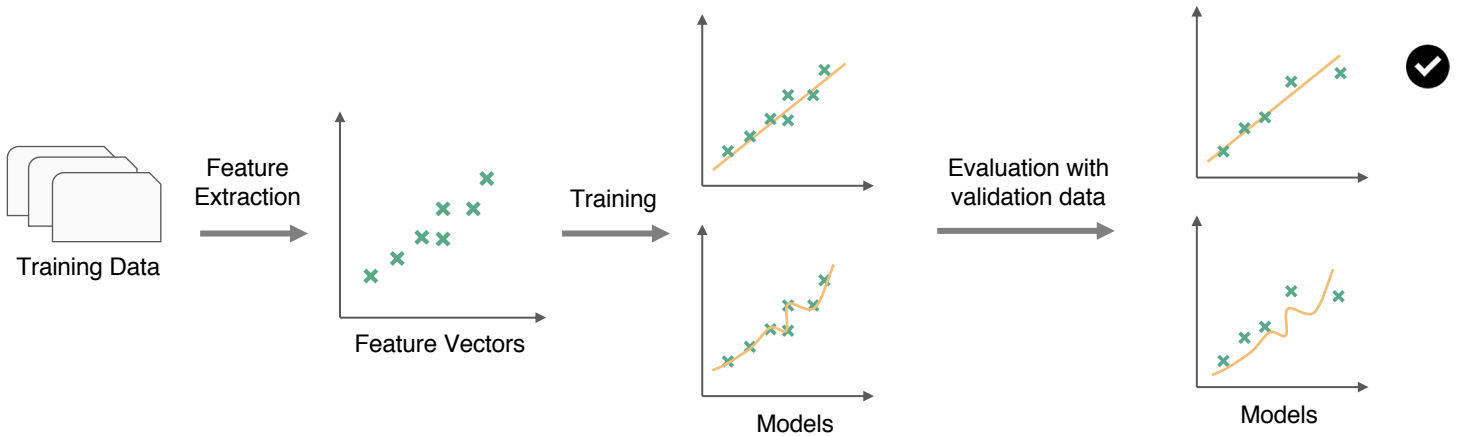
Classification (Predict the category)

Identifying to which category an object belongs to



Regression (Predict the value)

Predicting a continuous-valued attribute associated with an object



Source of original image: open.sap.com

:: Example - How to predict an answer with simple model

Example: predict the price of a diamond

Suppose I want to shop for a diamond. And you want to get an idea of how much it will cost. I take a notepad and pen into the jewelry store, and I write down the price of all of the diamonds in the case and how much they weigh in carats. Starting with the first diamond - it's 1.01 carats and $7,366. Now I go through and do this for all the other diamonds in the store.[1]

How much for the diamond ?

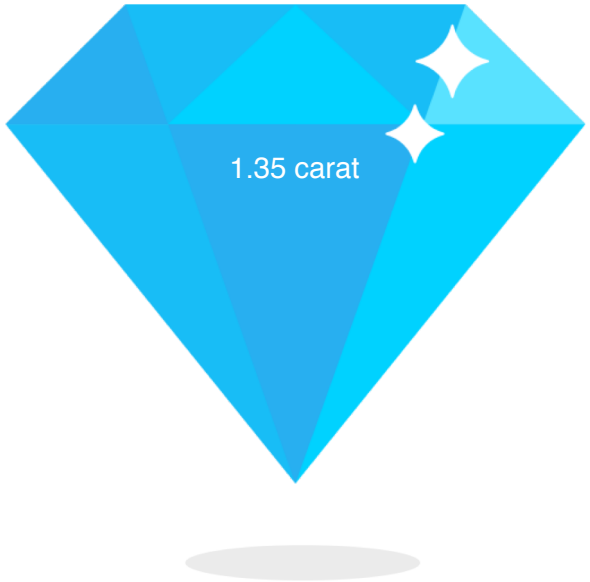


Image: Icon made by Freepik

1 Source: Microsoft, https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-for-beginners-predict-an-answer-with-a-simple-model

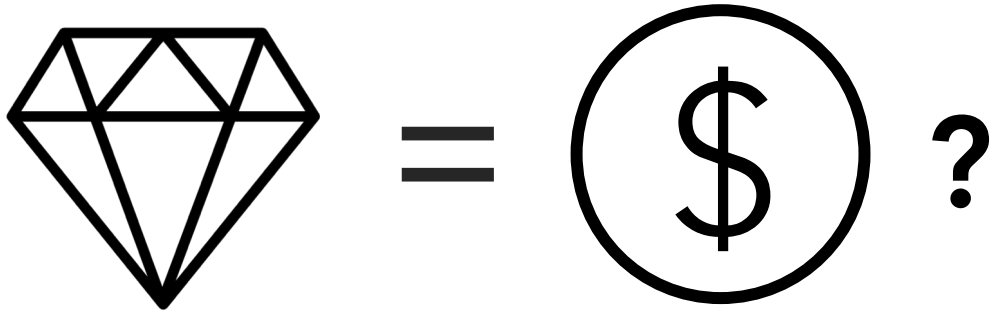
:: Example - How to predict an answer with simple model

Example: predict the price of a diamond *- continued*

Price of diamonds in jewelry store

Carats	Price ($)
1.01	7,366
0.49	985
0.31	544
1.51	9,140
0.37	493
0.73	3,011
1.53	11,413
0.56	1,814
0.41	876
0.74	2,690
0.63	1,190
0.6	4,172
2.06	11,764
1.1	4,682
1.31	6,172

“ How much will it cost to buy a **1.35** carat diamond ? ”



1.35 carat **How much?**

Our list doesn't have a 1.35 carat diamond in it, so we'll have to use the rest of our data to get an answer to the question.

Source: Microsoft, https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-for-beginners-predict-an-answer-with-a-simple-model

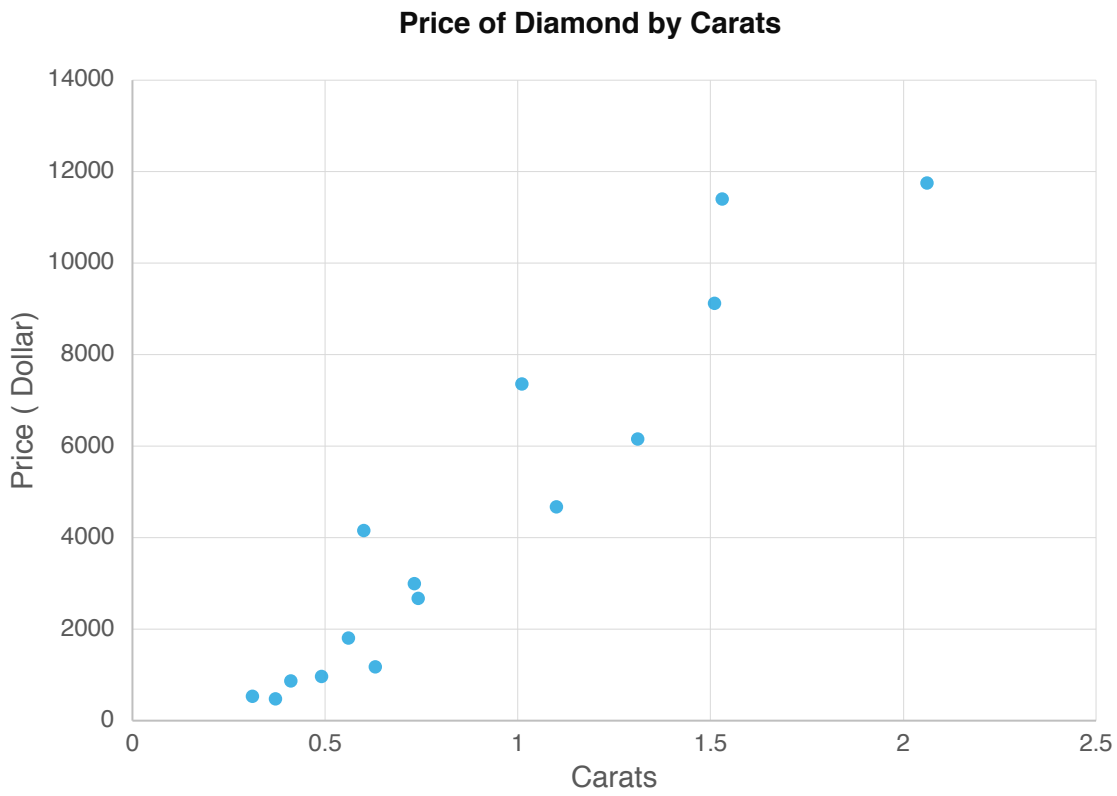
:: Example - How to predict an answer with simple model

Example: predict the price of a diamond *- continued*

Price of diamonds in jewelry store

Carats	Price ($)
1.01	7,366
0.49	985
0.31	544
1.51	9,140
0.37	493
0.73	3,011
1.53	11,413
0.56	1,814
0.41	876
0.74	2,690
0.63	1,190
0.6	4,172
2.06	11,764
1.1	4,682
1.31	6,172

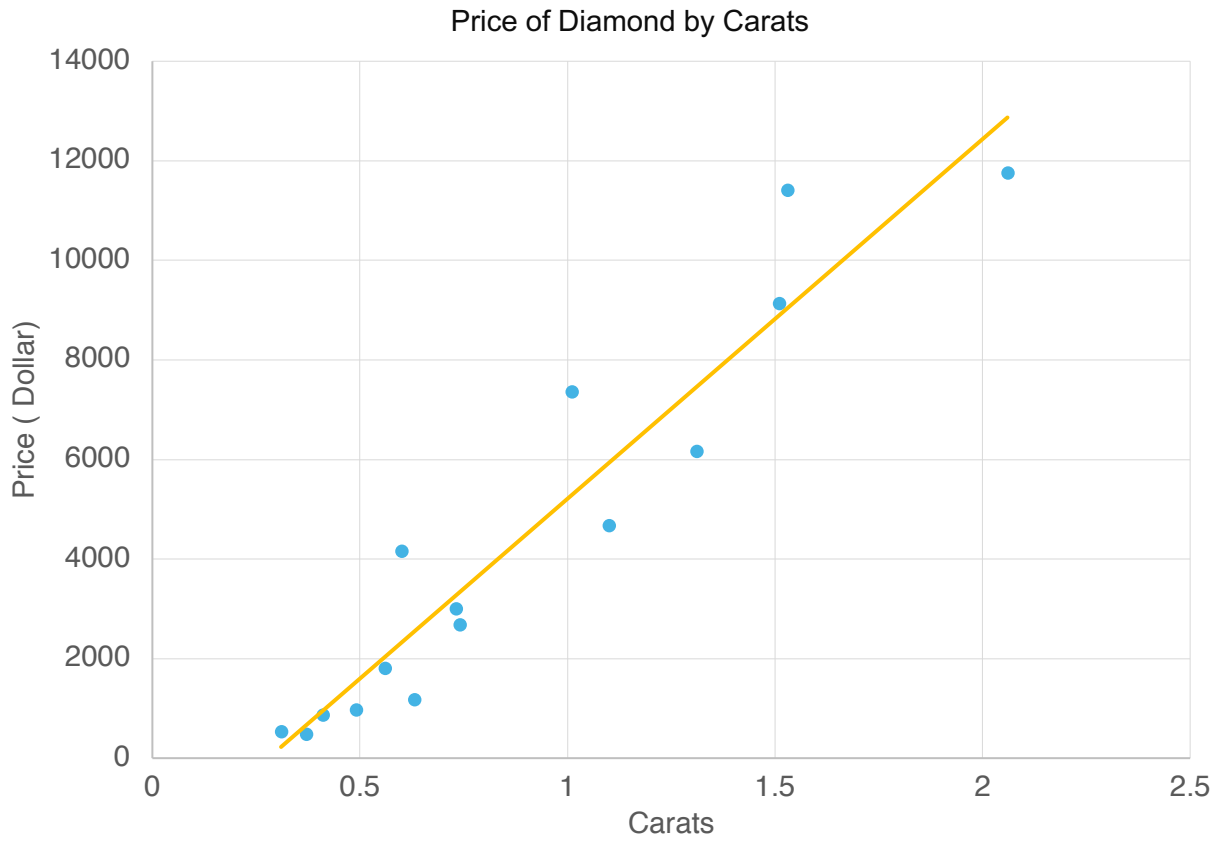
Plot the data



Source: Microsoft, https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-for-beginners-predict-an-answer-with-a-simple-model

:: Example - How to predict an answer with simple model

Predict the price of a diamond by Linear regression



Linear regression is an approach for modeling the relationship between a continuous dependent variable y and one or more predictors X

By drawing a line, we created a model

The fact that all the dots don't go exactly through the line is OK. Data scientists explain this by saying that there's the model - that's the line - and then each dot has some noise or variance associated with it.

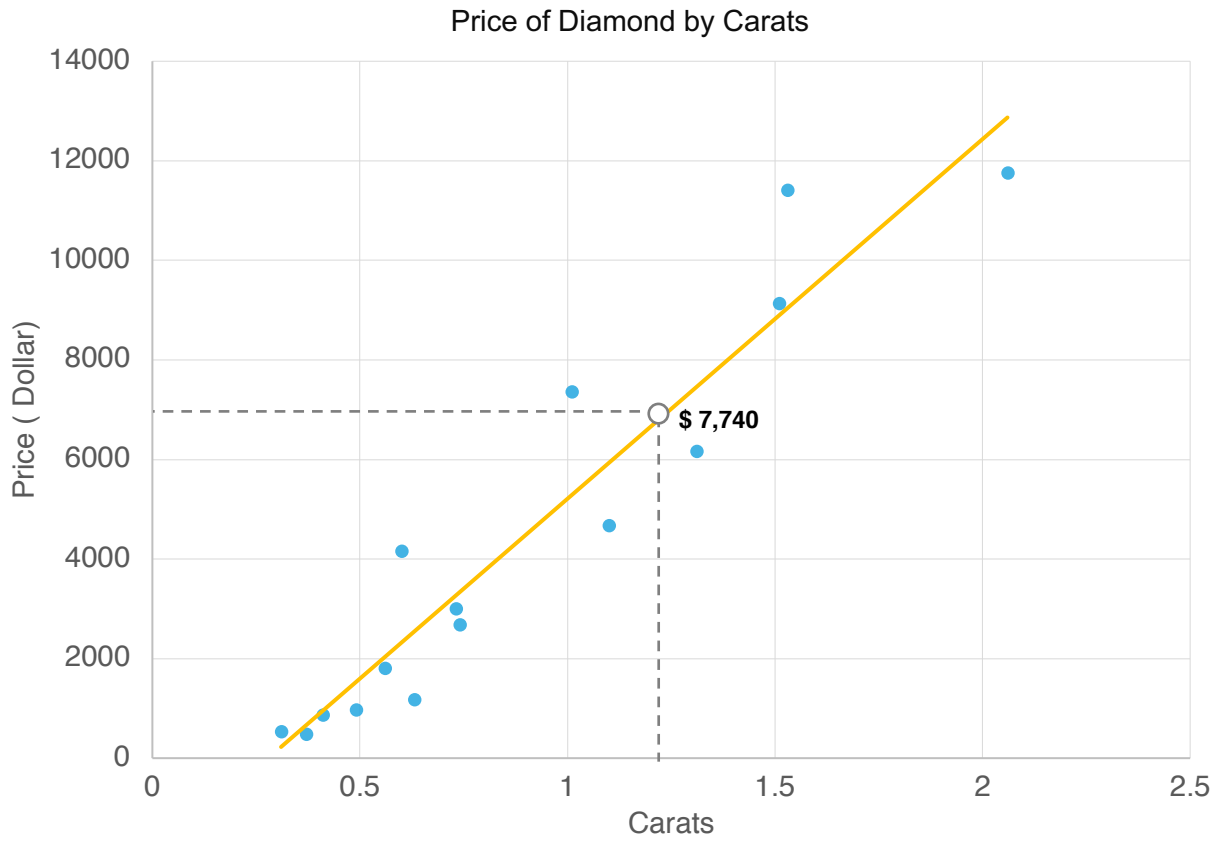
There's the underlying perfect relationship, and then there's the gritty, real world that adds noise and uncertainty.

Because we're trying to answer the question How much? this is called a regression. And because we're using a straight line, it's a linear regression.

Source: Microsoft, https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-for-beginners-predict-an-answer-with-a-simple-model

:: Example - How to predict an answer with simple model

Predict the price of a diamond by Linear regression



Linear regression is an approach for modeling the relationship between a continuous dependent variable y and one or more predictors X

Use the model to find the answer

Now we have a model and we ask it our question: How much will a 1.35 carat diamond cost?

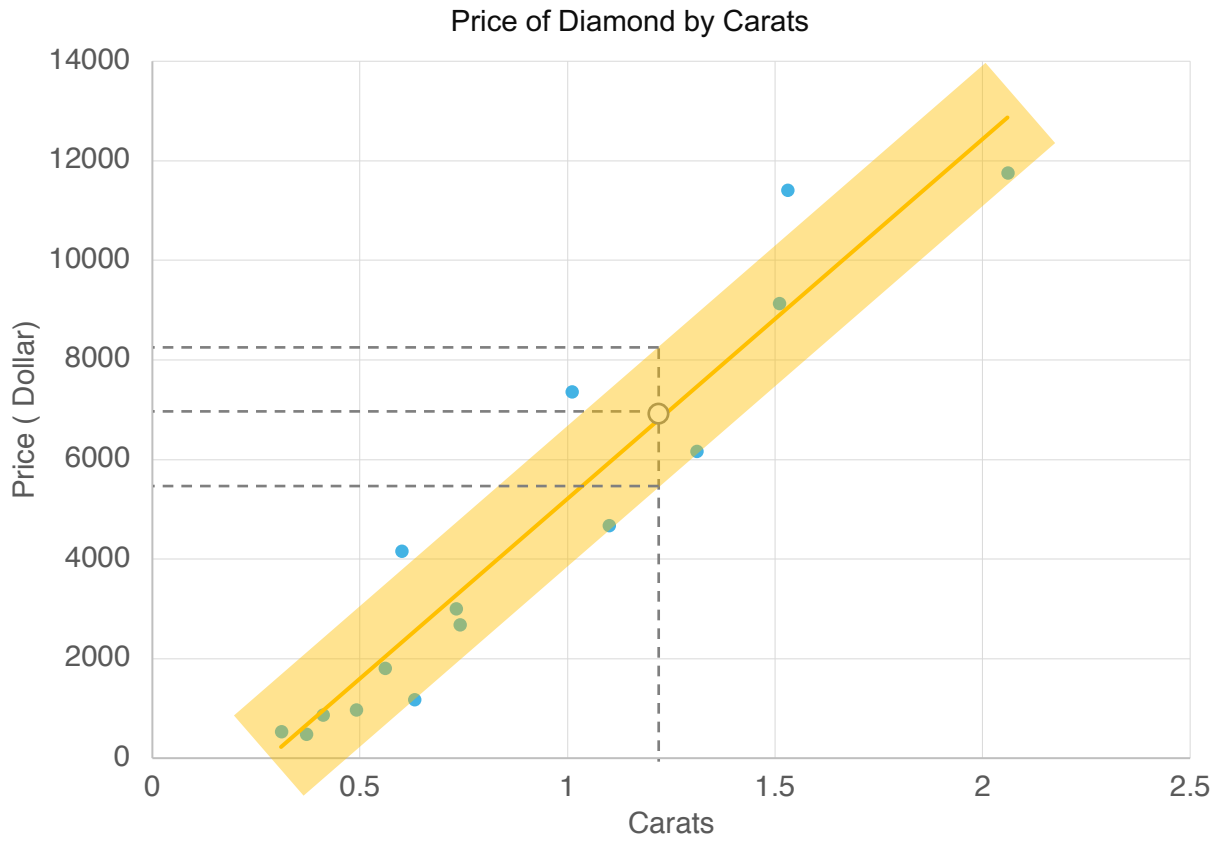
To answer our question, we eyeball 1.35 carats and draw a vertical line. Where it crosses the model line, we eyeball a horizontal line to the dollar axis. It hits right at $ 7740.

Boom! That's the answer: A 1.35 carat diamond costs about $7740.

Source: Microsoft, https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-for-beginners-predict-an-answer-with-a-simple-model

:: Example - How to predict an answer with simple model

Predict the price of a diamond by Linear regression



Linear regression is an approach for modeling the relationship between a continuous dependent variable y and one or more predictors X

Create a confidence interval

It's natural to wonder how precise this prediction is. It's useful to know whether the 1.35 carat diamond will be very close to $7400, or a lot higher or lower.

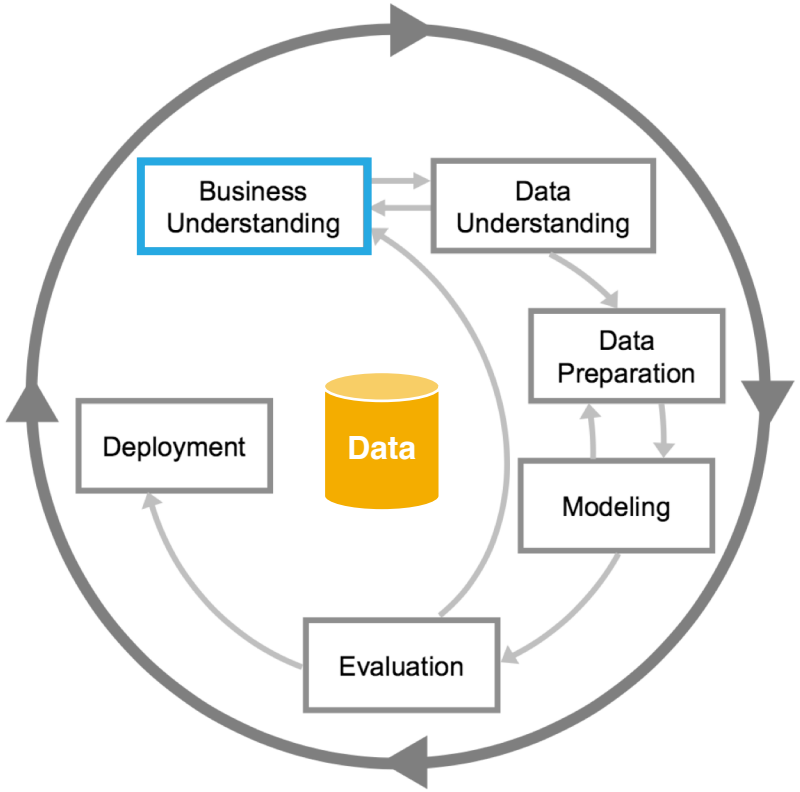
To figure this out, let's draw an envelope around the regression line that includes most of the dots. This envelope is called our confidence interval: We're pretty confident that prices fall within this envelope, because in the past most of them have. We can draw two more horizontal lines from where the 1.35 carat line crosses the top and the bottom of that envelope.

Now we can say something about our confidence interval: We can say confidently that the price of a 1.35 carat diamond is about $7,740 - but it might be as low as $6,300 and it might be as high as $9,000.

Source: Microsoft, https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-for-beginners-predict-an-answer-with-a-simple-model

02 | Business Understanding

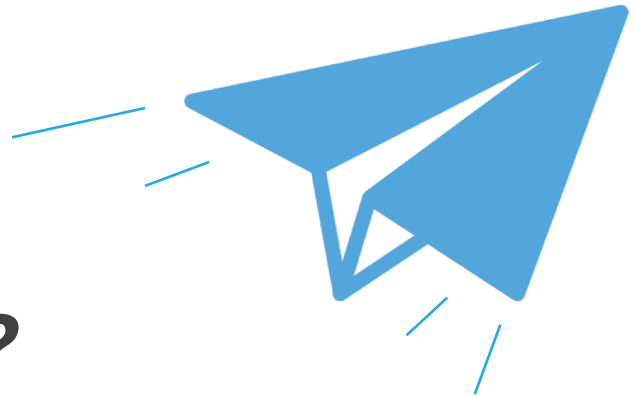
:: Understand the business



Get a clear understanding of the problem you're going to solve, how it impacts your organization, and your goals for addressing it. Tasks in this phase include: [1]

- Identifying your business goals
- Assessing your situation
- Defining your data mining goals
- Producing your project plan

Source: http://www.dummies.com/programming/big-data/data-science/phases-of-the-data-mining-process/



How to start ?

Ask a question you can answer with data

:: Ask a sharp question

A sharp question can be answered with a name or a number

- **Sharp Questions**

  Sharp questions can be answered with a name or a number.

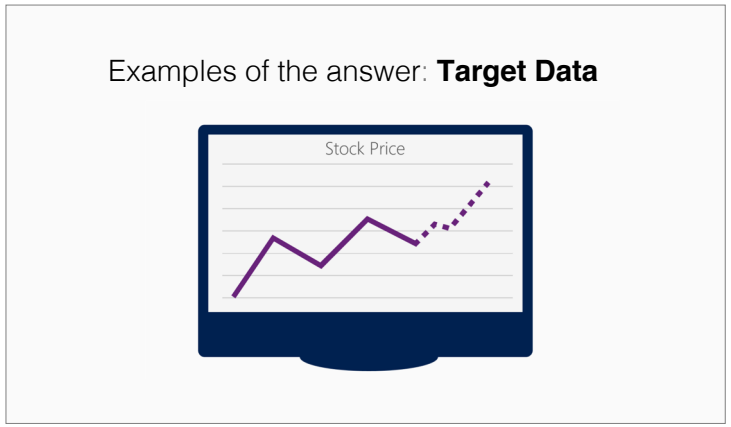
  - What will my stock's sale price be next week ?
  - Which car in my fleet is going to fail first ?

- **Vague Questions**

  Vague questions can't be answered with a name or a number.

  - How can I increase my profits ?
  - What can my data tell me about my business ?

Once you formulate your question, check to see whether you have examples of the answer in your data.



- These examples of answers are called a target. A target is what we are trying to predict about future data points, whether it's a category or a number.

- If you don't have any target data, you'll need to get some. You won't be able to answer your question without it.

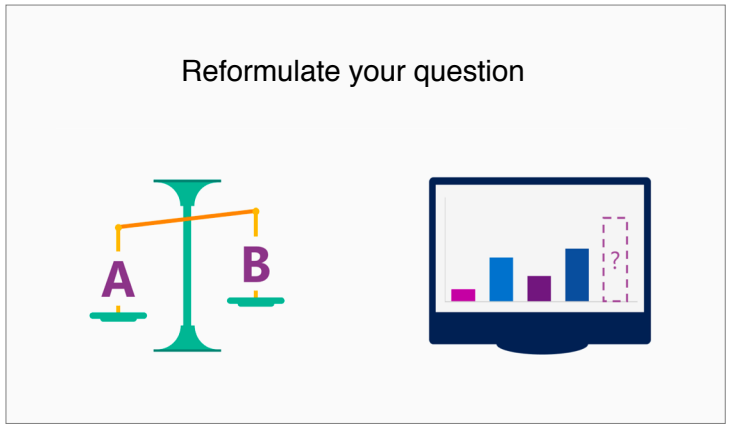
Source: Microsoft

:: Reformulate your question

how you ask a question is a clue to which algorithm can give you an answer

- **Reformulate your question**
  You can reformulate your question to use the algorithm that gives you the most useful answer.

Sometimes you can reword your question to get a more useful answer.



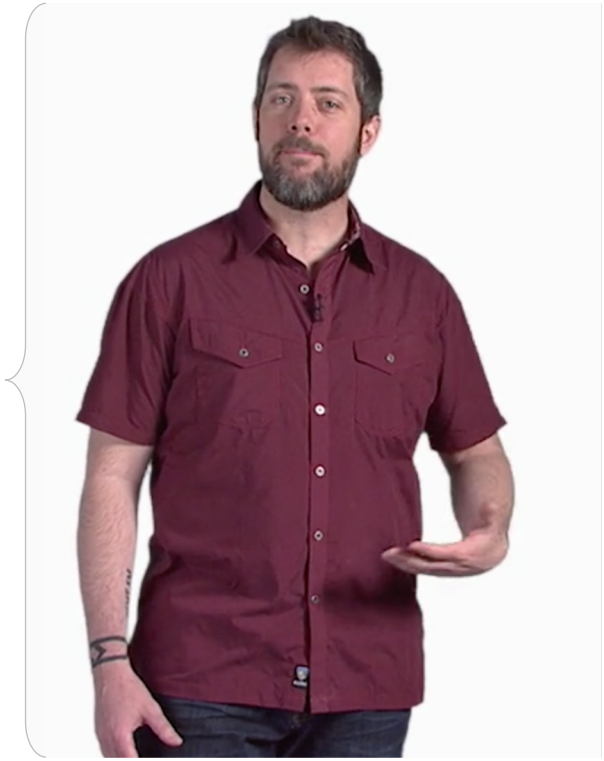
- The question "Is this data point A or B?" predicts the category (or name or label) of something. To answer it, we use a classification algorithm.
- The question “How much?” or “How many?” predicts an amount. To answer it, we use a regression algorithm

Source: Microsoft

:: The 5 questions data science can answer

It might surprise you, but there are only five questions that data science answers

- Is this A or B ?
- Is this weird ?
- How much – or – How many ?
- How is this organized ?
- What should I do next ?



Good Video (5 minutes)

https://azure.microsoft.com/en-us/documentation/videos/data-science-for-beginners-series-the-5-questions-data-science-answers/

Source: Microsoft, https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-for-beginners-the-5-questions-data-science-answers

:: Question 1: Is this A or B ?

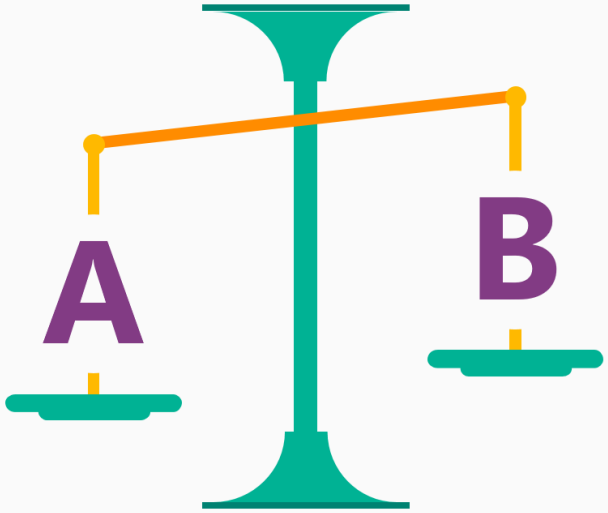
Is this A or B ?

Use Classification Algorithms

For example:

- Will this tire fail in the next 1,000 miles: Yes or no?

- Which brings in more customers: a $5 coupon or a 25% discount?



Source: Microsoft, https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-for-beginners-the-5-questions-data-science-answers

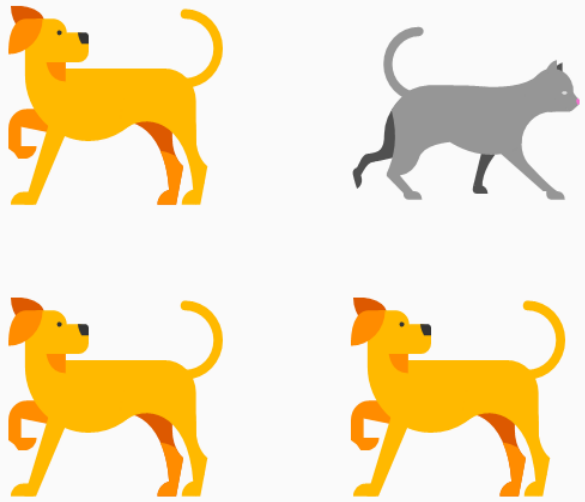
:: Question 2: Is this weird ?

Is this weird ?

Use Anomaly Detection Algorithms

Your credit card company analyzes your purchase patterns, so that they can alert you to possible fraud.

Charges that are "weird" might be a purchase at a store where you don't normally shop or buying an unusually pricey item.



Source: Microsoft, https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-for-beginners-the-5-questions-data-science-answers

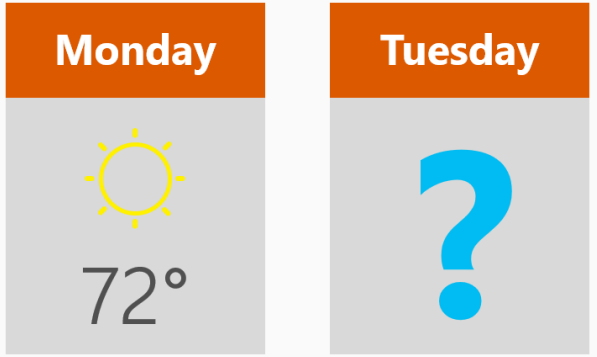
:: Question 3: How much ? or How many ?

How much ? How many ?

Use Regression Algorithms

Regression algorithms make numerical predictions, such as:

- What will the temperature be next Tuesday?
- What will my fourth quarter sales be?
  They help answer any question that asks for a number.



Source: Microsoft, https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-for-beginners-the-5-questions-data-science-answers

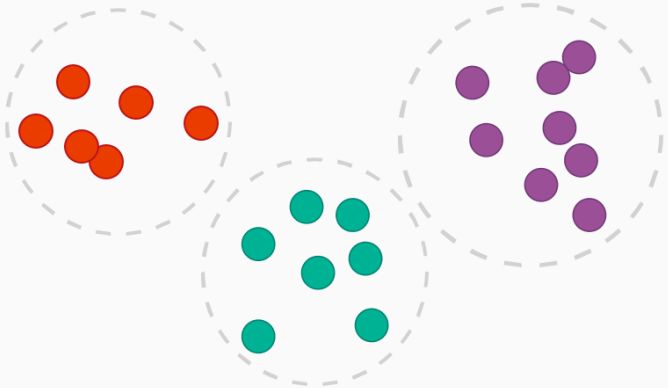
:: Question 4: How is this organized ?

How is this organized ?

Use Clustering Algorithms

Common examples of clustering questions are:

- Which viewers like the same types of movies?
- Which printer models fail the same way?



Sometimes you want to understand the structure of a data set - How is this organized? For this question, you don't have examples that you already know outcomes for. There are a lot of ways to tease out the structure of data. One approach is clustering. It separates data into natural "clumps," for easier interpretation. With clustering, there is no one right answer.

Source: Microsoft, https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-for-beginners-the-5-questions-data-science-answers

:: Question 5: What should I do now?

What should I do now ?

Use Reinforcement Learning Algorithms

Questions it answers are always about what action should be taken - usually by a machine or a robot. Examples are:

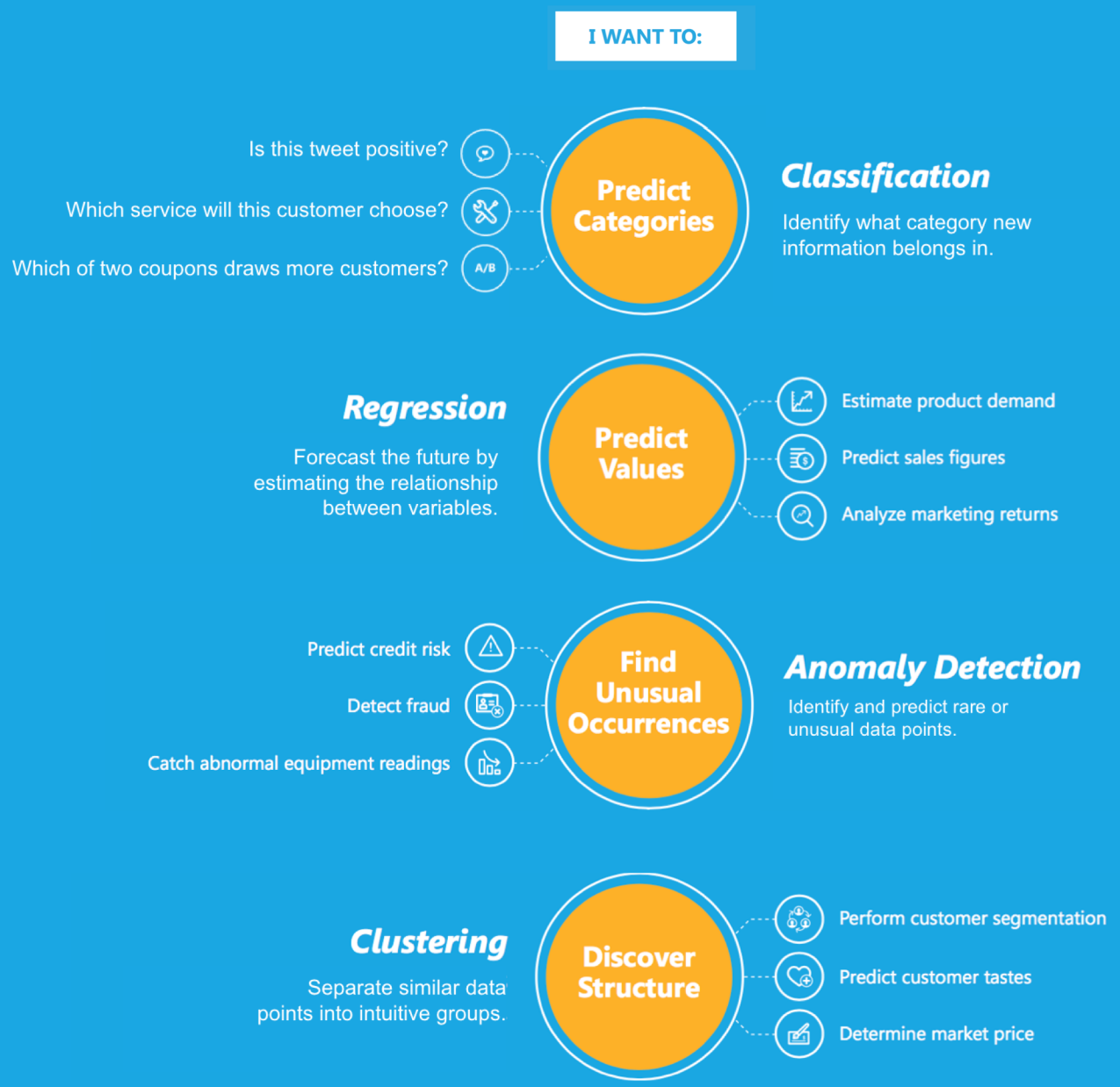
- If I'm a self-driving car: At a yellow light, brake or accelerate?
- For a robot vacuum: Keep vacuuming, or go back to the charging station

reinforcement learning algorithms learn from outcomes, and decide on the next action.

Typically, reinforcement learning is a good fit for automated systems that have to make lots of small decisions without human guidance.

Source: Microsoft, https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-for-beginners-the-5-questions-data-science-answers

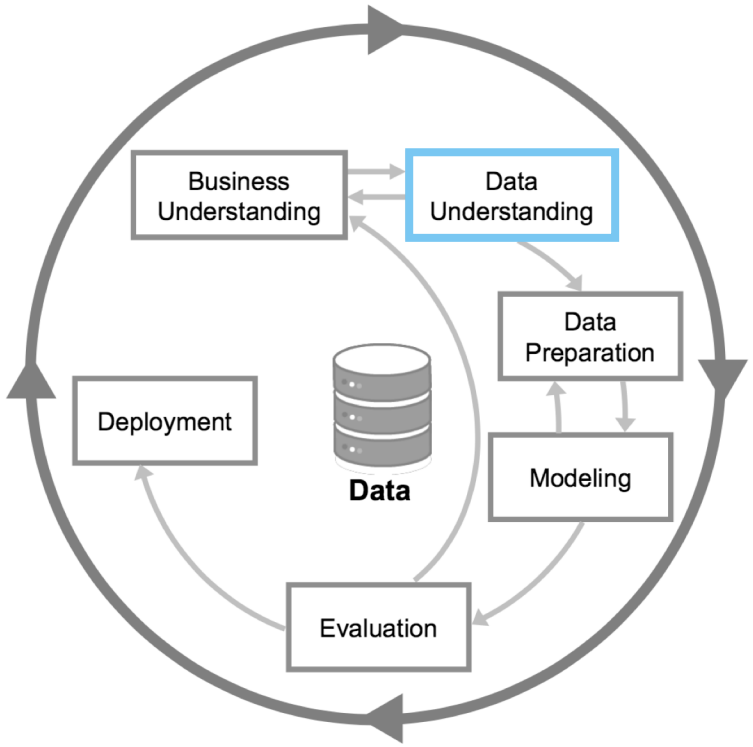
So, What do you want to find out?



http://azuremlsimpleds.azurewebsites.net/simpleds/

03 | Data Understanding

:: Data understanding



- **Collect Initial data**
  - Access and load the data
  - Joining data from multiple sources and rationalizing it into one dataset.
- **Describe data**
  - Descriptive statistics
  - The structure of the data
- **Explore data** (*Initial data analysis / Exploratory Data An*alysis)
  - Data exploration offers an early view into the data
  - Visualizing the data to look for patterns in the data
  - A number of data issues can be uncovered during this step
  - Possibly formulate hypotheses that could lead to new data collection and experiments.
- **Verify Data Quality**
  - errors, outliers, and missing observations

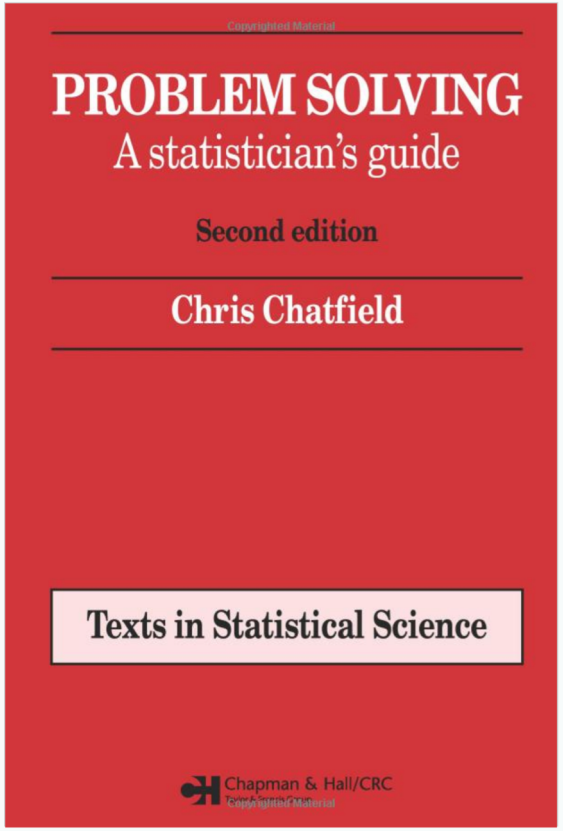
Source: https://www.the-modeling-agency.com/crisp-dm.pdf

:: Data understanding

Explore Data - Initial Data Analysis (IDA)

- **Initial data analysis (IDA) is an essential part of nearly every analysis. It includes analysis of:**
  - The structure of the data
  - The quality of the data
    - errors, outliers, and missing observations
  - Descriptive statistics
  - Graphs

- **The data are modified according to the analysis:**
  - Adjust extreme observations,
  - Estimate missing observations
  - Transform variables
  - Bin data
  - form new variables.



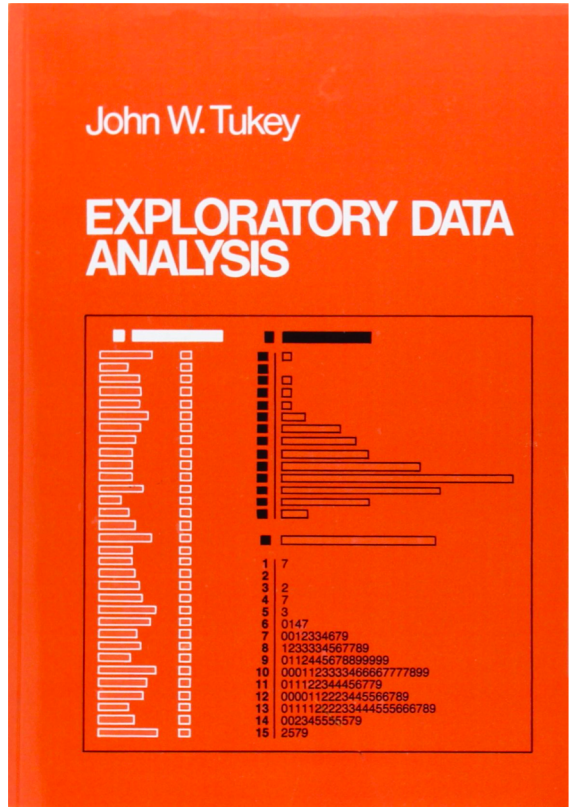
Source: open.SAP.com

:: Data understanding

Explore Data - Exploration Data Analysis (EDA)

- **Exploratory data analysis is an approach to analyzing data for the purpose of formulating hypotheses that are worth testing**
  - We often use data visualization techniques.
  - Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data and possibly formulate hypotheses that could lead to new data collection and experiments.
  - in many reference books, EDA now seems to encompass IDA

- **It is important to understand what you CAN DO before you learn to measure how WELL you seem to have done it**
  - *" To learn about data analysis, it is right that each of us try many things that do not work – that we tackle more problems than we make expert analyses of. We often learn less from an expertly done analysis than from one where, by not trying something, we missed an opportunity to learn more."*



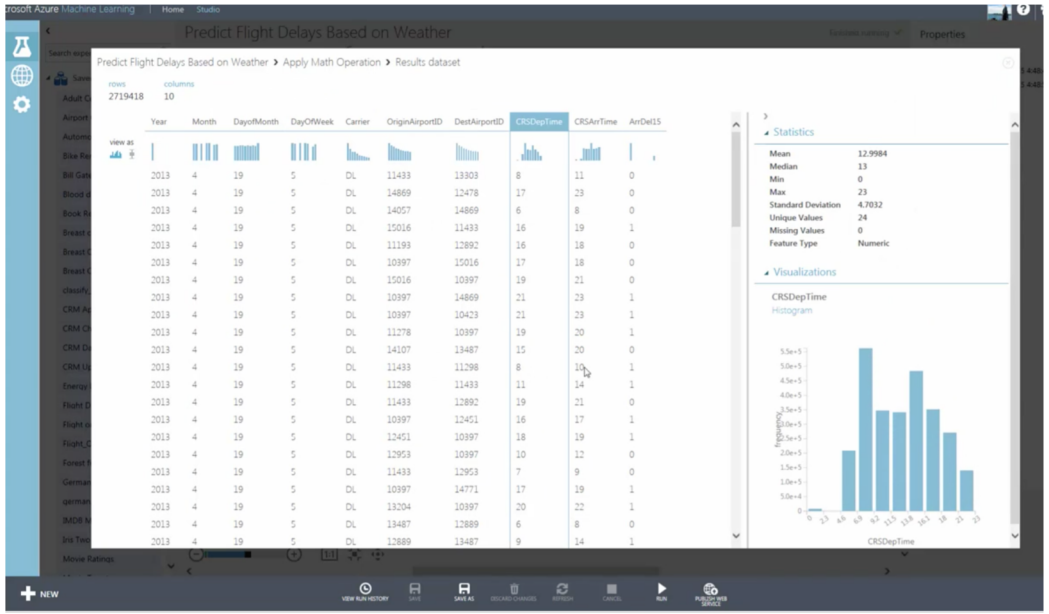
Source: open.SAP.com

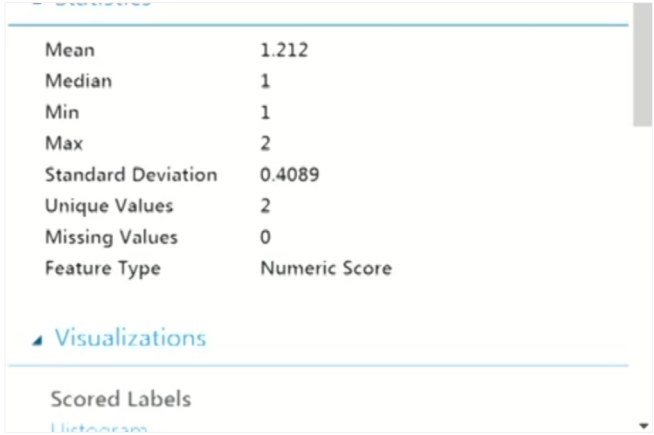
:: Data understanding

Describe the data

Examine the “gross” or “surface” properties of the acquired data and report on the results. The Statistical Summary Chart shows the distribution of each variable and provides descriptive statistics.

Example: Descriptive Statistics





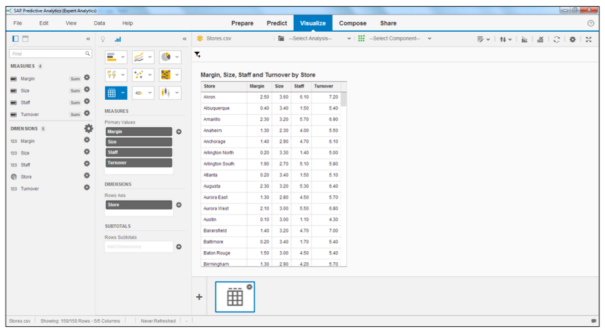
Data description report

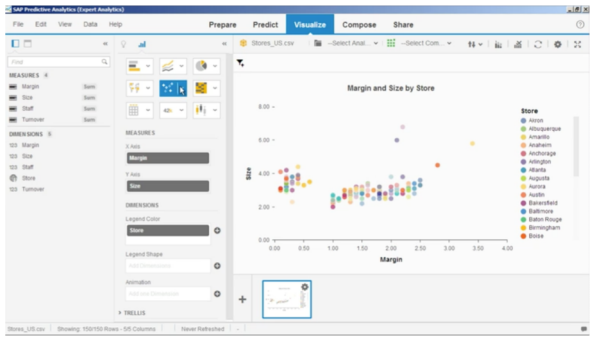
Describe the data that has been acquired including its format, its quantity (for example, the number of records and fields in each table), the identities of the fields and any other surface features which have been discovered. Evaluate whether the data acquired satisfies your requirements.

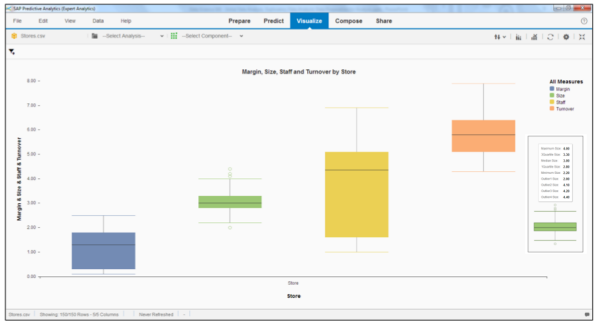
Source of image: Microsoft Azure Machine Learning

:: Data understanding

Example: Data Exploration and Visualization







- Box-plots can help identify outliers
- Density plots and histogram show the spread of the data
- Scatter plots can describe bivariate relationships

:: Data understanding

Let's take a closer look at the data

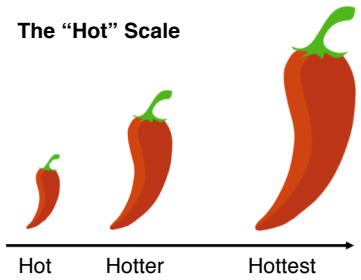
Categorical variable / Nominal variable[1]

- It's a discrete (categorical), qualitative variable that characterizes, describes, or names an element of a population.
- The order of the categories does not matter

ID	Account Name	Age	Gender	Annual Income	Membership	Satisfaction level
1	Jim Liang	43	Male	125,000	Gold	very dissatisfied
2	Steven Ive	35	Male	100,000	Gold	dissatisfied
3	Sherry Gao	38	Female	145,000	Silver	neutral
4	Peter Lorenz	31	Male	96,000	Gold	satisfied
5	Bill McCartney	25	Male	85,000	Bronze	very satisfied
6	John Carter	18	Male	234,000	Silver	very dissatisfied
7	Kelly Mills	19	Female	97,000	Bronze	very satisfied
8	Bono Sinead	57	Male	135,000	Gold	neutral
9	James Scott	44	Male	460,000	Silver	very satisfied
10	Jens Schneider	29	Male	150,000	Gold	satisfied

Ordinal variable [2]

- An ordinal variable is a discrete (categorical), qualitative variable that has order. For example, "Gold, Silver, Bronze"
- The order of the categories does matter



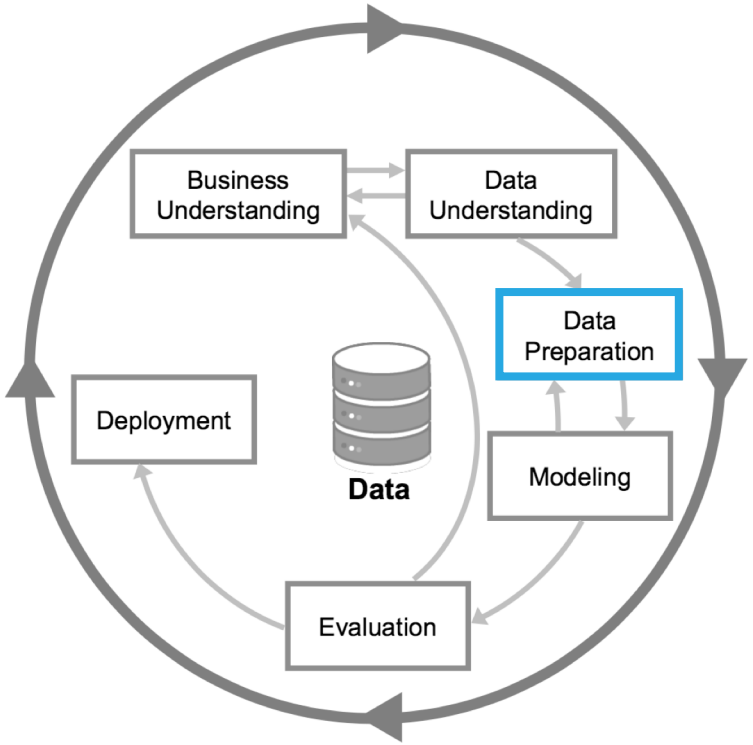
Continuous variable [3]

- It's is a quantitative variable.
- It is a real number that can take any value (with fractions/ decimal places) between two specific Numbers.
- It accommodates all basic arithmetic operations (addition, subtraction, multiplication, and division).

1,2, 3 & Image Source: SAP

04 | Data Preparation

:: Data Preparation

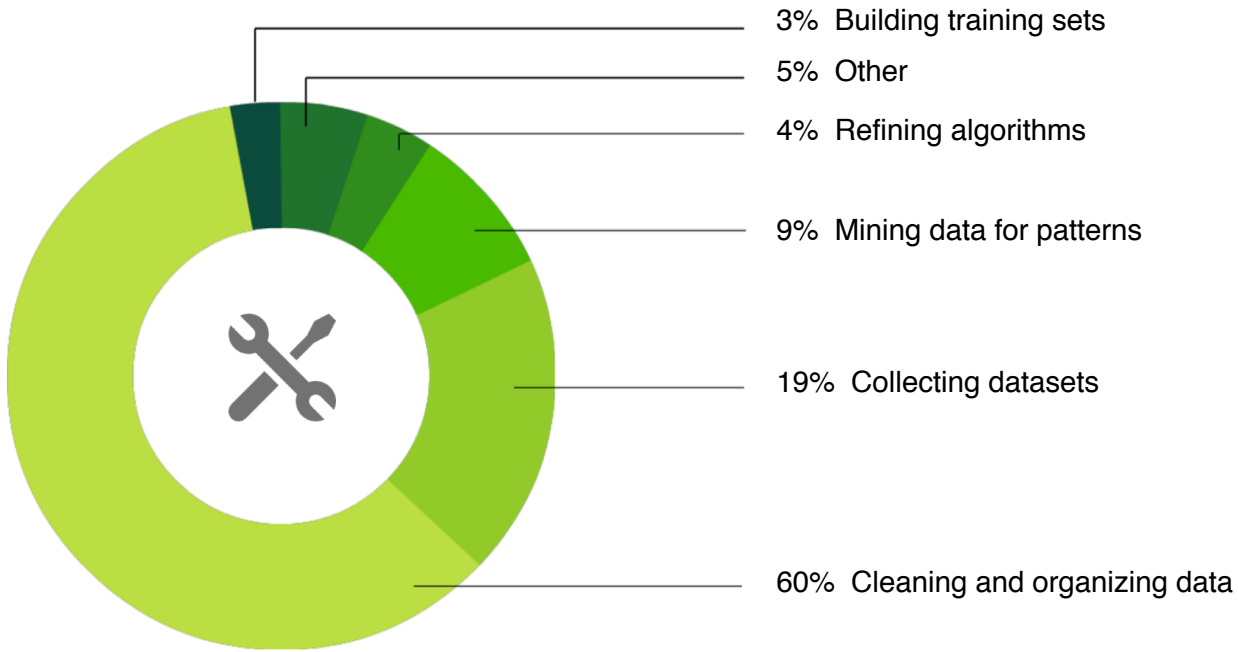


The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools..

Source: https://www.the-modeling-agency.com/crisp-dm.pdf

:: Preparing data is time-consuming

What data scientists spend the most time doing



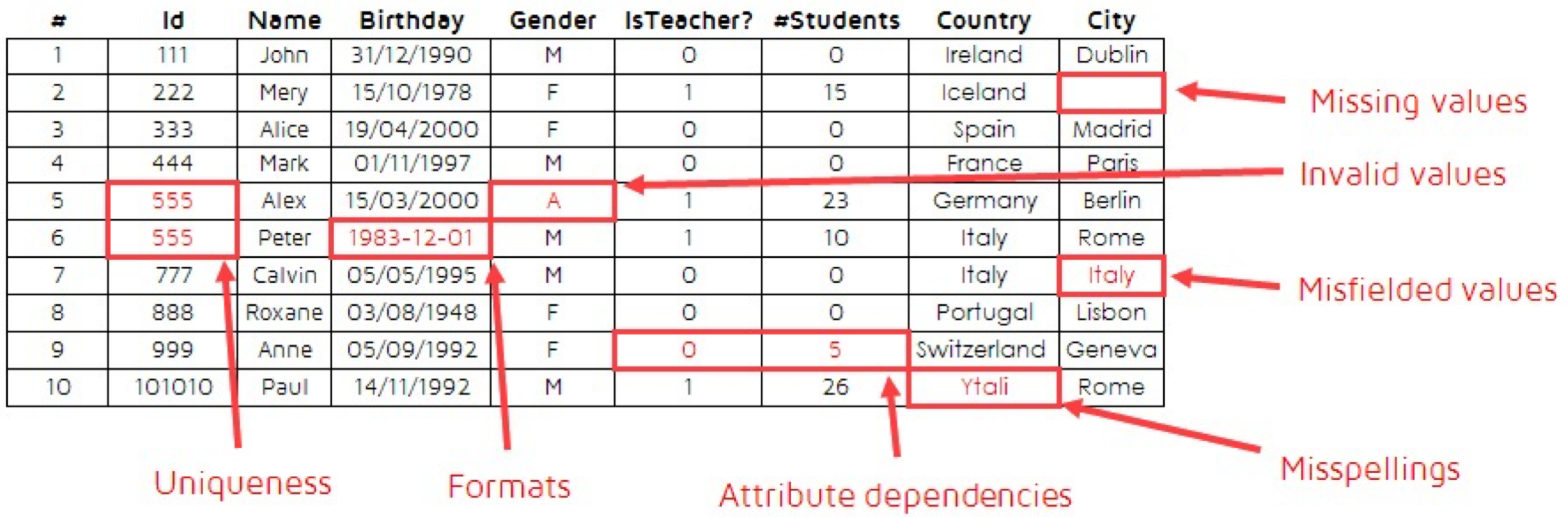
CrowdFlower Data Science Report 2016

*New York Times article reported that data scientists spend from **50% to 80%** of their time mired in the more mundane task of collecting and preparing unruly data before it can be explored for useful nuggets.*

Messy data is by far the most time-consuming aspect of the typical data scientist's work flow.

:: Data in the real world is dirty

Data is rarely clean and often you can have data quality issues



#	Id	Name	Birthday	Gender	IsTeacher?	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222	Mery	15/10/1978	F	1	15	Iceland	
3	333	Alice	19/04/2000	F	0	0	Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	555	Alex	15/03/2000	A	1	23	Germany	Berlin
6	555	Peter	1983-12-01	M	1	10	Italy	Rome
7	777	Calvin	05/05/1995	M	0	0	Italy	Italy
8	888	Roxane	03/08/1948	F	0	0	Portugal	Lisbon
9	999	Anne	05/09/1992	F	0	5	Switzerland	Geneva
10	101010	Paul	14/11/1992	M	1	26	Ytali	Rome

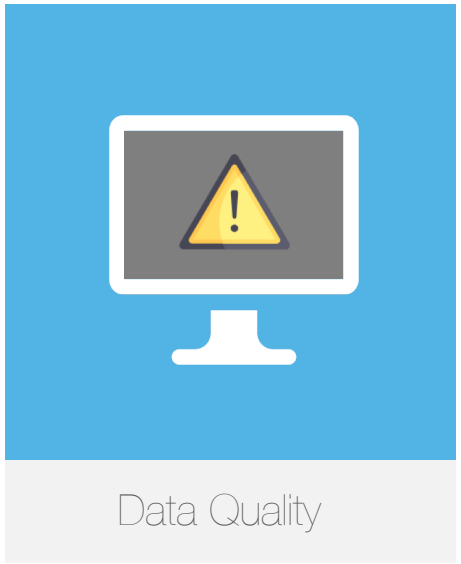
The typical data quality issues that arise are:

- **Incomplete**: Data lacks attributes or containing missing values.
- **Noisy**: Data contains erroneous records or outliers.
- **Inconsistent**: Data contains conflicting records or discrepancies.

Source: https://quantdare.com/data-cleansing-and-transformation/

:: Data in the real world is dirty

What kind of issues affect the quality of data?



- **Invalid values**

  Some datasets have well-known values, e.g. gender must only have “F” (Female) and “M” (Male). In this case it’s easy to detect wrong values.

- **Formats**

  The most common issue. It’s possible to get values in different formats like a name written as “Name, Surname” or “Surname, Name”.

- **Attribute dependencies**

  When the value of a feature depends on the value of another feature. For example, if we have some school data, the “number of students” is related to whether the person “is teacher?”. If someone is not a teacher he/she can’t have any students.

- **Uniqueness**

  It’s possible to find repeated data in features that only allow unique values. For example, we can’t have two products with the same identifier.

- **Missing values**

  Some features in the dataset may have blank or null values.

- **Misspellings**

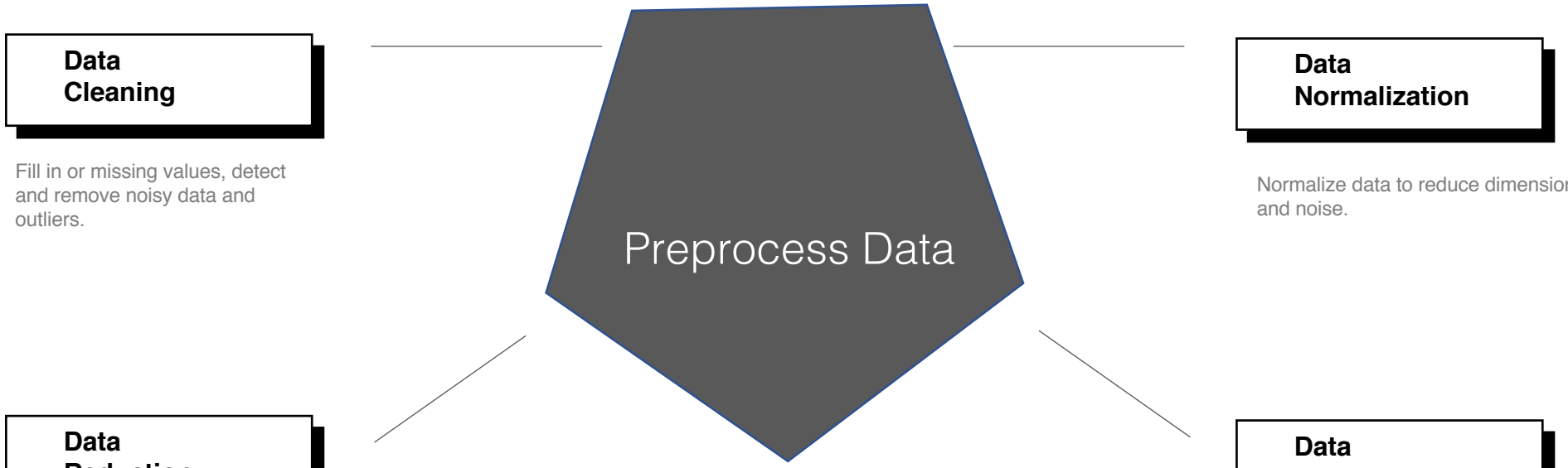
  Incorrectly written values.

- **Misfielded values**

  When a feature contains the values of another.

Source: https://quantdare.com/data-cleansing-and-transformation/
Icon Source: Icon made by Freepik

:: Preprocessing data to avoid "garbage in, garbage out"



Data Cleaning

Fill in or missing values, detect and remove noisy data and outliers.

Data Normalization

Normalize data to reduce dimensions and noise.

Data Reduction

Sample data records or attributes for easier data handling.

Data Discretization

Convert continuous attributes to categorical attributes for ease of use with certain machine learning methods.

Text Cleaning

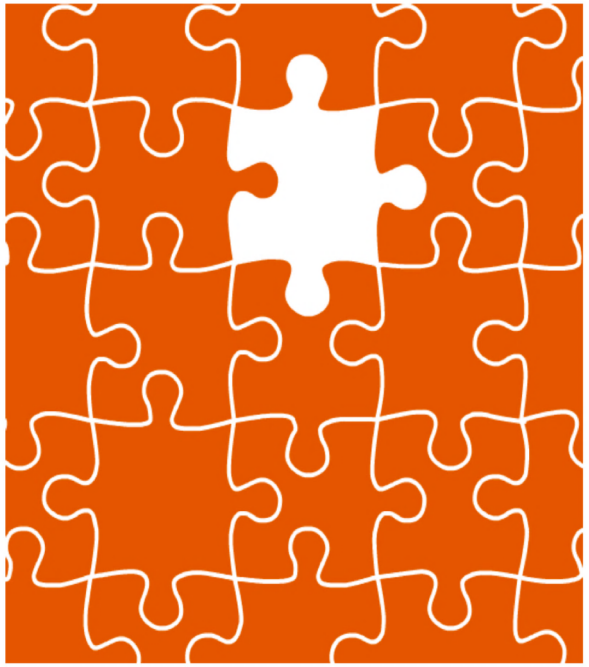
Remove embedded characters which may cause data misalignment, for e.g., embedded tabs in a tab-separated data file, embedded new lines which may break records, etc.

Source: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-prepare-data

:: Preprocessing data - Clean your data

Why to deal with missing values?

- Missing values in a dataset can be due to error or because observations that were not recorded
- When missing value are present, certain algorithms may not work or you may not have the desired result.
- Missing data affects some models more than others[1]
- Even for models that handle missing data, they can be sensitive to it (missing data for certain variables can result in poor predictions)[2]



Missing value is probably the most common problems in data mining/machine learning

1&2 Source: https://channel9.msdn.com/Events/OpenSourceTW/DevDays-Asia-2017/AI12

:: Preprocessing data - Clean your data

How to deal with missing values?

Typical missing value handling methods are:

- **Deletion**

  Remove records with missing values

- **Dummy substitution**

  Replace missing values with a dummy value: e.g, unknown for categorical or 0 for numerical values.

- **Mean substitution**

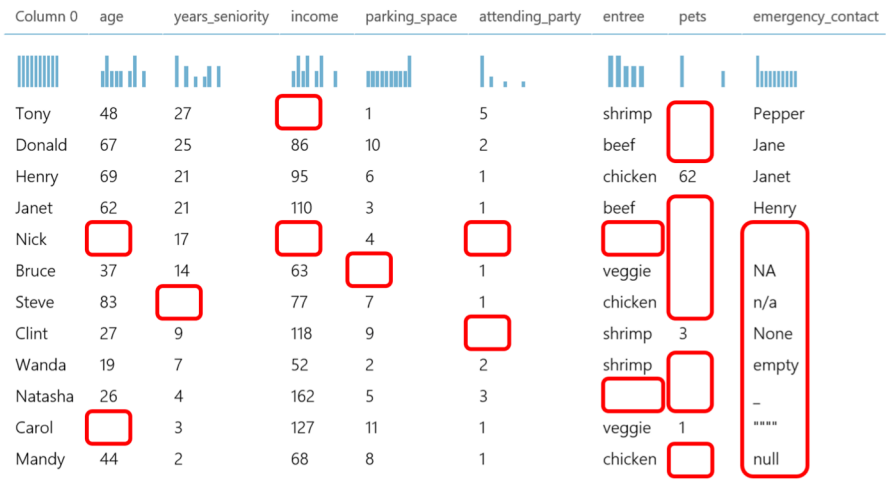
  If the missing data is numerical, replace the missing values with the mean.

- **Frequent substitution**

  If the missing data is categorical, replace the missing values with the most frequent item

- **Regression substitution**

  Use a regression method to replace missing values with regressed values.



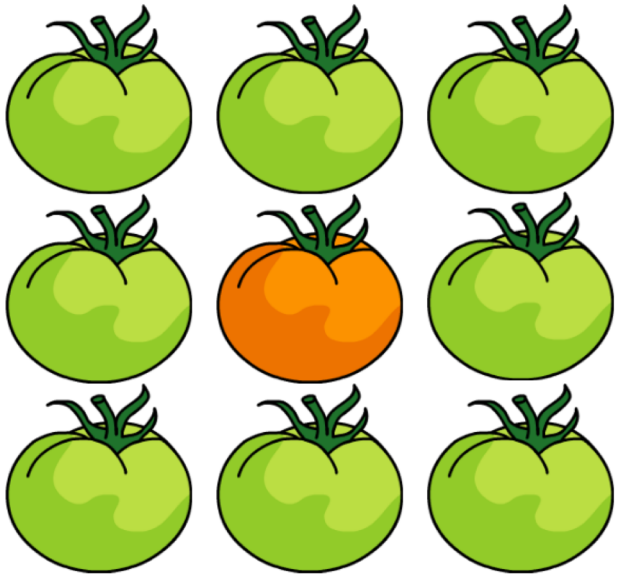
Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick		17		4				
Bruce	37	14	63		1	veggie		NA
Steve	83		77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			_
Carol		3	127	11	1	veggie	1	""""
Mandy	44	2	68	8	1	chicken		null

Source: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-prepare-data

:: Preprocessing data - Clean your data

What you should know about the outliers/anomalies

- Outliers may bring about problems by distorting the predictive model.
- What's an outlier is somewhat subjective.[1]
- Outliers can be very common in multidimensional data.[2]
- Some models are less sensitive (more robust) to outliers than others.[3]
- Outliers can be result of bad data collection, or they can legitimate extreme (or unusual) values.[4]
- Sometimes outliers are the interesting data points we want to model, and other times they just get in the way.[5]

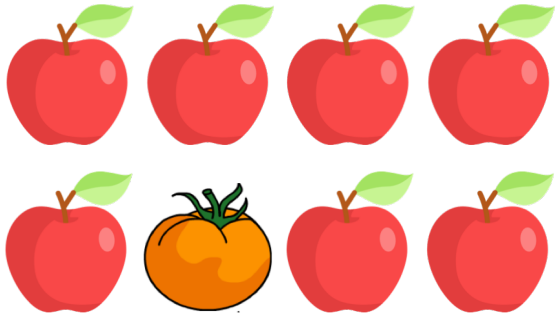


An outlier is a data point that distinctly separate from the rest of the data.

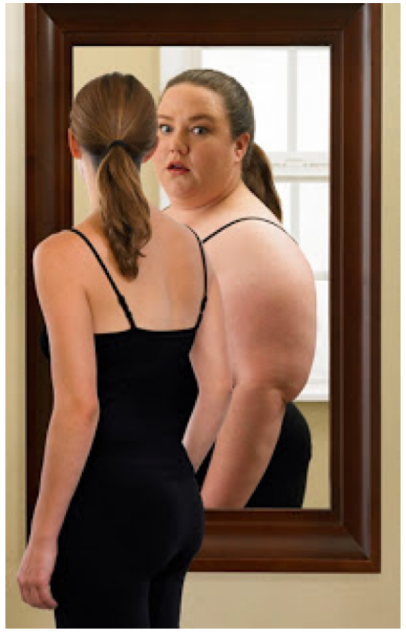
1,2,3,4 &5 Source: https://channel9.msdn.com/Events/OpenSourceTW/DevDays-Asia-2017/AI12

:: Preprocessing data - Clean your data

Cause of outliers



Data from different classes



Data measurement and collection Errors

Natural variation

Source: Overview of Outlier Detection Techniques and Applications by Ying, DFKI

:: Preprocessing data - Clean your data

How to deal with outliers?

The choice of how to deal with an outlier should depend on the cause.

- **Keep outliers**

  Outliers should not necessarily be omitted from the analysis as they may be genuine observations in the data.

  In many applications, outliers provide crucial information. For example, in a credit card fraud detection app, they indicated purchases that fall outside a customer's usual buying patterns.[1]

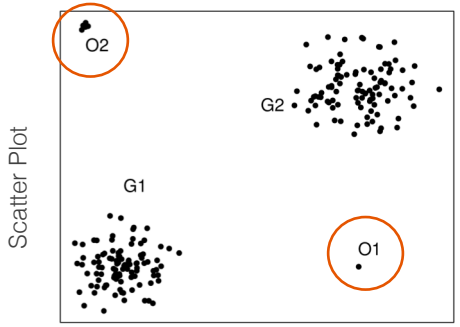
- **Exclude outliers**

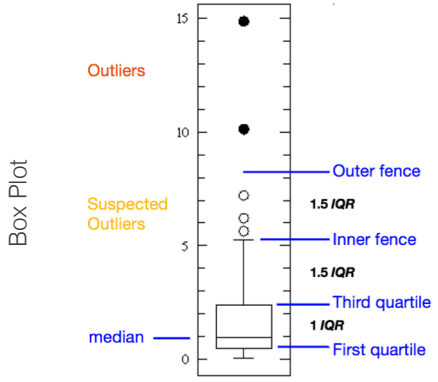
  There are two common approaches to exclude outliers [2]

  - Trimming/Truncation: Trimming discards the outliers
  - Winsorising: Winsorising replaces the outliers with the nearest "non-suspect" data.

Popular plots for outlier detection:

Scatter Plot & **Box Plot**





1 source : Mathworks

2 source: https://en.wikipedia.org/wiki/Outlier#Working_with_outliers

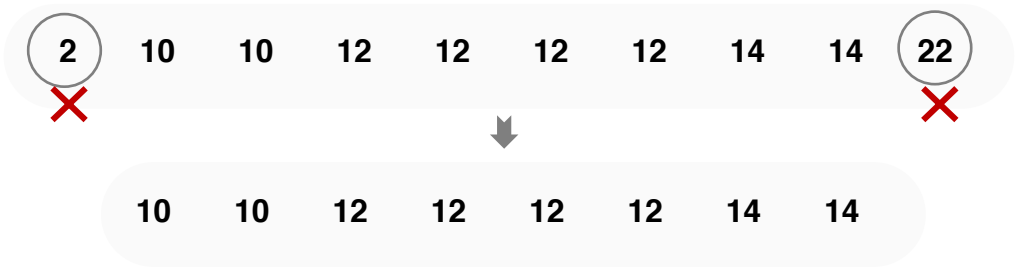
Box Plot's source: http://www.physics.csbsju.edu/stats/box2.html

:: Preprocessing data - Clean your data

Examples on how to deal with outliers

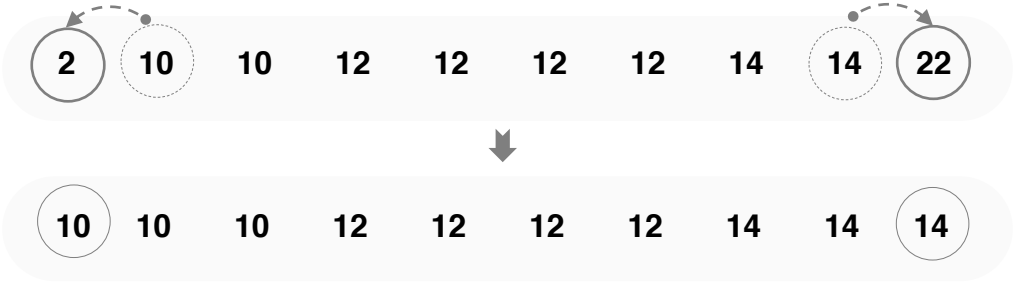
- **Example for Trimming**

  Eliminate the outliers “2” & “22”



- **Example for Winsorising**

  Assign outlier the next highest or lowest value found in the sample that is not an outlier. In this example, “10” & “14” are not outliers and used to replace the outliers “2” & “22”.



- **Trimming or Winsorizing less than 5% of data points**

  It will not likely affect the hypothesis testing outcome.

- **Trimming or Winsorizing great than 5% of data points**

  Trimming or Winsorizing great than 5% may affect the outcome results.

  - Reduce the power of analysis
  - Makes samples less representative
  - May affect normalcy of data
  - Consider transforming data, choosing an alternate outcome variable or data analysis technique.

Source : https://www.youtube.com/watch?v=ls-JHKUDL2k

:: Preprocessing data - Data normalization

How to normalize data?

Data normalization re-scales numerical values to a specified range. Popular data normalization methods include:

- **Min-Max Normalization**:

  Linearly transform the data to a range, say between 0 and 1, where the min value is scaled to 0 and max value to 1.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- **Z-score Normalization (** or **Standardization)**:

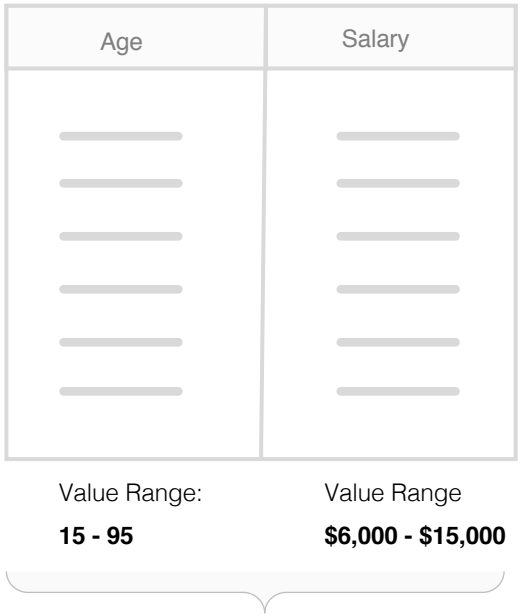
  Scale data based on mean and standard deviation: divide the difference between the data and the mean by the standard deviation. Feature standardization makes the values of each feature in the data have zero-mean (when subtracting the mean in the numerator) and unit-variance [1].

$$z = \frac{X - \mu}{\sigma}$$

$$\text{Z-Score} = \frac{\text{Value} - \text{mean}}{\text{Standard Deviation}}$$

- **Decimal scaling**:

  Scale the data by moving the decimal point of the attribute value.



Different scales

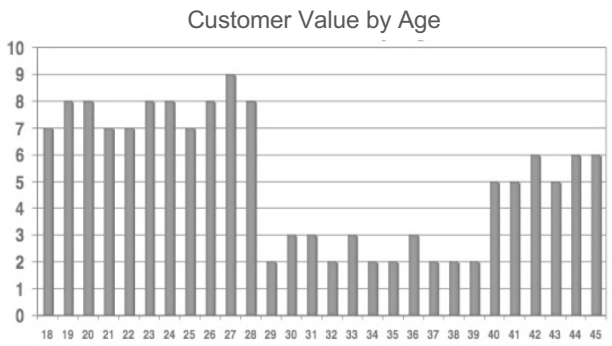
Different scales in a dataset may be problematic in some cases where certain machine algorithms require data to be in the same scale.

Source: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-prepare-data
1 Source: https://en.wikipedia.org/wiki/Feature_scaling

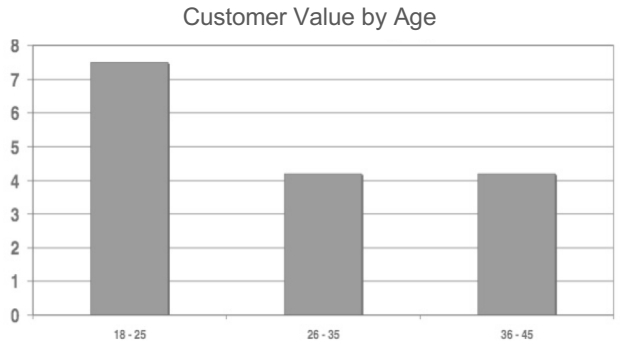
:: Preprocessing data - Data discretization

How to discretize data?

A numeric variable may have many different values and for some algorithms this may lead to very complex models. You can convert continuous attributes by “binning” to categorical attributes for ease of use with certain machine learning methods.



Raw data – No binning



Manual binning

Discretization is the process of putting values into buckets so that there are a limited number of possible states. The buckets themselves are treated as ordered and discrete values. You can discretize both numeric and string columns[1].

Binning

- Binning helps to improve model performance. It captures non-linear behavior of continuous variables.
- It minimizes the impact of outliers. It removes “noise” from large numbers of distinct values
- It makes the models more explainable – grouped values are easier to display and understand. It improves model build speed – predictive algorithms build much faster as the number of distinct values decreases.

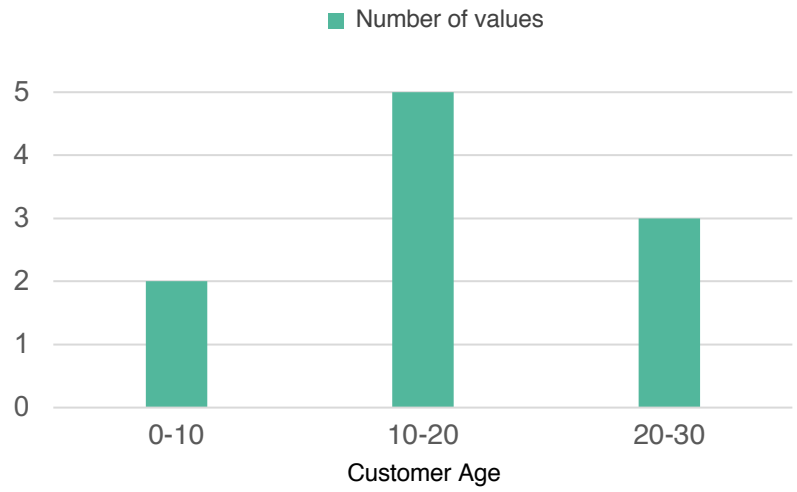
1 source: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-prepare-data 2 Source: https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/discretization-methods-data-mining

:: Preprocessing data - Data discretization

How to discretize data? - continued

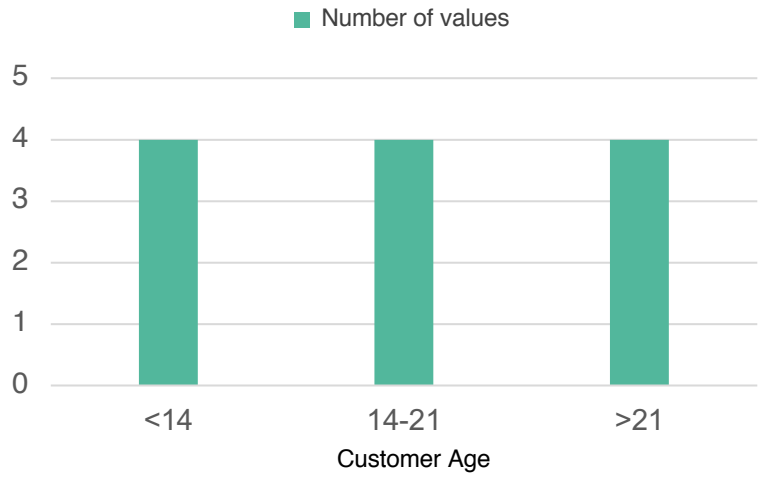
Data can be discretized by converting continuous values to categorical attributes or intervals. Some ways of doing this are:

- **Equal-Width Binning (by distance)**



Divide the range of all possible values of an attribute into N groups of the same size, and assign the values that fall in a bin with the bin number[1].

- **Equal-Height Binning (by frequency)**



Divide the range of all possible values of an attribute into N groups, each containing the same number of instances, then assign the values that fall in a bin with the bin number [2].

1 & 2 Source: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-prepare-data

:: Preprocessing data - Data reduction

How to reduce data ?

There are various methods to reduce data size for easier data handling. Depending on data size and the domain, the following methods can be applied:

- **Record Sampling**
  Sample the data records and only choose the representative subset from the data.

- **Attribute Sampling**
  Select only a subset of the most important attributes from the data.

- **Aggregation**
  Divide the data into groups and store the numbers for each group. For example, the daily revenue numbers of a restaurant chain over the past 20 years can be aggregated to monthly revenue to reduce the size of the data.

Sample Data

If the dataset you plan to analyze is large, it's usually a good idea to down-sample the data to reduce it to a smaller but representative and more manageable size. This facilitates data understanding, exploration, and feature engineering.

More data can result in much longer running times for algorithms and larger computational and memory requirements. You can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole dataset.

:: Preprocessing data - Text cleaning

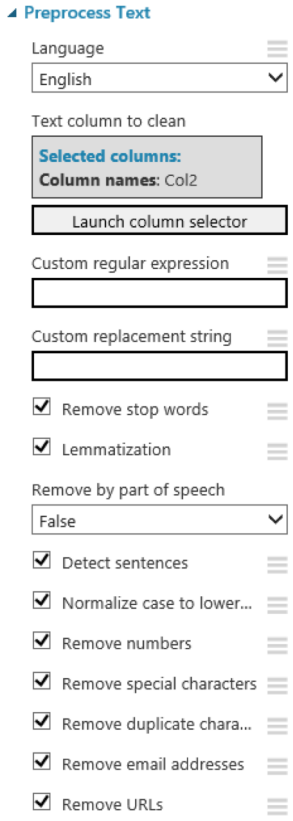
How to clean text data ?

Improper text encoding handling while writing/reading text leads to information loss, inadvertent introduction of unreadable characters, e.g., nulls, and may also affect text parsing.

Unstructured text such as tweets, product reviews, or search queries usually requires some preprocessing before it can be analyzed.

For example：

- replacing special characters and punctuation marks with spaces
- normalizing case
- removing duplicate characters
- removing user-defined or built-in stop-words
- word stemming



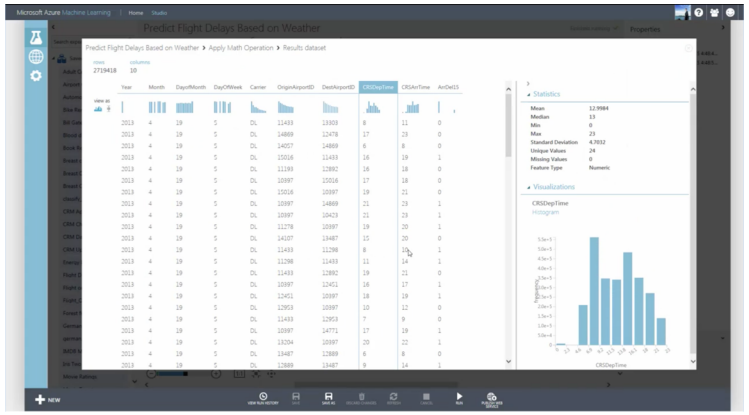
Example for Cleaning and preprocessing text dataset in Azure Machine Learning Studio

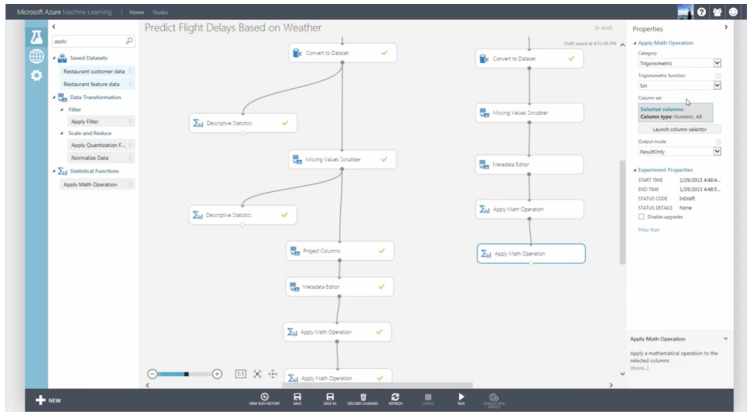
We remove stopwords - common words such as "the" or "a" - and numbers, special characters, duplicated characters, email addresses, and URLs. We also convert the text to lowercase, lemmatize the words, and detect sentence boundaries that are then indicated by "III" symbol in pre-processed text.

Source: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-text-analytics-module-tutorial

:: Preprocessing data (Demo)

Video: Preprocessing Data in Azure Machine Learning Studio (10 minutes)





https://azure.microsoft.com/en-us/resources/videos/preprocessing-data-in-azure-ml-studio/

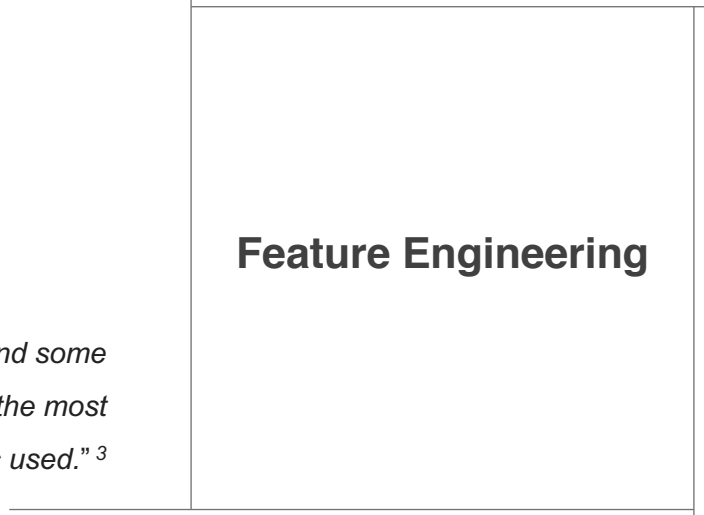
:: Feature Engineering

Feature Engineering is the key task in machine learning

There is no formal definition of feature engineering. It means different things to different people. In Google's definition, the process of extracting features from raw data is called feature engineering. In Microsoft's documentation feature engineering is more about feature construction.

" Coming up with features is difficult, time-consuming, requires expert knowledge. "Applied machine learning" is basically feature engineering. " [1]

- Andrew Ng (吴恩达)



"…some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used." [3]

" The algorithms we used are very standard for Kagglers. We spent most of our efforts in feature engineering. We were also very careful to discard features likely to expose us to the risk of over-fitting our model." [2]

1 source: Andrew Ng, Machine Learning and AI via Brain simulations
2 source: Xavier Conort, "Q&A with Xavier Conort"
3 source: Pedro Domingos, A Few Useful Things to Know about Machine Learning

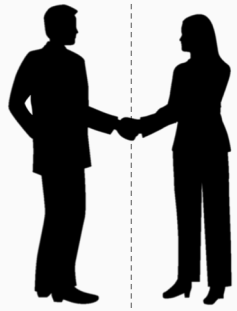
:: Feature Engineering

Feature Engineering is a sort of art

Feature engineers requires a creative combination of domain expertise and insights obtained from the data exploration step.

This is a balancing act of finding and including informative variables while avoiding too many unrelated variables.

Informative variables improve our result; unrelated variables introduce unnecessary noise into the model [2].



The more a data scientist interacts with the domain expert, the better the feature engineering process[1]

1Source:https://assets.cdn.sap.com/sapcom/docs/2015/09/2a16f496-3f7c-0010-82c7-eda71af511fa.pdf
2 source : Microsoft

:: What's feature ?

Features (input)

Also known as **Attributes**、**Explanatory Variables**、**Independent Variables, Predictors**

Target (output)

Also known as **Dependent Variables**

Example

	Bedrooms	Sq. feet	Neighborhood	Sales price
Training Data	3	2000	Normaltown	$250,000
	2	800	Hipesteron	$300,000
	2	850	Normaltown	$150,000
	1	550	Normaltown	$78,000
	4	2000	Skid Row	$150,00
	3	2000	Hipesteron	$220,00
Test Data	3	900	Hipesteron	$890,00
	4	2100	Normaltown	$270,000
New Data	2	450	Hipesteron	?
	4	1950	Normaltown	?

unlabeled examples

This unknown value should be predicted by the model

a **feature** is an individual measurable property or characteristic of a phenomenon being observed[1].

In this example, there are a couple of features like bedrooms, house size, neighborhood which are used as input for the modeling. The target is the sales price of the house.

There are some known price for some houses (Training data & Test Data).

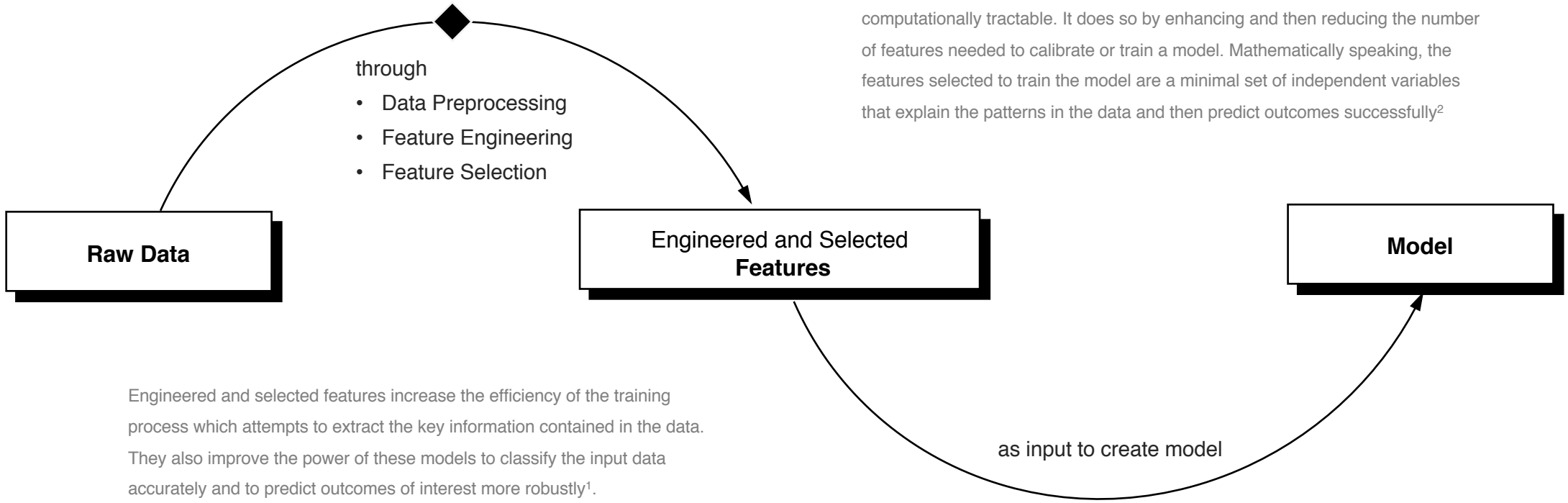
A model can be built to predict the the sales price of the houses which are still unknown.

In a nutshell, once we've trained our model with labeled examples, we use that model to predict the label on unlabeled examples.[2]

1 Source : Wikipedia

2. https://developers.google.com/machine-learning/crash-course/framing/ml-terminology

:: Feature Engineering can augment your data

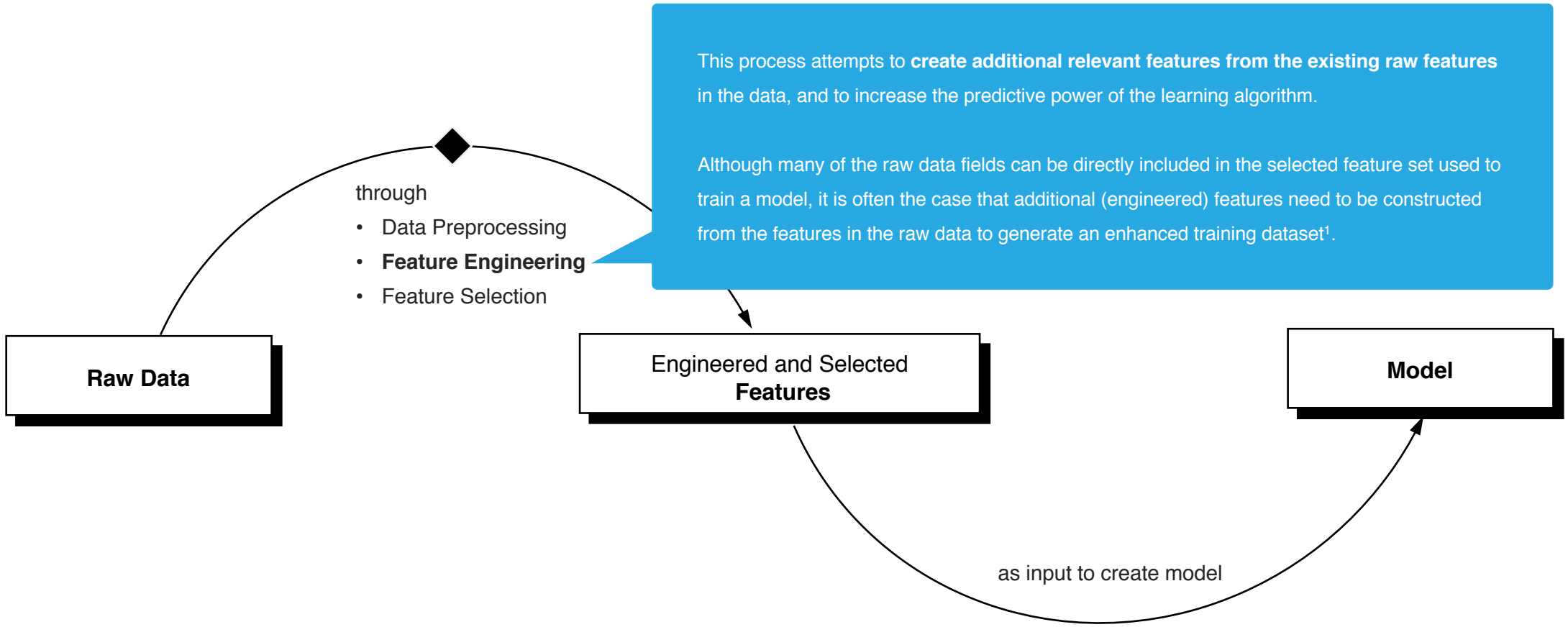


Feature engineering and selection can also combine to make the learning more computationally tractable. It does so by enhancing and then reducing the number of features needed to calibrate or train a model. Mathematically speaking, the features selected to train the model are a minimal set of independent variables that explain the patterns in the data and then predict outcomes successfully[2]

Engineered and selected features increase the efficiency of the training process which attempts to extract the key information contained in the data. They also improve the power of these models to classify the input data accurately and to predict outcomes of interest more robustly[1].

1 & 2 Source: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-create-features

:: Feature Engineering can augment your data



Note: There is no formal definition of Feature Engineering. Someone include feature selection, even data preprocessing as a part of Feature Engineering.

- *In Google's definition, Feature engineering means transforming raw data into a feature vector*
- *In Microsoft's documentation feature engineering mainly focuses on feature construction.*

1 Source: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-create-features

:: Feature Engineering - example

Example 1

Additional features

Emp_code	Gender	Date	New_Day	New_Month	New_Year
A001	Male	21-Sep-11	21	9	2011
A002	Female	27-Feb-13	27	2	2013
A003	Female	14-Nov-12	14	11	2012
A004	Male	07-Apr-13	7	4	2013
A005	Female	21-Jan-11	21	1	2011
A006	Male	26-Apr-13	26	4	2013
A007	Male	15-Mar-12	15	3	2012

Emp_code	Gender	Var_Male	Var_Female
A001	Male	1	0
A002	Female	0	1
A003	Female	0	1
A004	Male	1	0
A005	Female	0	1
A006	Male	1	0
A007	Male	1	0

Additional features

Feature / Variable creation

Feature creation is a process to generate a new variables / features based on existing variable(s).

For example, say, we have date(dd-mm-yy) as an input variable in a data set. We can generate new variables like day, month, year, week, weekday that may have better relationship with target variable. This step is used to highlight the hidden relationship in a variable:

Creating dummy variables

One of the most common application of dummy variable is to convert categorical variable into numerical variables. Dummy variables are also called Indicator Variables. It is useful to take categorical variable as a predictor in statistical models. Categorical variable can take values 0 and 1.

Let's take a variable 'gender'. We can produce two variables, namely, "Var_Male" with values 1 (Male) and 0 (No male) and "Var_Female" with values 1 (Female) and 0 (No Female). We can also create dummy variables for more than two classes of a categorical variables with n or n-1 dummy variables.

source: https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/

:: Feature Engineering - example

- Date features: year, month, week of month, etc.
- Time features
- Season features
- Weekday-and-weekend features
- Holiday features: New Year, U.S. Labor Day, U.S. Thanksgiving, Cyber Monday, Christmas, etc.
- Fourier features to capture seasonality

Example 2

Raw Data

Retail Forecasting: Step 3 of 6, feature engineering › Execute R Script › Result

rows 2475 columns 5

ID1	ID2	time	value	RDPI
1	2	2010-11-06T00:00:00Z		11080.5
1	2	2010-11-13T00:00:00Z	130	11114.7
1	2	2010-11-20T00:00:00Z	222	11114.7
1	2	2010-11-27T00:00:00Z	166	11114.7
1	2	2010-12-04T00:00:00Z	174	11114.7
1	2	2010-12-11T00:00:00Z	236	11101.2
1	2	2010-12-18T00:00:00Z	350	11101.2
1	2	2010-12-25T00:00:00Z	216	11101.2
1	2	2011-01-01T00:00:00Z	230	11101.2
1	2	2011-01-08T00:00:00Z	268	11128.3
1	2	2011-01-15T00:00:00Z	332	11128.3
1	2	2011-01-22T00:00:00Z	280	11128.3
1	2	2011-01-29T00:00:00Z	214	11128.3
1	2	2011-02-05T00:00:00Z	220	11128.3
1	2	2011-02-12T00:00:00Z	276	11160.8
1	2	2011-02-19T00:00:00Z	218	11160.8
1	2	2011-02-26T00:00:00Z	226	11160.8
1	2	2011-03-05T00:00:00Z	246	11160.8
1	2	2011-03-12T00:00:00Z	340	11239
1	2	2011-03-19T00:00:00Z	212	11239
1	2	2011-03-26T00:00:00Z	300	11239
1	2	2011-04-02T00:00:00Z	214	11239
1	2	2011-04-09T00:00:00Z	348	11297.4
1	2	2011-04-16T00:00:00Z	390	11297.4
1	2	2011-04-23T00:00:00Z	346	11297.4
1	2	2011-04-30T00:00:00Z	402	11297.4
1	2	2011-05-07T00:00:00Z	306	11297.4
1	2	2011-05-14T00:00:00Z	318	11329
1	2	2011-05-21T00:00:00Z	286	11329

Retail Forecasting: Step 3 of 6, feature engineering › Execute R Script › Result Dataset

Additional features are created through feature engineering

rows 2475 columns 22

ID1	ID2	time	value	RDPI	year	month	weekofmonth	weekofyear	USNewYearsDay	USLaborDay	USThanksgivingDay	CyberMonday	ChristmasDay	FreqCos1	FreqSin1	FreqCos2	FreqSin2
1	2	2010-11-06T00:00:00Z		11080.5	2010	11	1	45	false	false	false	false	false	1	0	1	0
1	2	2010-11-13T00:00:00Z	130	11114.7	2010	11	2	46	false	false	false	false	false	0.992709	0.120537	0.970942	0.239316
1	2	2010-11-20T00:00:00Z	222	11114.7	2010	11	3	47	false	false	false	false	false	0.970942	0.239316	0.885456	0.464723
1	2	2010-11-27T00:00:00Z	166	11114.7	2010	11	4	48	false	false	true	false	false	0.935016	0.354605	0.748511	0.663123
1	2	2010-12-04T00:00:00Z	174	11114.7	2010	12	1	49	false	false	false	true	false	0.885456	0.464723	0.568065	0.822984
1	2	2010-12-11T00:00:00Z	236	11101.2	2010	12	2	50	false	false	false	false	false	0.822984	0.568065	0.354605	0.935016
1	2	2010-12-18T00:00:00Z	350	11101.2	2010	12	3	51	false	false	false	false	false	0.748511	0.663123	0.120537	0.992709
1	2	2010-12-25T00:00:00Z	216	11101.2	2010	12	4	52	false	false	false	false	true	0.663123	0.748511	-0.120537	0.992709
1	2	2011-01-01T00:00:00Z	230	11101.2	2011	1	1	1	true	false	false	false	false	0.568065	0.822984	-0.354605	0.935016
1	2	2011-01-08T00:00:00Z	268	11128.3	2011	1	2	2	false	false	false	false	false	0.464723	0.885456	-0.568065	0.822984
1	2	2011-01-15T00:00:00Z	332	11128.3	2011	1	3	3	false	false	false	false	false	0.354605	0.935016	-0.748511	0.663123
1	2	2011-01-22T00:00:00Z	280	11128.3	2011	1	4	4	false	false	false	false	false	0.239316	0.970942	-0.885456	0.464723
1	2	2011-01-29T00:00:00Z	214	11128.3	2011	1	5	5	false	false	false	false	false	0.120537	0.992709	-0.970942	0.239316
1	2	2011-02-05T00:00:00Z	220	11128.3	2011	2	1	6	false	false	false	false	false	0	1	-1	0
1	2	2011-02-12T00:00:00Z	276	11160.8	2011	2	2	7	false	false	false	false	false	-0.120537	0.992709	-0.970942	-0.239316
1	2	2011-02-19T00:00:00Z	218	11160.8	2011	2	3	8	false	false	false	false	false	-0.239316	0.970942	-0.885456	-0.464723
1	2	2011-02-26T00:00:00Z	226	11160.8	2011	2	4	9	false	false	false	false	false	-0.354605	0.935016	-0.748511	-0.663123
1	2	2011-03-05T00:00:00Z	246	11160.8	2011	3	1	10	false	false	false	false	false	-0.464723	0.885456	-0.568065	-0.822984
1	2	2011-03-12T00:00:00Z	340	11239	2011	3	2	11	false	false	false	false	false	-0.568065	0.822984	-0.354605	-0.935016

The template provides two ID fields, which she associates with the store ID and product/SKU ID.

Source: https://gallery.cortanaintelligence.com/Experiment/Retail-Forecasting-Step-3-of-6-feature-engineering-2

:: Feature Engineering - example

Example 3

Raw Data

amazon review › Book Reviews from Amazon › dataset

rows 10000 columns 2

Col1	Col2
2	Ok~ but I think the Keirsey Temperment Test is more accurate - and cheaper. This book has its good points. If anything, it helps you put into words what you want from a supervisor, but it is not very accurate. The online test does not account for a difference between when 2 of their options are bo...
2	Repellent Sale of Conservativism The fatalistic view of Conservativism is expressed by Kirk even before the first chapter concludes as he states, " [Edmund] Burke, could he see our century, never would concede that a consumption-society, so near suicide, is the end for which Providence has prepared ...
1	I had a bad feeling about this! And I was right! I was intrigued by the title, which supposedly links Jedi wisdom to Christianity. Well, after 60 pages or so, I have got the feeling that the Staub is trying to wrap Jedi in Christian cloth and failing at that. The author speaks of the difficulty i...
	Lost Credability, QUICKLY!! I admit, I haven't finished this book. A friend recommended it to

amazon review › Feature Hashing › Transformed dataset

256 (2^8) features are created through feature hashing

rows 10000 columns 258

Col1	Col2	Col2_HashingFeature_1	Col2_HashingFeature_2	Col2_HashingFeature_3	Col2_HashingFeature_4	Col2_HashingFeature_5	Col2_HashingFeature_6	Col2_HashingFeature_7	Col2_HashingFeature_8	Col2_HashingFeature_9
2	Ok~ but I think the Keirsey Temperment Test is more accurate - and cheaper. This book has its good points. If anything, it helps you put into words what you want from a supervisor, but it is not very accurate. The online test does not account for a difference between when 2 of their options are bo...	3	3	0	4	0	3	1	3	4
2	Repellent Sale of Conservativism The fatalistic view of Conservativism is expressed by Kirk even before the first chapter concludes as he states, " [Edmund] Burke, could he see our century, never would concede that a consumption-society, so near suicide, is the end for which Providence has	9	5	6	13	1	10	4	4	1

The input dataset contains two columns: the book rating ranging from 1 to 5, and the actual review content. The goal of this Feature Hashing module is to retrieve a bunch of new features that show the occurrence frequency of the corresponding word(s)/phrase(s) within the particular book review. Instead of associating each text feature (words/phrases) to a particular index, this method functions by applying a hash function to the features and using their hash values as indices directly.

Source: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-create-features

:: Feature Engineering - example

Example 4

Bike Rental dataset
The dataset represents the number of bike rentals within a specific hour of a day in the years 2011 and year 2012 and contains 17379 rows and 17 columns.

Raw Data

rows 17379 columns 17

instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1	2011-01-01T00:00:00	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0	3	13	16
2	2011-01-01T00:00:00	1	0	1	1	0	6	0	1	0.22	0.2727	0.8	0	8	32	40
3	2011-01-01T00:00:00	1	0	1	2	0	6	0	1	0.22	0.2727	0.8	0	5	27	32
4	2011-01-01T00:00:00	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0	3	10	13
5	2011-01-01T00:00:00	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0	0	1	1
6	2011-01-01T00:00:00	1	0	1	5	0	6	0	2	0.24	0.2576	0.75	0.0896	0	1	1
7	2011-01-01T00:00:00	1	0	1	6	0	6	0	1	0.22	0.2727	0.8	0	2	0	2
8	2011-01-01T00:00:00	1	0	1	7	0	6	0	1	0.2	0.2576	0.86	0	1	2	3
9	2011-01-01T00:00:00	1	0	1	8	0	6	0	1	0.24	0.2879	0.75	0	1	7	8
10	2011-01-01T00:00:00	1	0	1	9	0	6	0	1	0.32	0.3485	0.76	0	8	6	14
11	2011-01-01T00:00:00	1	0	1	10	0	6	0	1	0.38	0.3939	0.76	0.2537	12	24	36
12	2011-01-01T00:00:00	1	0	1	11	0	6	0	1	0.36	0.3333	0.81	0.2836	26	30	56
13	2011-01-01T00:00:00	1	0	1	12	0	6	0	1	0.42	0.4242	0.77	0.2836	29	55	84
14	2011-01-01T00:00:00	1	0	1	13	0	6	0	2	0.46	0.4545	0.72	0.2985	47	47	94
15	2011-01-01T00:00:00	1	0	1	14	0	6	0	2	0.46	0.4545	0.72	0.2836	35	71	106
16	2011-01-01T00:00:00	1	0	1	15	0	6	0	2	0.44	0.4394	0.77	0.2985	40	70	110
17	2011-01-01T00:00:00	1	0	1	16	0	6	0	2	0.42	0.4242	0.82	0.2985	41	52	93

The dataset "Bike Rental UCI dataset" is used as the raw input data. This dataset is based on real data from the Capital Bikesh are company that maintains a bike rental network in Washington DC in the United States.

Predict the demand for the bikes

The objective of this experiment is to predict the demand for the bikes, that is, the number of bike rentals within a specific month/day/hour.

The raw feature set contains weather conditions (temperature/humidity/wind speed) and the type of the day (holiday/weekday).

The field to predict is "cnt", a count which represents the bike rentals within a specific hour and which ranges ranges from 1 to 977.

It can be tempting to includes many raw data fields in the feature set, but more often, you need to construct additional features from the raw data to provide better predictive power. This is called feature engineering.

Source: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-create-features

:: Feature Engineering - example

Example 4 - Continued

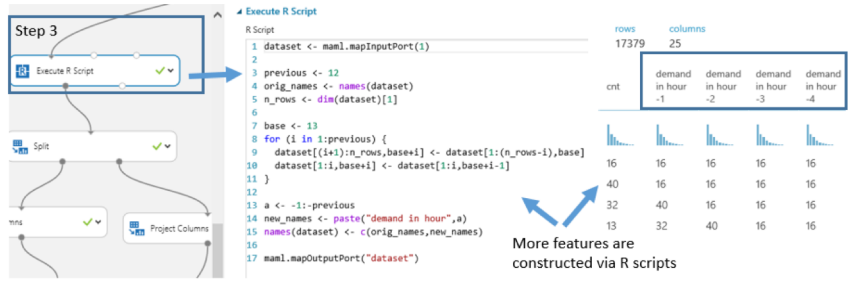
Because our goal was to construct effective features in the training data, we built four models using the same algorithm, but with four different training datasets.

The four training datasets that we constructed were all based on the same raw input data, but we added different additional features to each training set.

- **Set A** = weather + holiday + weekday + weekend features for the predicted day
- **Set B** = number of bikes that were rented in each of the previous 12 hours
- **Set C** = number of bikes that were rented in each of the previous 12 days at the same hour
- **Set D** = number of bikes that were rented in each of the previous 12 weeks at the same hour and the same day

Besides feature set A, which already exist in the original raw data, the other three sets of features are created through the feature engineering process. Each of these feature sets captures different aspects of the problem:

- Feature set B captures very recent demand for the bikes.
- Feature set C captures the demand for bikes at a particular hour.
- Feature set D captures demand for bikes at a particular hour and particular day of the week.

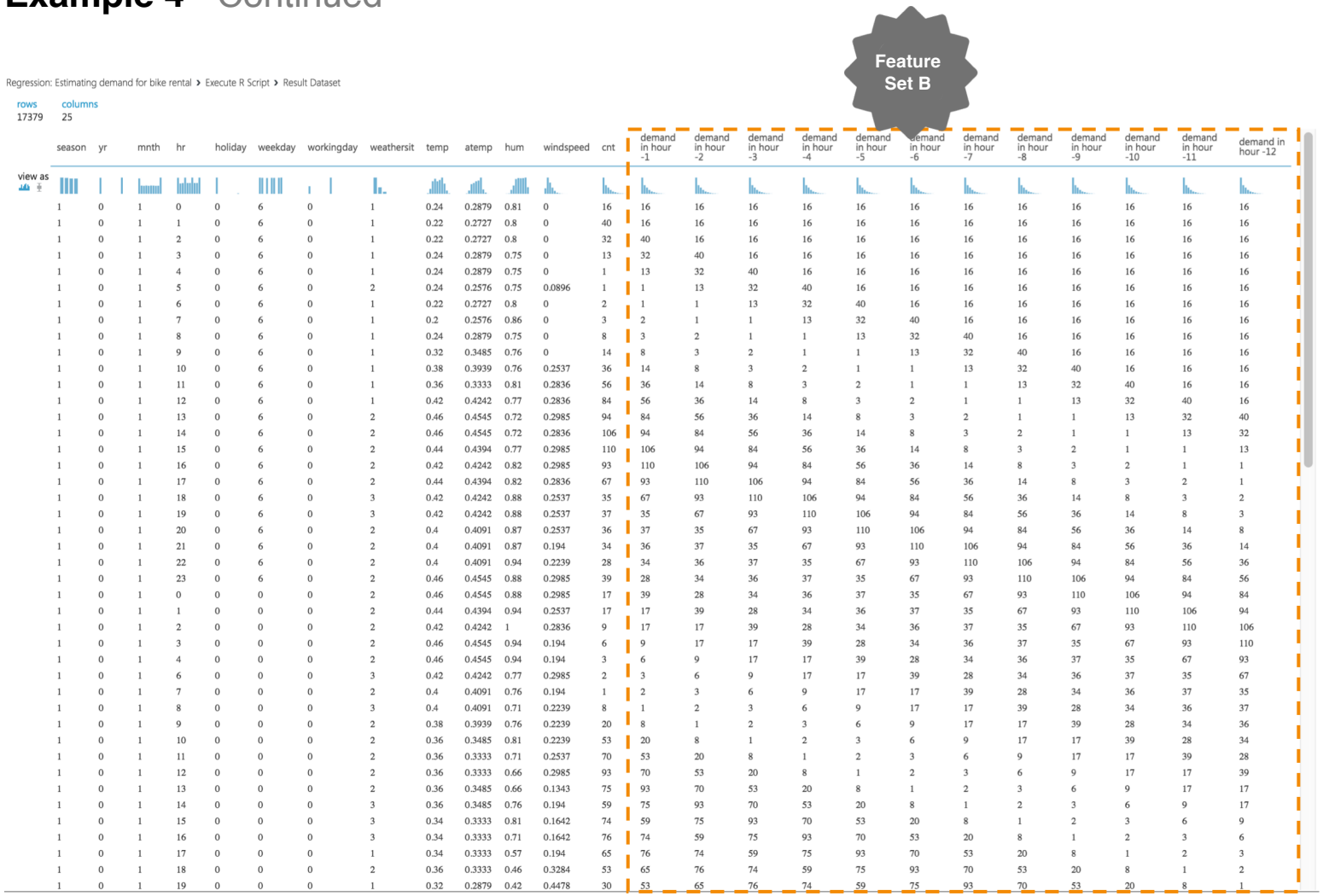


For example, use the Execute R Script module to construct different sets of derived features and to append the new features to each dataset.

Source: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-create-features

:: Feature Engineering - example

Example 4 - Continued



Training set 2: feature sets A+B

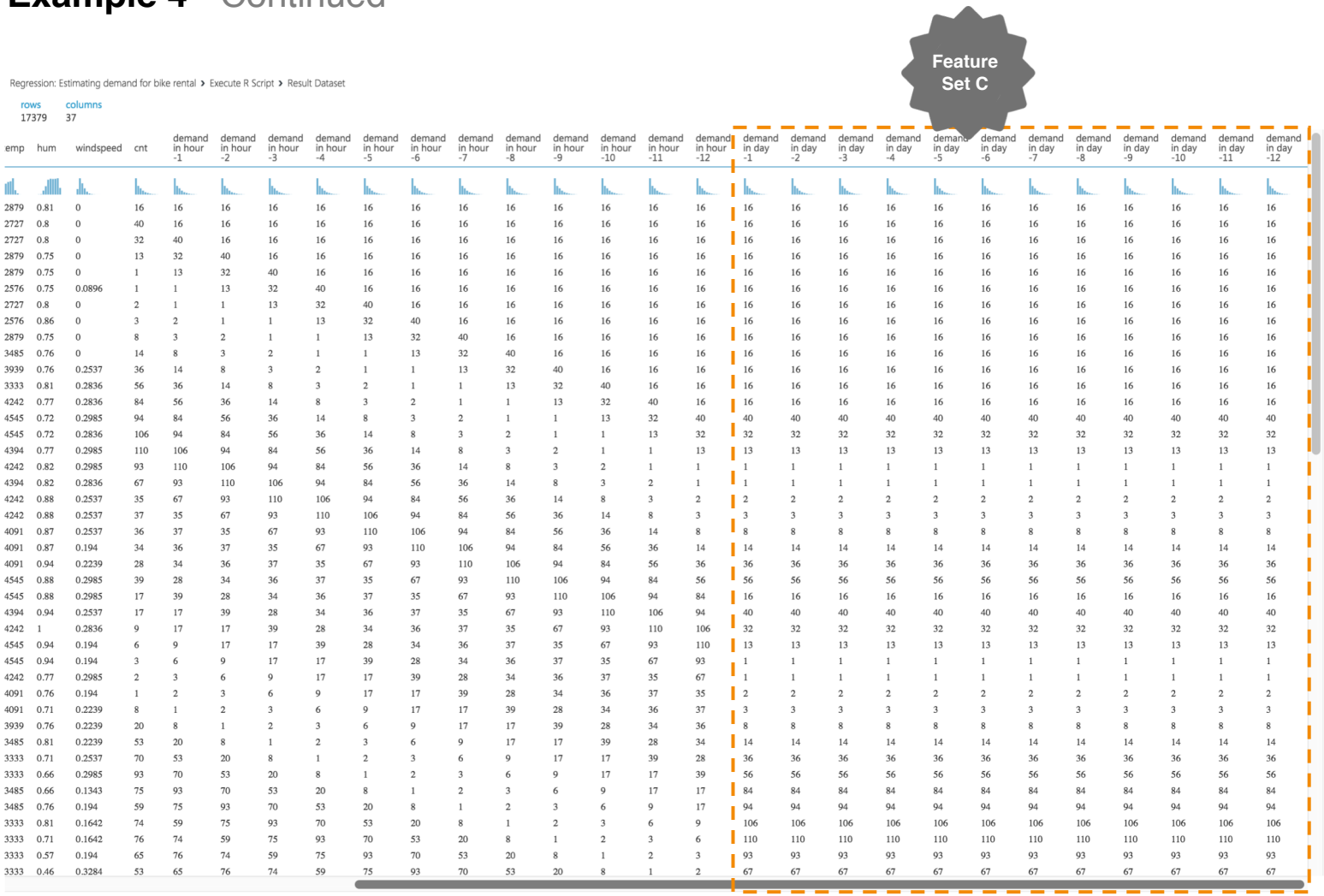
Additional features are created via feature engineering

- number of bikes that were rented in each of the previous 12 hours

Source: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-create-features

:: Feature Engineering - example

Example 4 - Continued



emp	hum	windspeed	cnt	demand in hour -1	demand in hour -2	demand in hour -3	demand in hour -4	demand in hour -5	demand in hour -6	demand in hour -7	demand in hour -8	demand in hour -9	demand in hour -10	demand in hour -11	demand in hour -12	demand in day -1	demand in day -2	demand in day -3	demand in day -4	demand in day -5	demand in day -6	demand in day -7	demand in day -8	demand in day -9	demand in day -10	demand in day -11	demand in day -12
2879	0.81	0	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16
2727	0.8	0	40	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16
2727	0.8	0	32	40	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16
2879	0.75	0	13	32	40	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16
2879	0.75	0	1	13	32	40	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16
2576	0.75	0.0896	1	1	13	32	40	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16
2727	0.8	0	2	1	1	13	32	40	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16
2576	0.86	0	3	2	1	1	13	32	40	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16
2879	0.75	0	8	3	2	1	1	13	32	40	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16
3485	0.76	0	14	8	3	2	1	1	13	32	40	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16
3939	0.76	0.2537	36	14	8	3	2	1	1	13	32	40	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16
3333	0.81	0.2836	56	36	14	8	3	2	1	1	13	32	40	16	16	16	16	16	16	16	16	16	16	16	16	16	16
4242	0.77	0.2836	84	56	36	14	8	3	2	1	1	13	32	40	16	16	16	16	16	16	16	16	16	16	16	16	16
4545	0.72	0.2985	94	84	56	36	14	8	3	2	1	1	13	32	40	40	40	40	40	40	40	40	40	40	40	40	40
4545	0.72	0.2836	106	94	84	56	36	14	8	3	2	1	1	13	32	32	32	32	32	32	32	32	32	32	32	32	32
4394	0.77	0.2985	110	106	94	84	56	36	14	8	3	2	1	1	13	13	13	13	13	13	13	13	13	13	13	13	13
4242	0.82	0.2985	93	110	106	94	84	56	36	14	8	3	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4394	0.82	0.2836	67	93	110	106	94	84	56	36	14	8	3	2	1	1	1	1	1	1	1	1	1	1	1	1	1
4242	0.88	0.2537	35	67	93	110	106	94	84	56	36	14	8	3	2	2	2	2	2	2	2	2	2	2	2	2	2
4242	0.88	0.2537	37	35	67	93	110	106	94	84	56	36	14	8	3	3	3	3	3	3	3	3	3	3	3	3	3
4091	0.87	0.2537	36	37	35	67	93	110	106	94	84	56	36	14	8	8	8	8	8	8	8	8	8	8	8	8	8
4091	0.87	0.194	34	36	37	35	67	93	110	106	94	84	56	36	14	14	14	14	14	14	14	14	14	14	14	14	14
4091	0.94	0.2239	28	34	36	37	35	67	93	110	106	94	84	56	36	36	36	36	36	36	36	36	36	36	36	36	36
4545	0.88	0.2985	39	28	34	36	37	35	67	93	110	106	94	84	56	56	56	56	56	56	56	56	56	56	56	56	56
4545	0.88	0.2985	17	39	28	34	36	37	35	67	93	110	106	94	84	16	16	16	16	16	16	16	16	16	16	16	16
4394	0.94	0.2537	17	17	39	28	34	36	37	35	67	93	110	106	94	40	40	40	40	40	40	40	40	40	40	40	40
4242	1	0.2836	9	17	17	39	28	34	36	37	35	67	93	110	106	32	32	32	32	32	32	32	32	32	32	32	32
4545	0.94	0.194	6	9	17	17	39	28	34	36	37	35	67	93	110	13	13	13	13	13	13	13	13	13	13	13	13
4545	0.94	0.194	3	6	9	17	17	39	28	34	36	37	35	67	93	1	1	1	1	1	1	1	1	1	1	1	1
4242	0.77	0.2985	2	3	6	9	17	17	39	28	34	36	37	35	67	1	1	1	1	1	1	1	1	1	1	1	1
4091	0.76	0.194	1	2	3	6	9	17	17	39	28	34	36	37	35	2	2	2	2	2	2	2	2	2	2	2	2
4091	0.71	0.2239	8	1	2	3	6	9	17	17	39	28	34	36	37	3	3	3	3	3	3	3	3	3	3	3	3
3939	0.76	0.2239	20	8	1	2	3	6	9	17	17	39	28	34	36	8	8	8	8	8	8	8	8	8	8	8	8
3485	0.81	0.2239	53	20	8	1	2	3	6	9	17	17	39	28	34	14	14	14	14	14	14	14	14	14	14	14	14
3333	0.71	0.2537	70	53	20	8	1	2	3	6	9	17	17	39	28	36	36	36	36	36	36	36	36	36	36	36	36
3333	0.66	0.2985	93	70	53	20	8	1	2	3	6	9	17	17	39	56	56	56	56	56	56	56	56	56	56	56	56
3485	0.66	0.1343	75	93	70	53	20	8	1	2	3	6	9	17	17	84	84	84	84	84	84	84	84	84	84	84	84
3485	0.76	0.194	59	75	93	70	53	20	8	1	2	3	6	9	17	94	94	94	94	94	94	94	94	94	94	94	94
3333	0.81	0.1642	74	59	75	93	70	53	20	8	1	2	3	6	9	106	106	106	106	106	106	106	106	106	106	106	106
3333	0.71	0.1642	76	74	59	75	93	70	53	20	8	1	2	3	6	110	110	110	110	110	110	110	110	110	110	110	110
3333	0.57	0.194	65	76	74	59	75	93	70	53	20	8	1	2	3	93	93	93	93	93	93	93	93	93	93	93	93
3333	0.46	0.3284	53	65	76	74	59	75	93	70	53	20	8	1	2	67	67	67	67	67	67	67	67	67	67	67	67

Training set 3: feature sets A+B+C

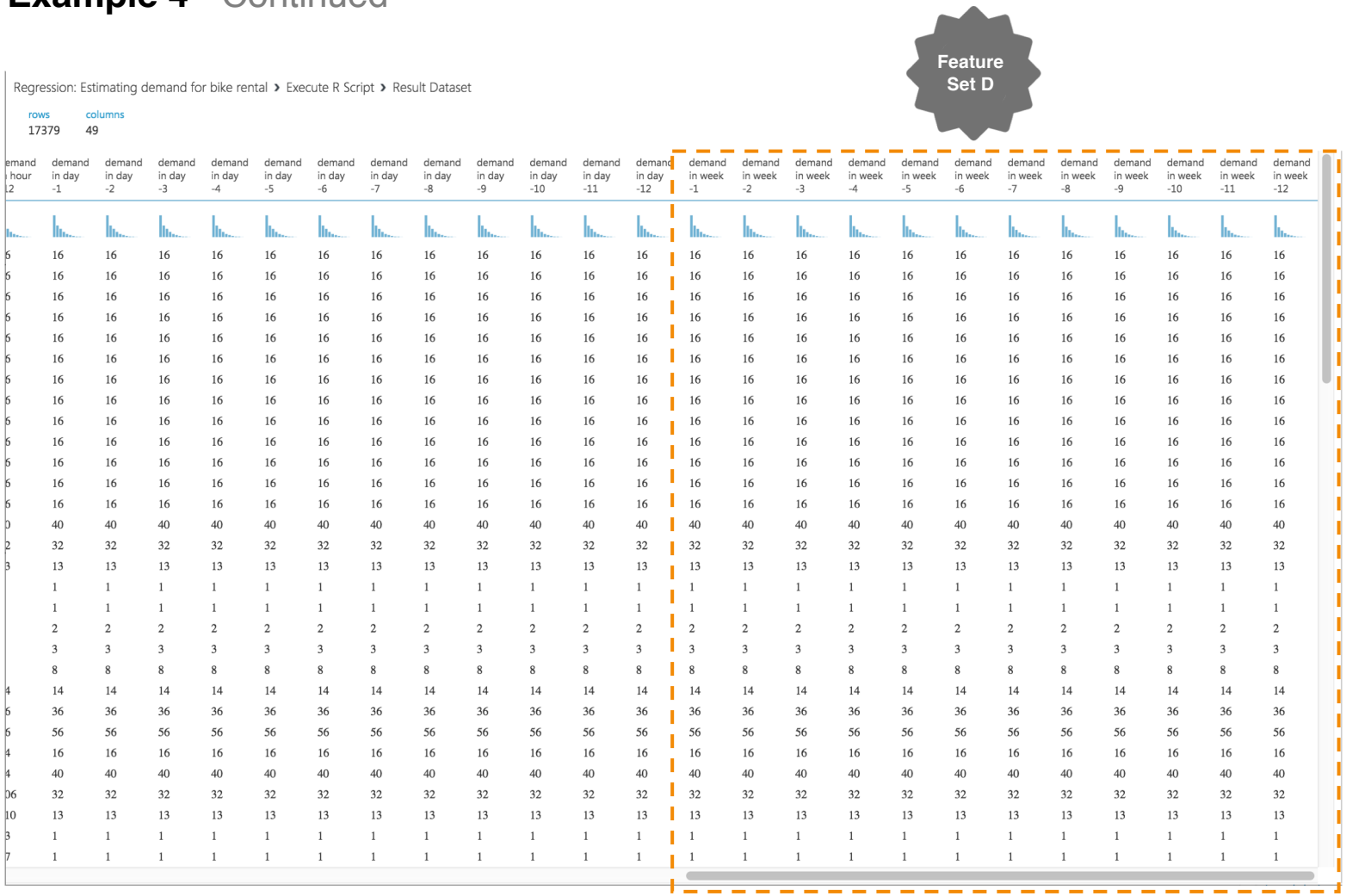
Additional features are created via feature engineering

- number of bikes that were rented in each of the previous 12 hours at the same hour

Source: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-create-features

:: Feature Engineering - example

Example 4 - Continued



Training set 4: feature sets A+B+C+D

Additional features are created via feature engineering

- number of bikes that were rented in each of the previous 12 hours at the same hour and the same day

Source: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-create-features

:: Feature Engineering - example

The comparison of the performance results of the four models which are based on 4 training sets

Features	Mean Absolute Error	Root Mean Squared Error	Coefficient of Determination
Feature Set A (training set1) baseline: weather + holiday + weekday + weekend features for the predicted day	89.7	124.9	0.6
Feature Set A+B (training set2) baseline + previous 12 hours demand	51.7	88.3	0.8
Feature Set A+B+C (training set3) baseline + previous 12 hours demand + previous 12 days at the same hour	47.6 	81.1 	0.8
Feature Set A+B+C+D (training set4) baseline + previous 12 hours demand + previous 12 days at the same hour + previous 12 weeks at the same hour and the same day demand	48.3	82.1	0.8

Same algorithm but different performance

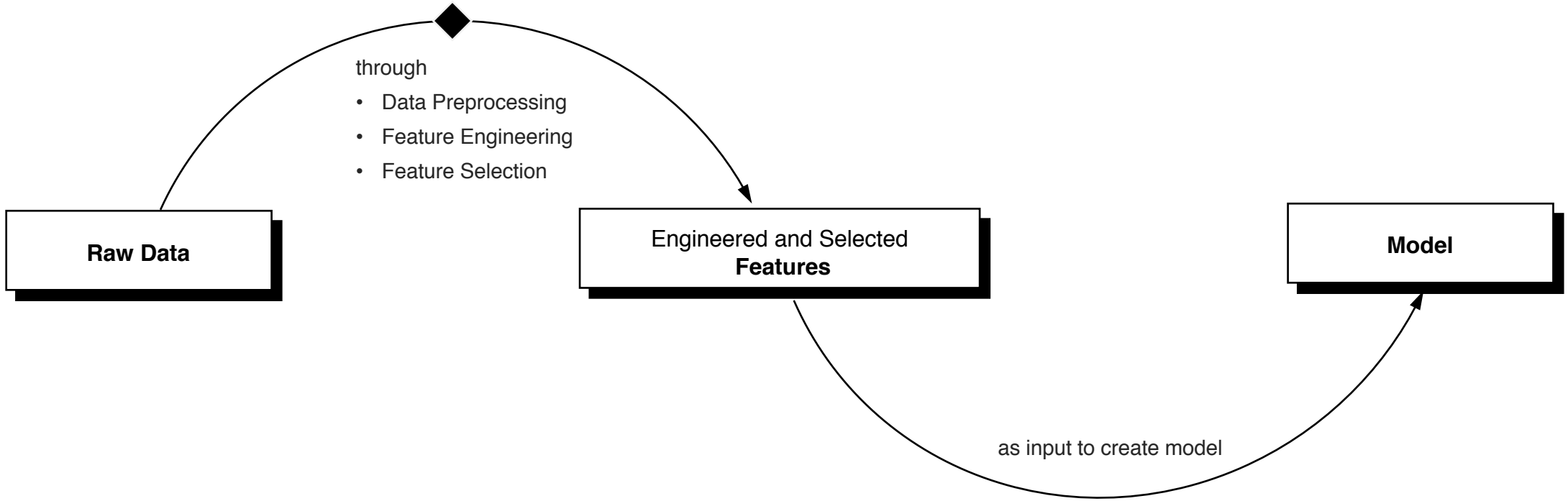
we used the Boosted Decision Tree Regression module, a commonly used nonlinear algorithm, to build the models.

To understand the performance of four models, see the comparison results in the following table.

- The best results were from the combination of features A+B+C and A+B+C+D.
- Feature set D does not provide additional improvement over A+B+C.

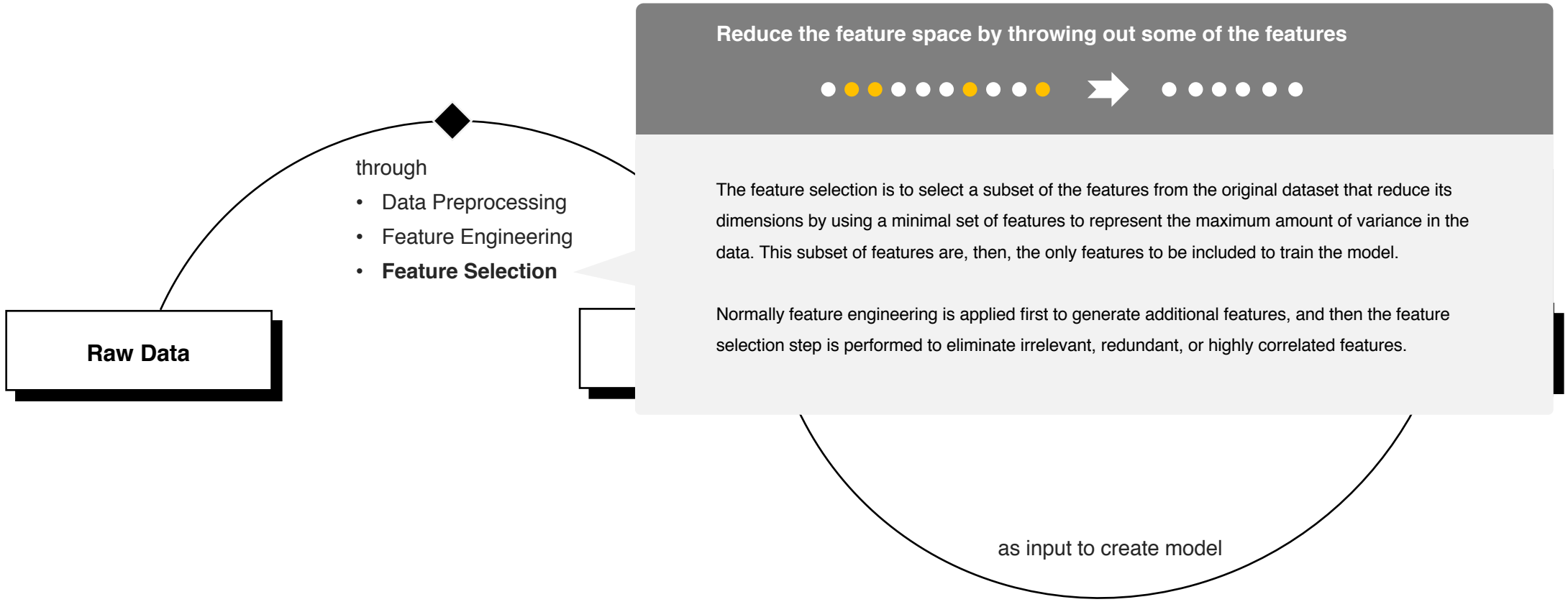
Source: https://gallery.cortanaintelligence.com/Experiment/Regression-Demand-estimation-4

:: Feature Engineering can augment your data



1 & 2 Source: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-create-features

:: Feature Engineering can augment your data



Mathematically speaking, the features selected to train the model are a minimal set of independent variables that explain the patterns in the data and then predict outcomes successfully[1].

Feature selection is especially useful when you're dealing with high-dimensional data or when your dataset contains a large number of features and a limited number of observations[2].

1 source: Microsoft 2 source: Mathworks, Applying Supervised Learning

:: Why should you perform Feature Selection ?

Often, data contains many features that are either redundant or irrelevant

Feature selection techniques are used for four reasons:

- **simplification of models** to make them easier to interpret by researchers/users
- **shorter training times** and speed up learning process
- **to avoid the curse of dimensionality** (See next slide for details)
- **enhanced generalization** by reducing overfitting

Often, data contains many features that are either *redundant* or *irrelevant*, and can be removed without incurring much loss of information[1].

Note: Feature selection is also known as *variable selection, attribute selection*

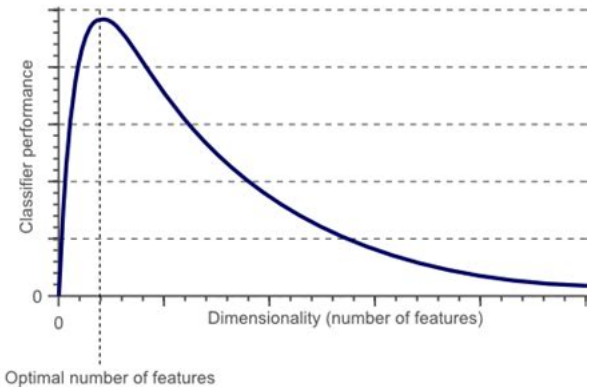
1 Source: https://en.wikipedia.org/wiki/Feature_selection

:: Why should you perform Feature Selection ?

Curse of dimensionality

The curse of dimensionality refers to how certain learning algorithms may perform poorly in high-dimensional data.

For example, after a certain point, increasing the dimensionality of the problem by adding new features would actually degrade the performance of our classifier. This is illustrated by the following figure, and is often referred to as 'The Curse of Dimensionality' [2].



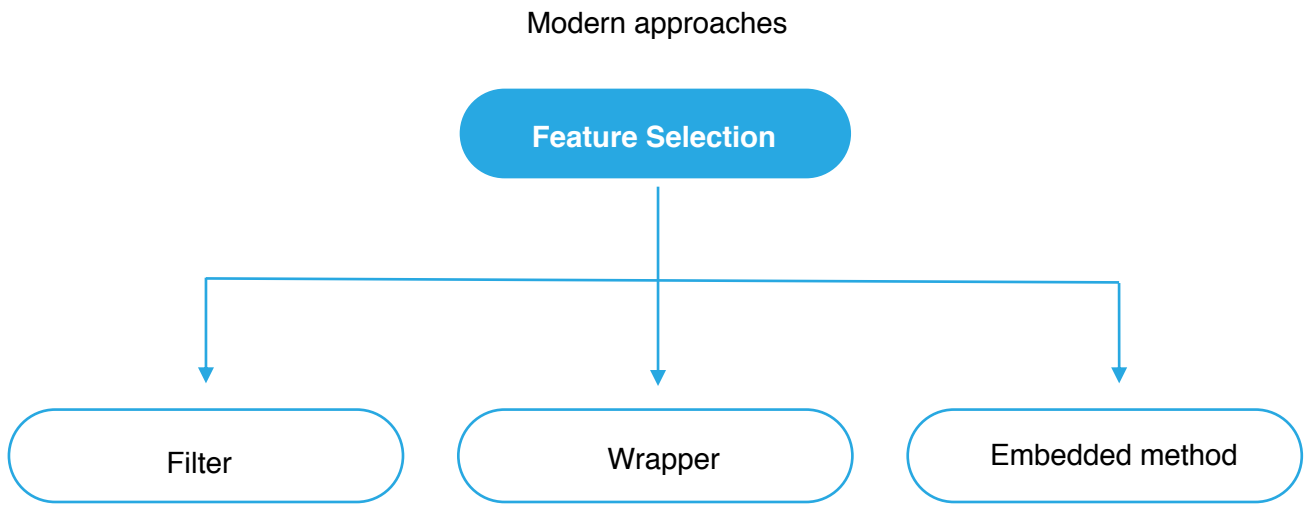
As you add more features, you may **exponentially** need more data to kind of fill out the space.

However, in practice the number of training examples is fixed.
It's not easy to add more data as you want.

1 Source: http://www.datasciencecentral.com/profiles/blogs/about-the-curse-of-dimensionality
Image Source: http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/

:: Approaches for Feature Selection

Modern appoaches for feature selection



Traditional approaches

- **Forward selection**
  starts with no variables in the model. You then iteratively add variables and test the predictive accuracy of the model until adding more variables no longer makes a positive effect.

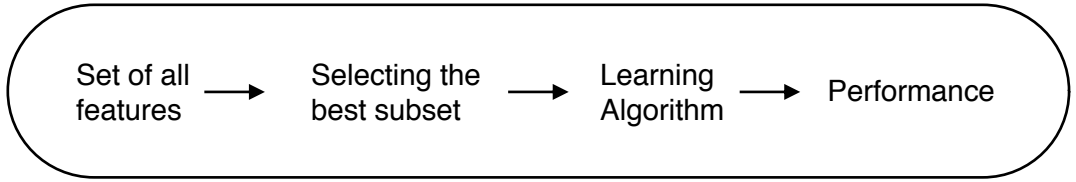
- **Backward elimination**
  proceed by removing variables and testing the predictive accuracy of the model.
  begins with all the variables in the model.

- **Stepwise regression ★**
  This is an algorithm that adds the best feature (or deletes the worst feature) in a series of iterative steps.

  At each stage in the process, after a new variable is added, a test is made to check if some variables can be deleted without appreciably increasing the error. The procedure terminates when the measure is (locally) maximized, or when the available improvement falls below some critical value.

:: Feature Selection – Filter method

Filter type methods select variables regardless of the model



Filter

By evaluating the correlation between each feature and the target attribute, these methods apply a statistical measure to assign a score to each feature. The features are then ranked by the score, which may be used to help set the threshold for keeping or eliminating a specific feature [1].

The statistical measures used in these methods include [2]

- Pearson correlation
- Mutual information
- Chi squared test
- And so on…

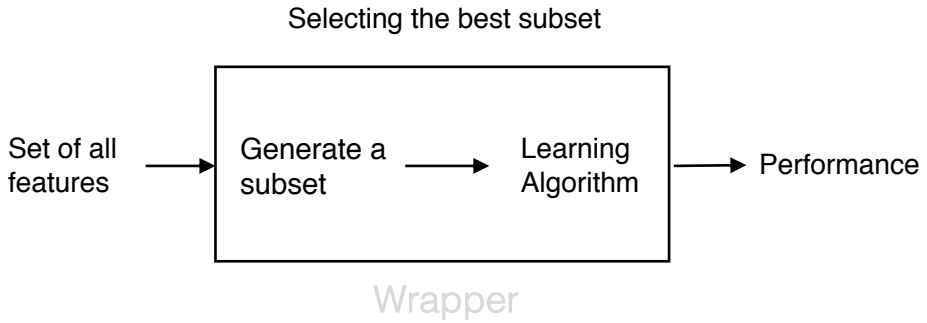
These methods are particularly effective in computation time and robust to overfitting.

However, filter methods tend to select redundant variables because they do not consider the relationships between variables. Therefore, they are mainly used as a pre-process method.

1&2 Source: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-select-features
Source: https://en.wikipedia.org/wiki/Feature_selection

:: Feature Selection – Wrapper method

Wrapper methods use a predictive model to score feature subsets



Wrapper methods consider the selection of a set of features as a search problem, where different combinations are prepared, evaluated and compared to other combinations. A predictive model is used to evaluate a combination of features and assign a score based on model accuracy.

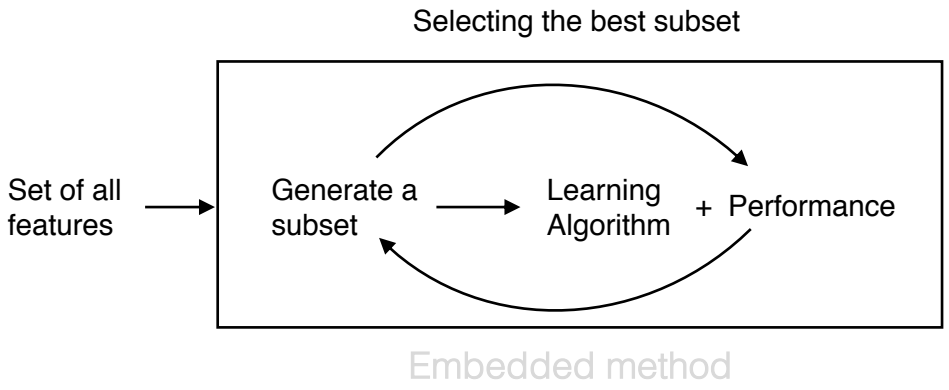
Wrapper methods evaluate subsets of variables which allows, unlike filter approaches, to detect the possible interactions between variables

As wrapper methods train a new model for each subset, they are very computationally intensive, but usually provide the best performing feature set for that particular type of model.

Source: https://en.wikipedia.org/wiki/Feature_selection

:: Feature Selection – Embedded method

Feature selection is a part of model construction



Embedded methods learn which features best contribute to the accuracy of the model while the model is being created.

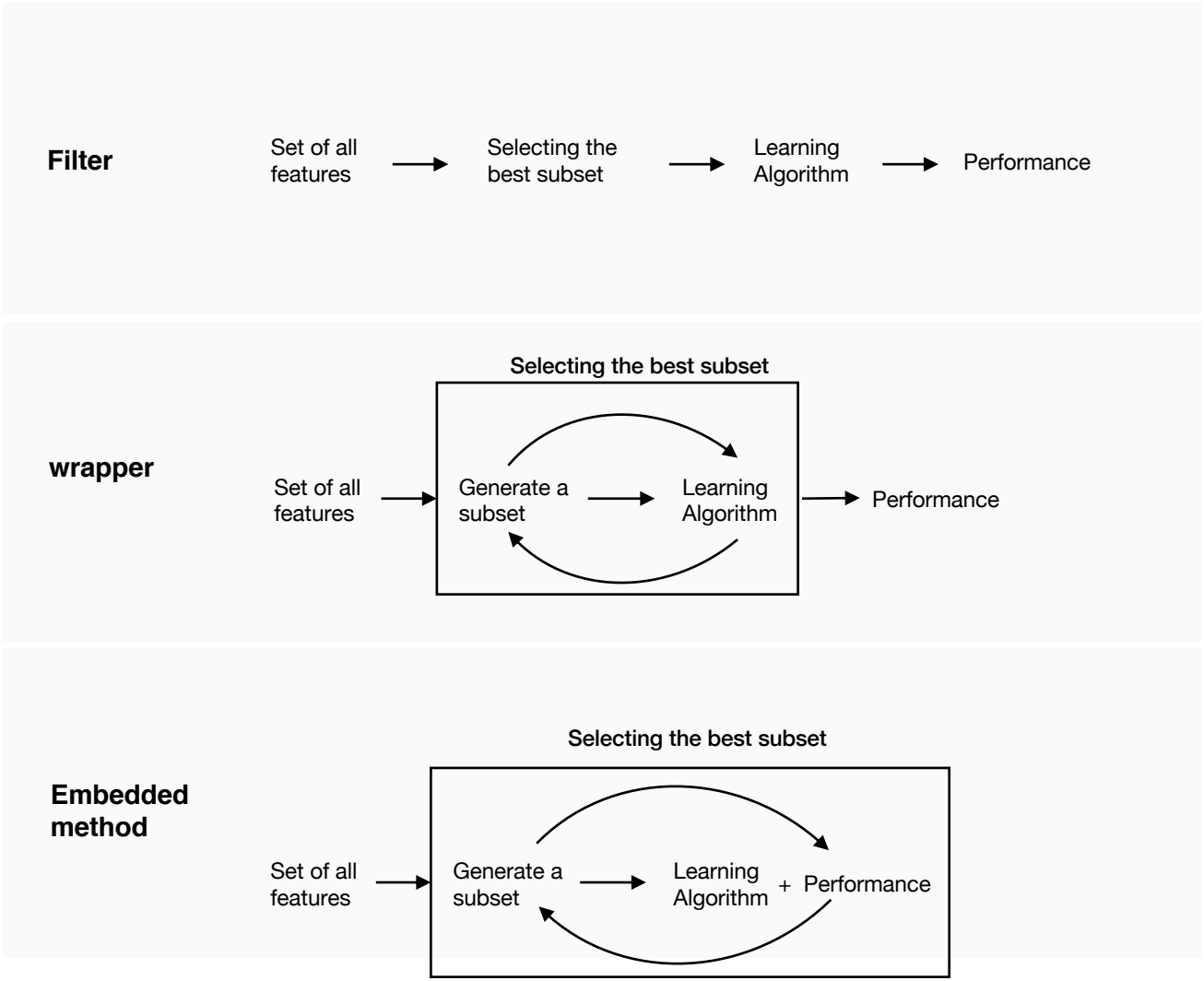
The most common type of embedded feature selection methods are regularization methods.

Regularization methods are also called penalization methods that introduce additional constraints into the optimization of a predictive algorithm (such as a regression algorithm) that bias the model toward lower complexity (fewer coefficients).

Some of the most popular examples of these methods are LASSO and RIDGE regression which have inbuilt penalization functions to reduce overfitting.

Source: https://en.wikipedia.org/wiki/Feature_selection

:: Summary: the modern approaches for Feature Selection



- Filter methods use statistical methods for evaluation of a subset of features while wrapper methods use cross validation.
- Filter methods are much faster compared to wrapper methods as they do not involve training the models. On the other hand, wrapper methods are computationally very expensive as well.
- Filter methods might fail to find the best subset of features in many occasions but wrapper methods can always provide the best subset of features.
- Using the subset of features from the wrapper methods make the model more prone to overfitting as compared to using subset of features from the filter methods.
- Embedded methods combine the qualities' of filter and wrapper methods. It's implemented by algorithms that have their own built-in feature selection methods.

Source: https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/

:: Feature Selection vs Dimension Reduction

Although feature selection does seek to reduce the number of features in the dataset used to train the model, it is not usually referred to by the term “dimensionality reduction”.

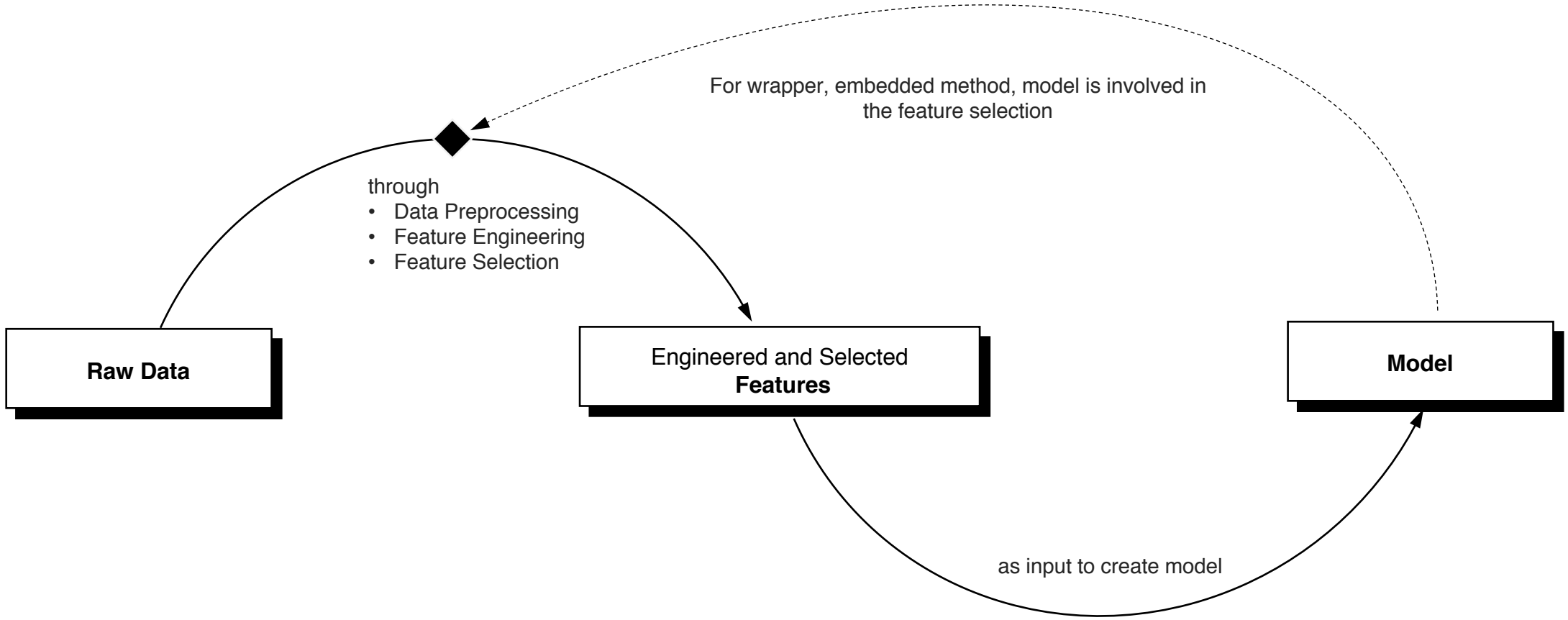
Feature selection methods extract a subset of original features in the data without changing them. [1]

Dimensionality reduction methods employ engineered features that can transform the original features and thus modify them.

Examples of dimensionality reduction methods include Principal Component Analysis, canonical correlation analysis, and Singular Value Decomposition. [2]

1 ,2 source: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-select-features

:: Recap: Transform raw data into the features



:: Feature Engineering’s definition has different versions

Google’s definition covers more scope than Microsoft’s definition

In Google's definition of feature engineering, it includes data prepressing, data augmentation, feature selection, etc.

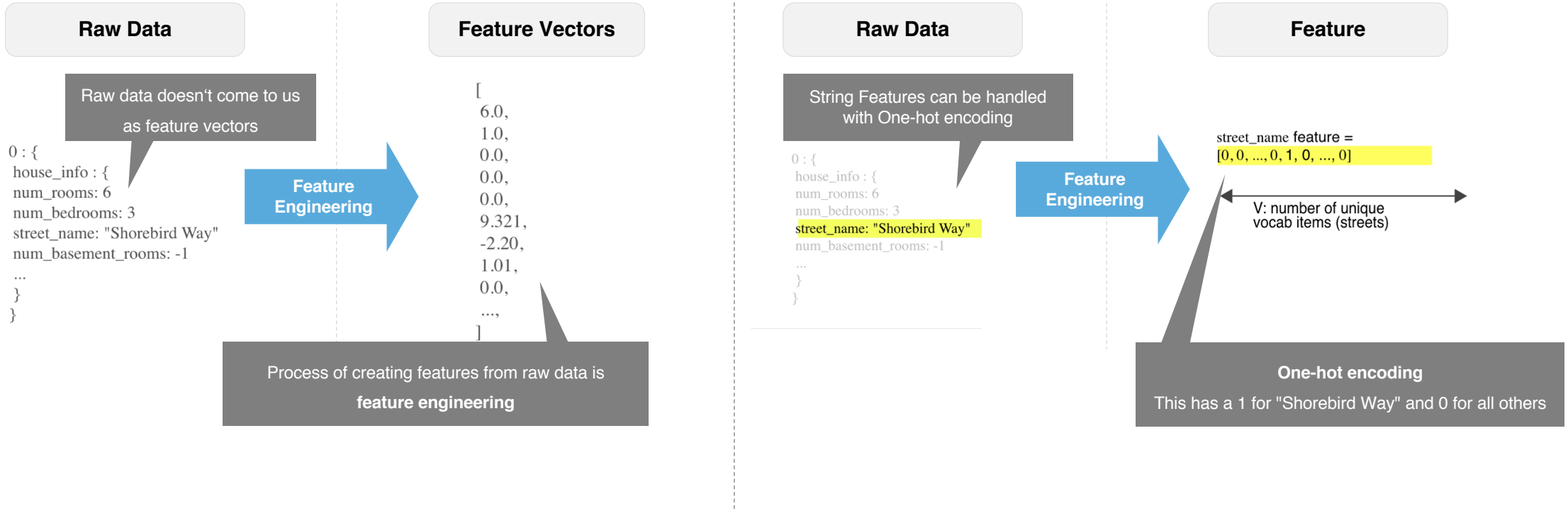
In previous slides, we talk about Microsoft’s definition of feature engineering.

Now let’s know how Google defines feature engineering

:: Let's take a look at how Google defines Feature Engineering

Feature engineering means transforming raw data into a feature vector [1]

Process of creating features from raw data is feature engineering. Expect to spend significant time doing feature engineering.[2]

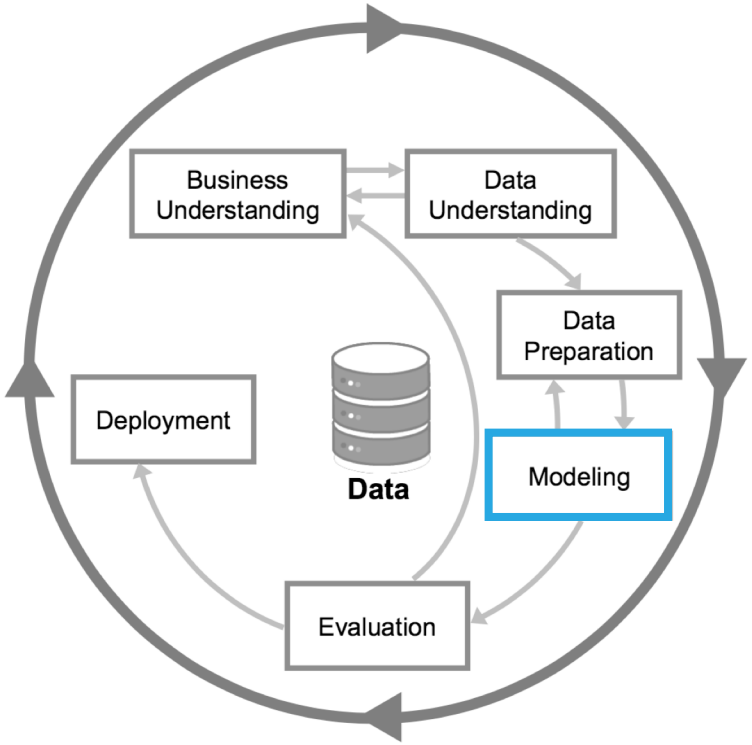


Further reading | **Cleaning Data (by Google) - 10 mins**
https://developers.google.com/machine-learning/crash-course/representation/cleaning-data

1&2 source: https://developers.google.com/machine-learning/crash-course/glossary#feature_engineering

05 | Modelling

:: Train the model



The modelling is an iterative process of building model which is a combination of data structure, algorithm, and mathematics to captures the relationship between features and target within the dataset

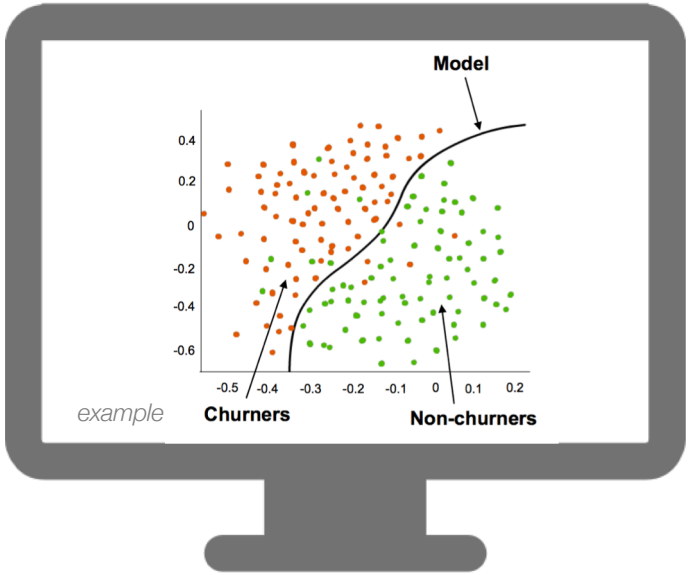
Source: https://www.the-modeling-agency.com/crisp-dm.pdf

:: What's a model ?

A model defines the relationship between features and label (i.e. target) [1]

Machine learning uses a model to capture the relationship between **feature vectors** and some **target variables** within a training data set.

A feature vector is a set of features or attributes that characterize a particular object, such as the number of bedrooms, bathrooms, and location of an apartment. The target is either a scalar value like rent price, or it's an integer classification such as "creditworthy" or "it's not cancer." [2]



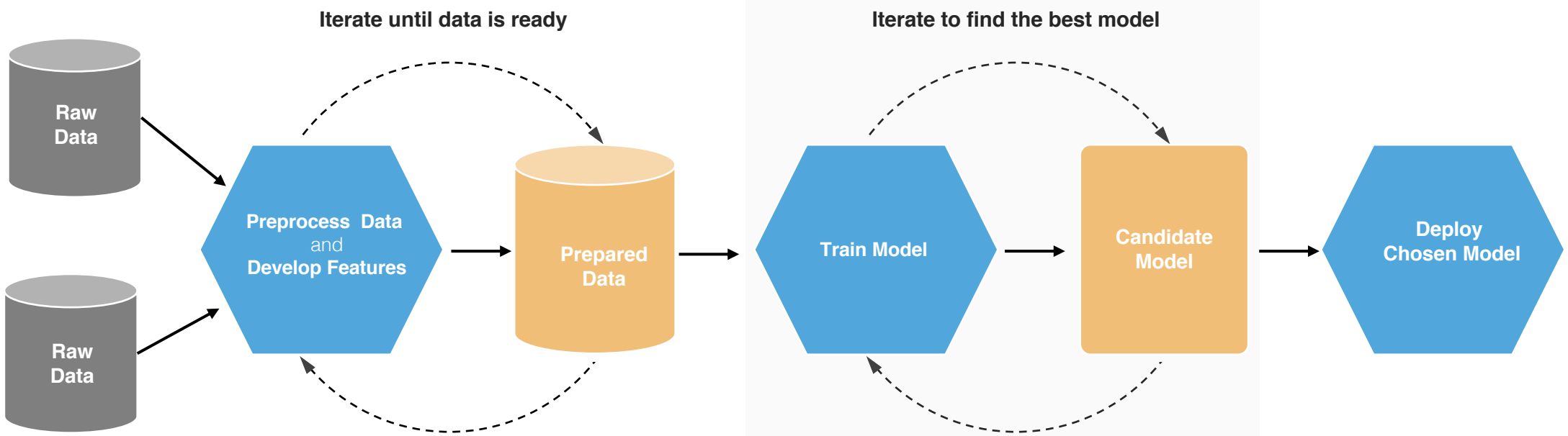
"All models are wrong, but some are useful."

- George Box

- *An machine learning model is a mathematical model that generates predictions by finding patterns in your data [3].*
- *At a high level, a model is a simplification of something more complex [4].*
- *A machine learning algorithm use data to automatically learn the rules. It simplifies the complexity of the data into relationships described by rules [5].*
- *Modelling is part art and part science [6].*

1 Source : Google, https://developers.google.com/machine-learning/crash-course/framing/ml-terminology
2Source : https://mlbook.explained.ai/intro.html#sec:2.2.3
3 source: Amazon, https://docs.aws.amazon.com/machine-learning/latest/dg/amazon-machine-learning-key-concepts.html
4,5,6 source Microsoft, https://channel9.msdn.com/Events/OpenSourceTW/DevDays-Asia-2017/AI11:

:: The process of Machine Learning



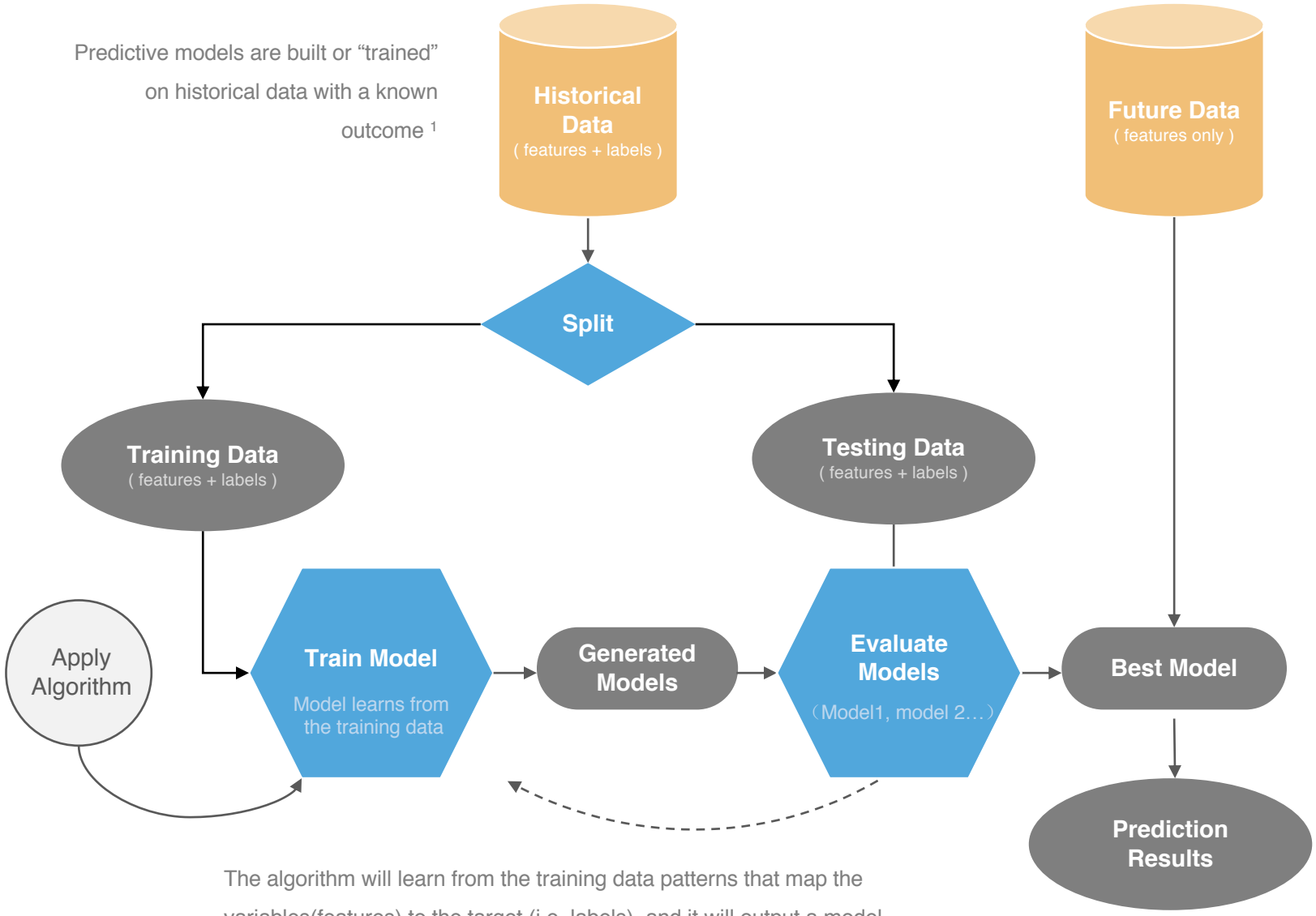
Train model, Evaluate Model & Optimization

- find the model that answers the question most accurately by comparing their success metrics
- Determine if your model is suitable for production

The diagram is revised version of the original diagram from Microsoft. Source of original diagram: INTRODUCING AZURE MACHINE LEARNING by Microsoft

:: How to develop a model ?

Predictive models are built or “trained” on historical data with a known outcome [1]



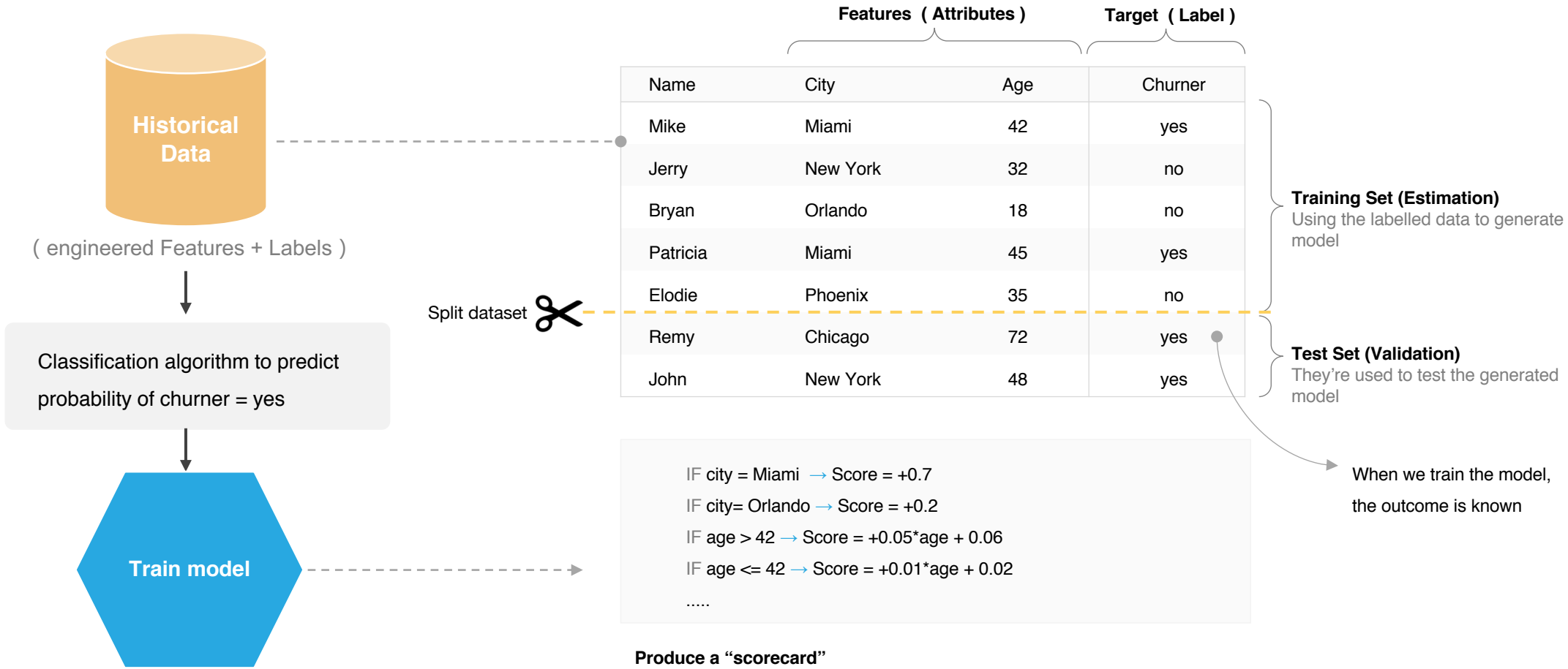
Once the model has been built, it’s applied onto new, more recent data which has an unknown outcome (because the outcome is in the future) [2].

The algorithm will learn from the training data patterns that map the variables(features) to the target (i.e. labels), and it will output a model that captures these relationships [3].

1&2 Source: SAP, https://open.sap.com/courses/ds1
3 source: Amazon, http://docs.aws.amazon.com/machine-learning/latest/dg/training-the-model.html

:: Example of building the model (Learning phase)

Predict if a customer is going to switch to another supplier



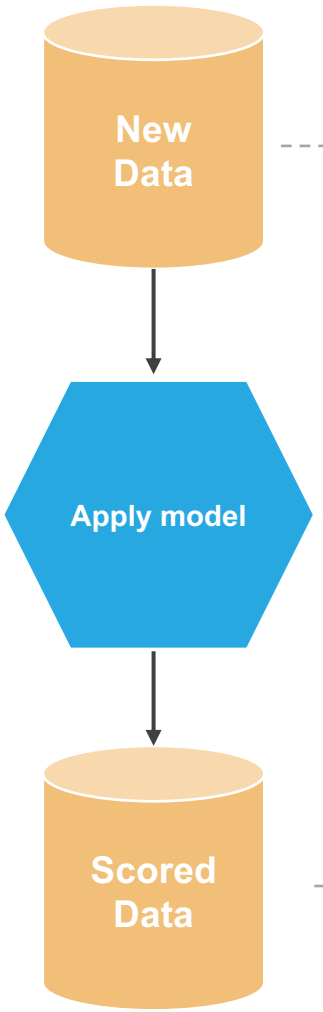
Produce a "scorecard"

Add up each component score to give an overall score for each customer.

This will equate to their churn probability.

Source: SAP, https://open.sap.com/courses/ds1

:: Example of using the model (Applying phase)



Name	City	Age	Churner
Marine	Miami	45	?
Julien	Miami	52	?
Fred	Orlando	20	?
Michelle	Boston	34	?
Nicolas	Phoenix	90	?

Unknown Outcome

Recent data, with customers who have not yet made a decision to churn or remain.

IF city = Miami → Score = +0.7
IF city= Orlando → Score = +0.2
IF age > 42 → Score = +0.05*age + 0.06
IF age <= 42 → Score = +0.01*age + 0.02
.....

Name	City	Age	Churner
Marine	Miami	45	**0.8**
Julien	Miami	52	**0.9**
Fred	Orlando	20	**0.6**
Michelle	Boston	34	**0.5**
Nicolas	Phoenix	90	**0.4**

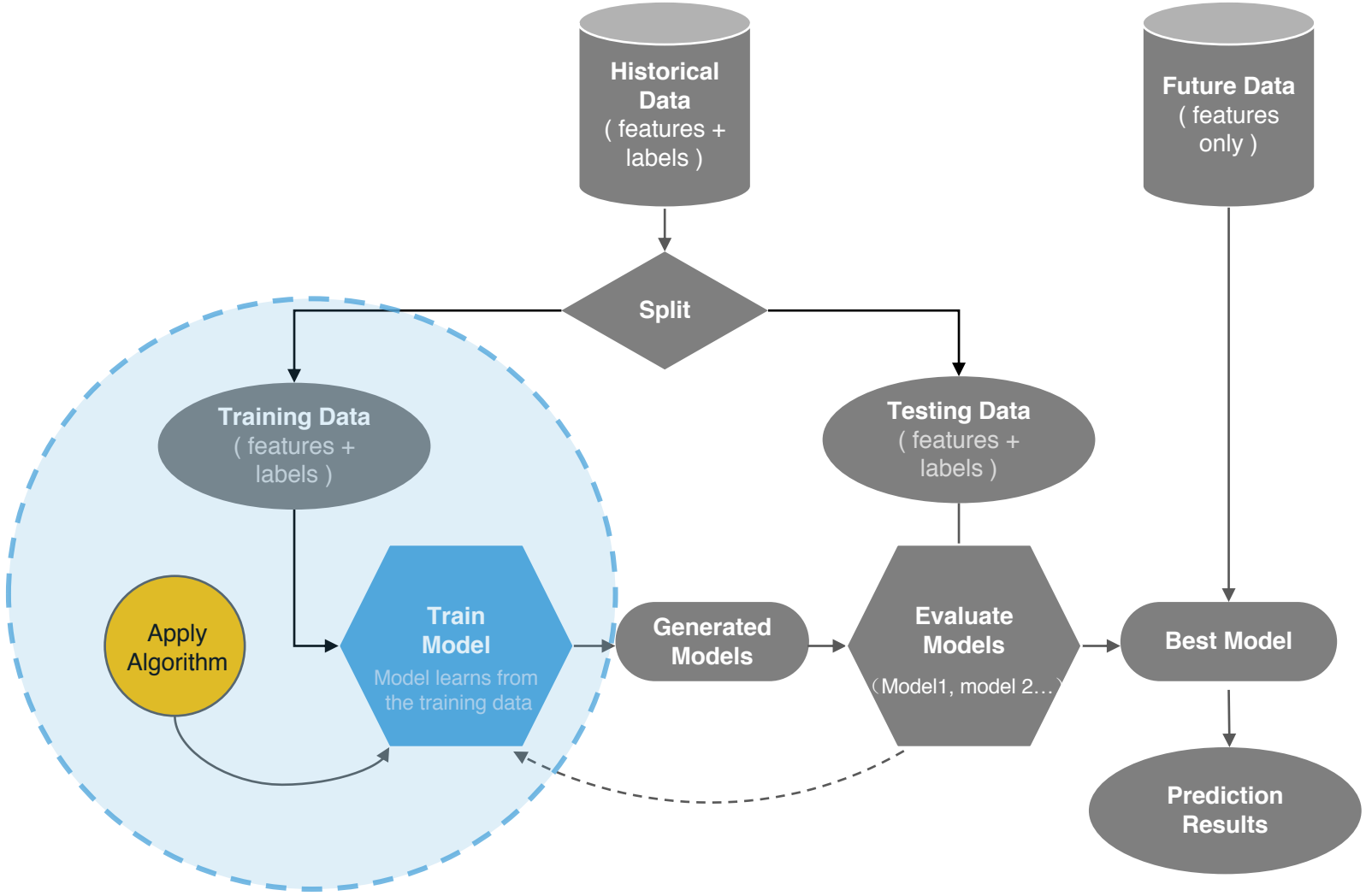
Predict Churn Probability

"Apply" the model onto new data to calculate the overall "score" or "probability" for each customer.

Source: SAP, https://open.sap.com/courses/ds1

:: Train the model

Let's take a closer look at the model training

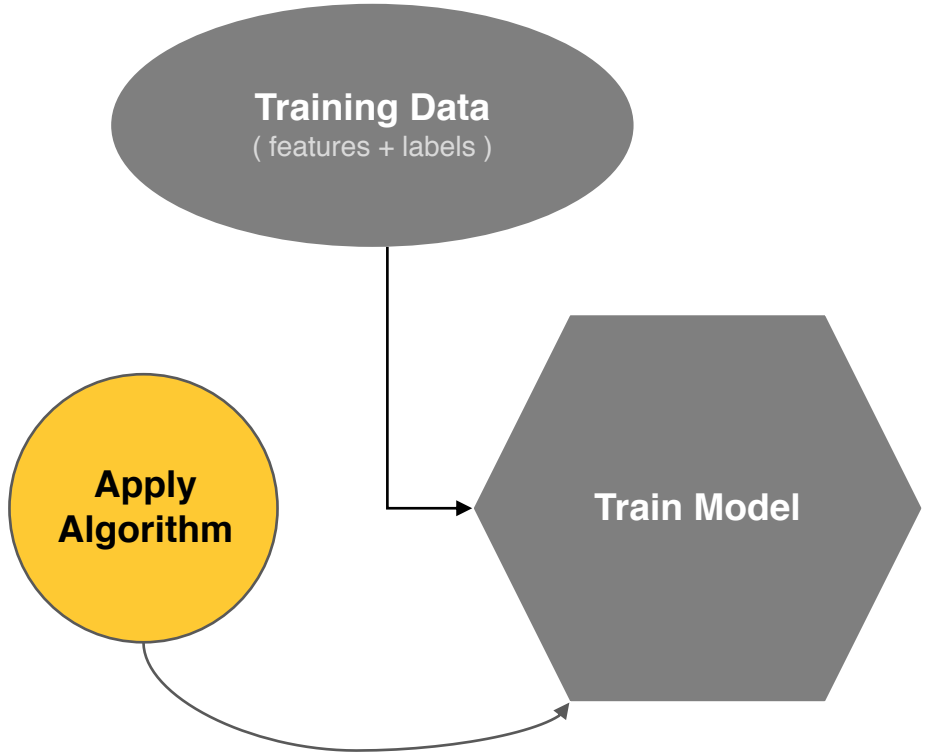


:: Train a model using algorithms and training data

Machine Learning uses algorithms to iteratively learn from data

Algorithm is a self-contained set of rules used to solve problems through data processing, math, or automated reasoning. Machine learning algorithms use computational methods to “learn” information directly from data without relying on a predetermined equation as a model. [1]

The algorithm will learn from the training data patterns that map the variables(features) to the target (i.e. labels), and it will output a model that captures these relationships [3].



1 source: Mathworks, Introducing Machine Learning
3 source: Amazon, http://docs.aws.amazon.com/machine-learning/latest/dg/training-the-model.html

:: Let’s take a look at an example for regression problem

Regression algorithm is a major types of machine learning algorithms

Let’s take a look at a very simple regression algorithm now.

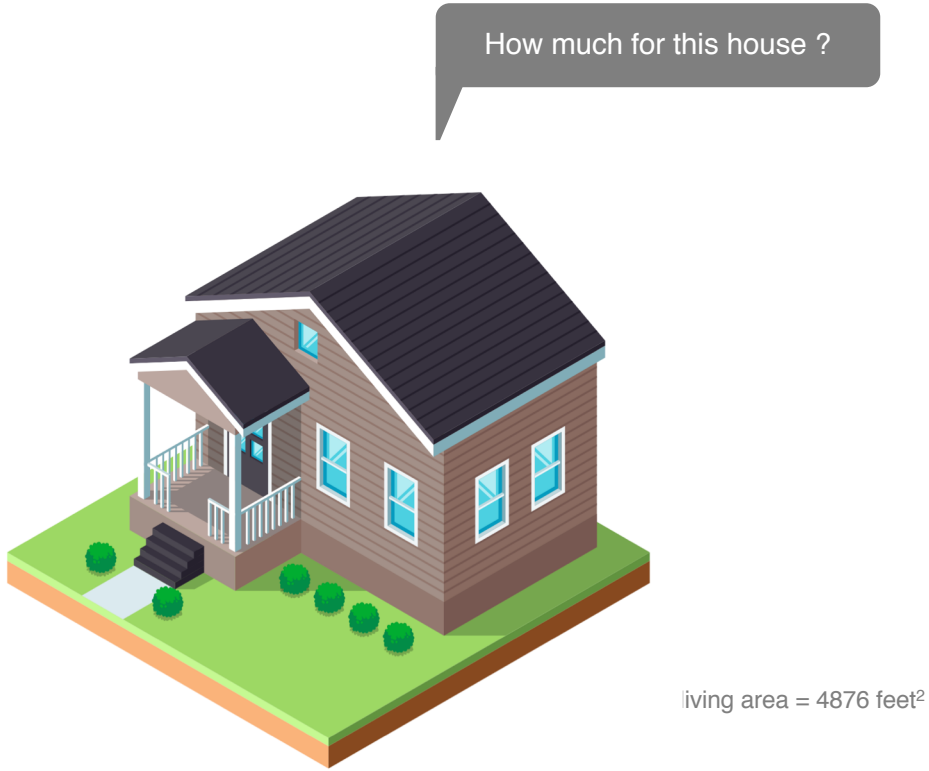
:: Let's take a look at an example for regression problem

Example: Predict house's price using linear regression

Suppose we have a dataset giving the living areas and prices of 21,613 houses from House Sales in King County, USA. Given data like this, we can learn to predict the prices of other houses in King County.

Living Area (Feet^2)	Price ($)
1180	221,900
2570	538,000
770	180,000
1960	604,000
1680	510,000
5420	1,225,000
1715	257,500
1060	291,850
1780	229,500
1890	323,000
3560	662,500
1160	468,000
1430	310,000
1370	400,000
1810	530,000
...	...
x	y

Data set

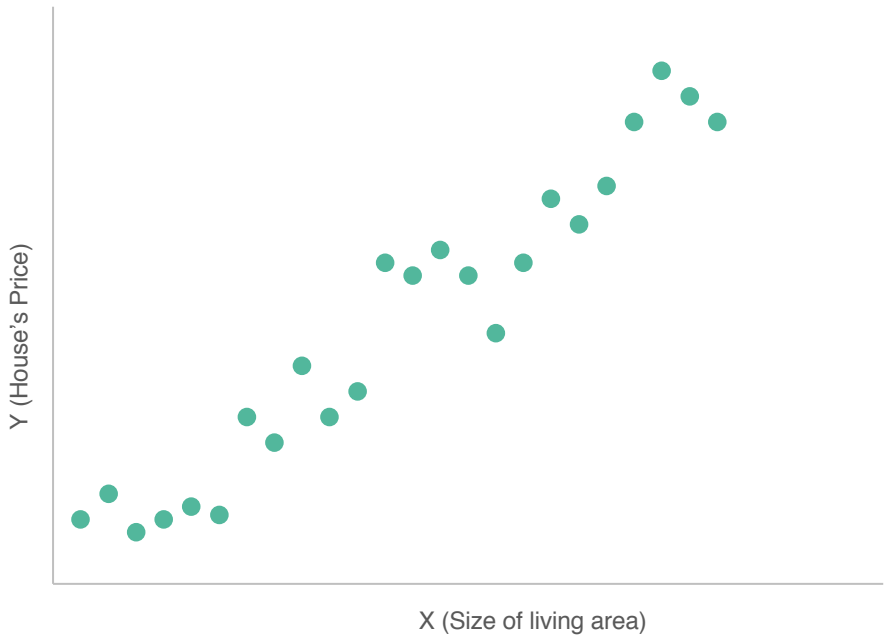


Picture of house: Designed by Freepik

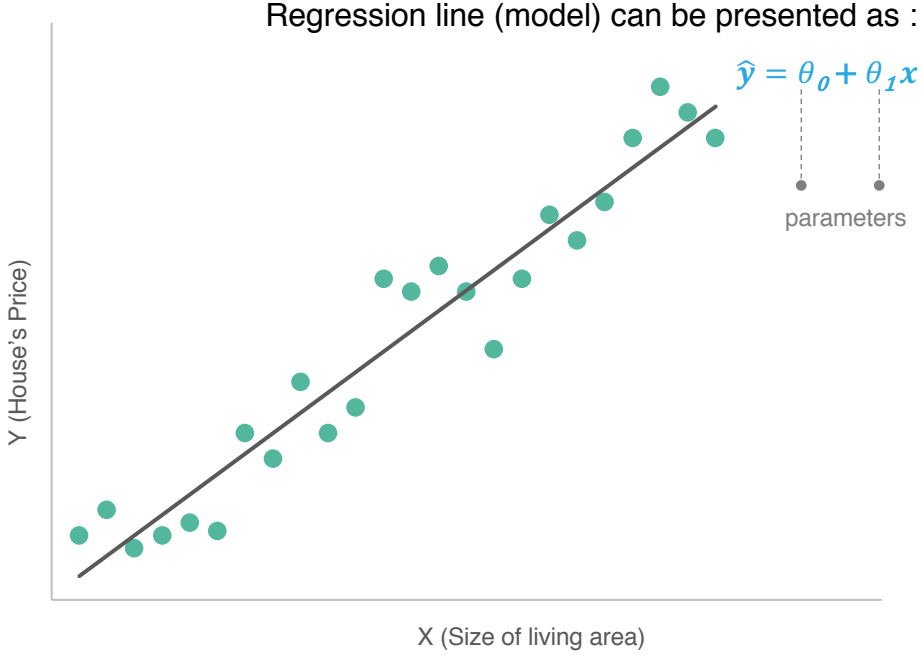
:: Example: Train a model using Linear regression algorithm

Use linear regression algorithm to approximate the relationship between x and y

Take linear regression as example, the algorithm is trying to find a best-fit line to represent the relationship between the input feature x and target y.



Regression line (model) can be presented as :



The straight line can be seen in the plot, showing how linear regression attempts to draw a straight line that will best minimize the residual sum of squares between the observed responses in the dataset, and the responses predicted by the linear approximation.

:: Sample python code for building a very simple model

```
import matplotlib.pyplot as plt
import numpy as np
from sklearn import datasets, linear_model

# Load the diabetes dataset
diabetes = datasets.load_diabetes()


# Use only one feature
diabetes_X = diabetes.data[:, np.newaxis, 2]

# Split the data into training/testing sets
diabetes_X_train = diabetes_X[:-20]
diabetes_X_test = diabetes_X[-20:]

# Split the targets into training/testing sets
diabetes_y_train = diabetes.target[:-20]
diabetes_y_test = diabetes.target[-20:]

# Create linear regression object
regr = linear_model.LinearRegression()

# Train the model using the training sets
regr.fit(diabetes_X_train, diabetes_y_train)

# The coefficients
print('Coefficients: \n', regr.coef_)
# The mean squared error
print("Mean squared error: %.2f"
      % np.mean((regr.predict(diabetes_X_test) - diabetes_y_test) ** 2))
# Explained variance score: 1 is perfect prediction
print('Variance score: %.2f' % regr.score(diabetes_X_test, diabetes_y_test))

# Plot outputs
plt.scatter(diabetes_X_test, diabetes_y_test,  color='black')
plt.plot(diabetes_X_test, regr.predict(diabetes_X_test), color='blue',
         linewidth=3)

plt.xticks(())
plt.yticks(())

plt.show()
```

- A model is built based on Linear regression algorithm.
- The model is trained using the training dataset

http://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html#sphx-glr-auto-examples-linear-model-plot-ols-py

Important:

Developing a model is a process of experimentation and incremental adjustment as much as it is of applying algorithms to a problem.

You should expect to spend a lot of time refining and modifying your model to get the best results.

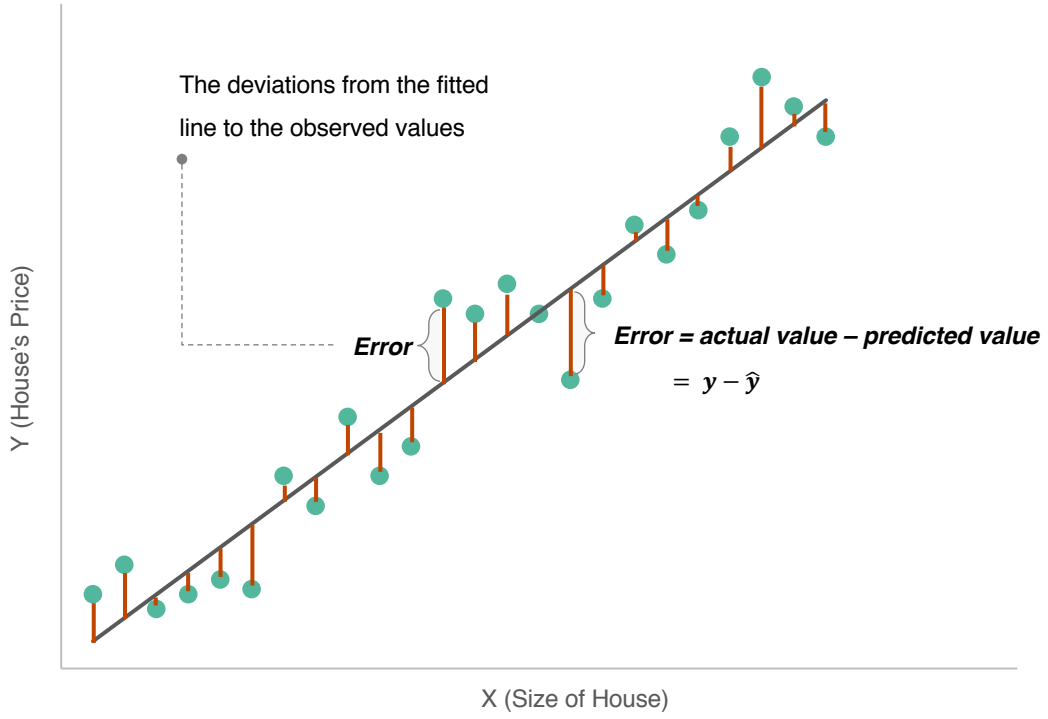
It's very important that you establish a threshold of success for your model before you begin because otherwise you may not know when to stop refining.

Source : https://cloud.google.com/ml-engine/docs/concepts/ml-solutions-overview

:: Usually there is deviation between the actual value and prediction

The deviations indicate how bad the model's prediction was on the training examples

Loss (i.e. error) is a number indicating how bad the model's prediction was on a single example. If the model's prediction is perfect, the loss is zero; otherwise, the loss is greater. [1]



Mean square error (MSE) is a commonly-used function to measure how large the loss is. It's called as **Loss function** or **Cost function**.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$\hat{y}_i$ *is the prediction*
y_i *is the actual value*

Mean square error (MSE) is the average squared loss per example over the whole dataset. We will elaborate the details of loss function later on.

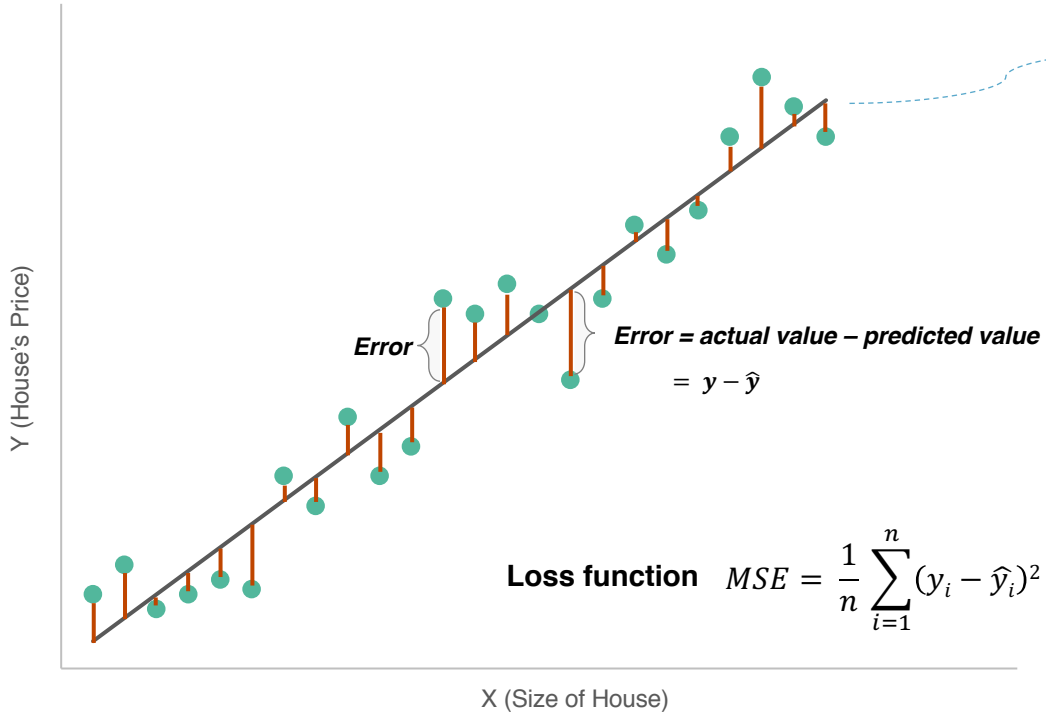
The smaller the Mean square error, the better the fit of the line to the data.

1, 2 Source: https://developers.google.com/machine-learning/crash-course/descending-into-ml/training-and-loss

:: Train the model to minimize the loss/error

Training the model is an iterative process of finding the “best” parameters to minimize the error

Training a model simply means learning (determining) good values for all the parameters of the model from labeled examples. In supervised learning, a machine learning algorithm builds a model by examining many examples and attempting to find a model that minimizes loss. [1]



Model:

$$\hat{y} = \theta_0 + \theta_1 x$$

parameter

The goal of training a model is to find a set of parameters that have low loss, on average, across all examples.[2]

In this linear regression example, the goal of the training is to find estimated value for the parameters θ_0 and θ_1 which would provide the “best” fit for the data points within the training set.

If you don’t understand the linear regression equation, don’t worry about it. You can find the detailed explanation in *Part 2 – popular algorithms*

1, 2 Source: https://developers.google.com/machine-learning/crash-course/descending-into-ml/training-and-loss

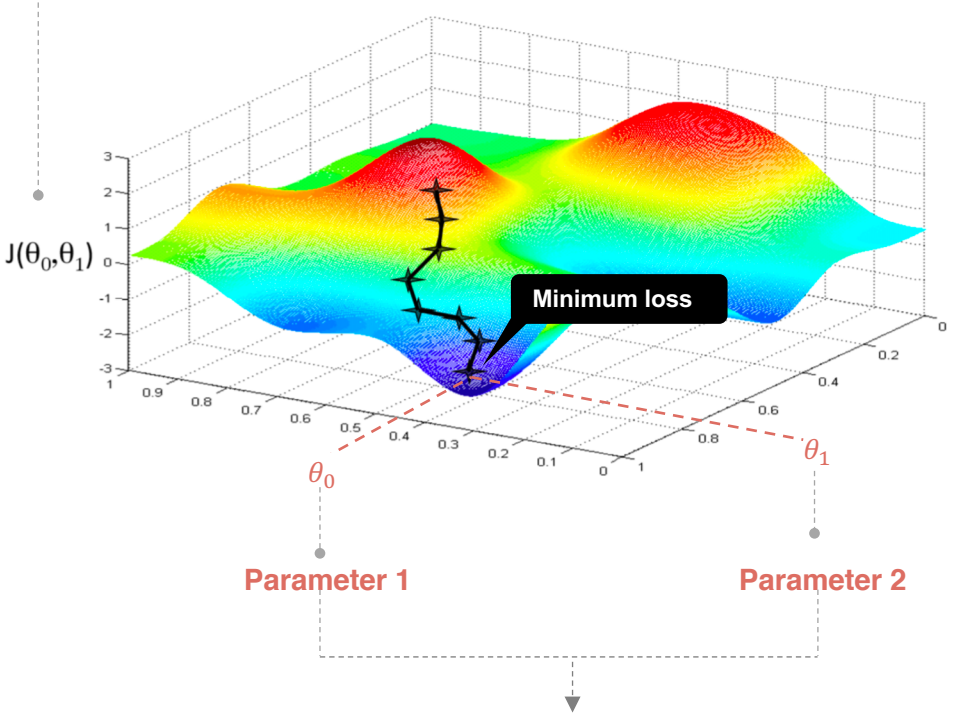
:: How does the model find the “best” parameters ?

Gradient Descent is one of the most common algorithms to find the good parameters

A Machine Learning model is trained by starting with an initial guess for the parameters (e.g. weights and bias in neural network) and iteratively adjusting those guesses until learning parameters with the lowest possible loss [1]

Loss function/cost function: $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

(error)



With these 2 specific parameter value, the loss (i.e. MSE) is almost smallest.

Usually, you iterate until overall loss stops changing or at least changes extremely slowly.

When that happens, we say that the model has **converged**.[2]

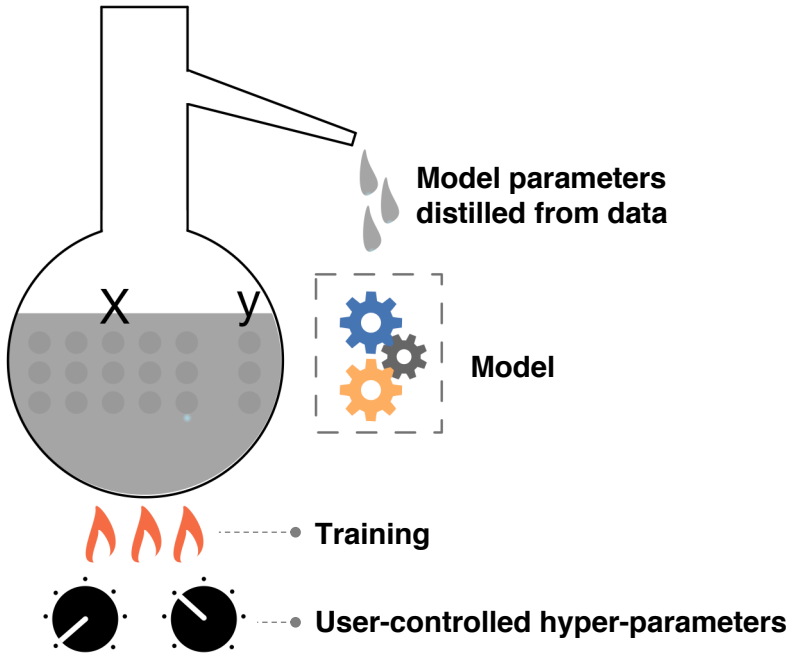
If you don’t get it, don’t worry. I will explain the details later in the *Part 2 – popular algorithms*

1,2 source: https://developers.google.com/machine-learning/crash-course/reducing-loss/an-iterative-approach

:: The model has parameters and hyper-parameters

Iteratively tuning the Hyperparameter so that the model can learn the “best” Parameters from data

The hyper-parameters are specified by the developer/data scientist while parameters are computed from the data via the algorithms.



- X is the feature vectors
- Y is the target variable

- Model’s parameters are the variables that your chosen machine learning technique uses to adjust to your data. They are internal to the model. They are estimated or learned from data. They are often not set manually by the practitioner.

- Hyperparameters control how a machine learning algorithm fits the model to the data. Hyper-parameters are specified by the programmer, not computed from the training data, and are often used to tune a model to improve accuracy for a particular data set.[1]

  The examples of hyper-parameters :
  - Number of layers, learning-rate in Neural network
  - Number of trees in Random forest

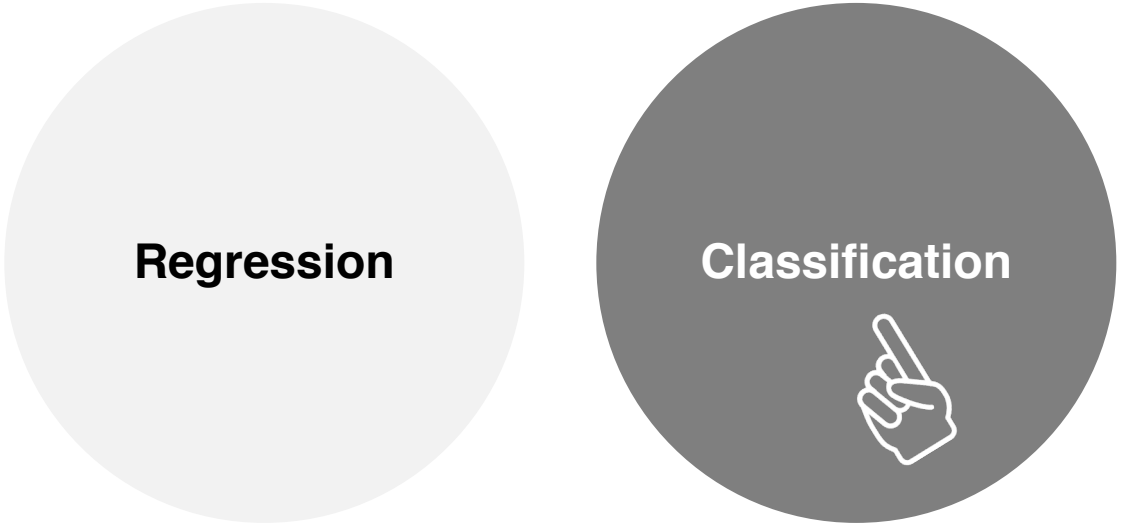
The detailed explanation of the difference will be elaborated later.

1 & Source of figure: A First Taste of Machine Learning, https://mlbook.explained.ai/intro.html#sec:2.2.3

:: Let’s take a look at another example for classification problem

Classification algorithm is another major types of machine learning algorithms

Just now we quickly go over the regression model example, now let’s move to the example of classification.

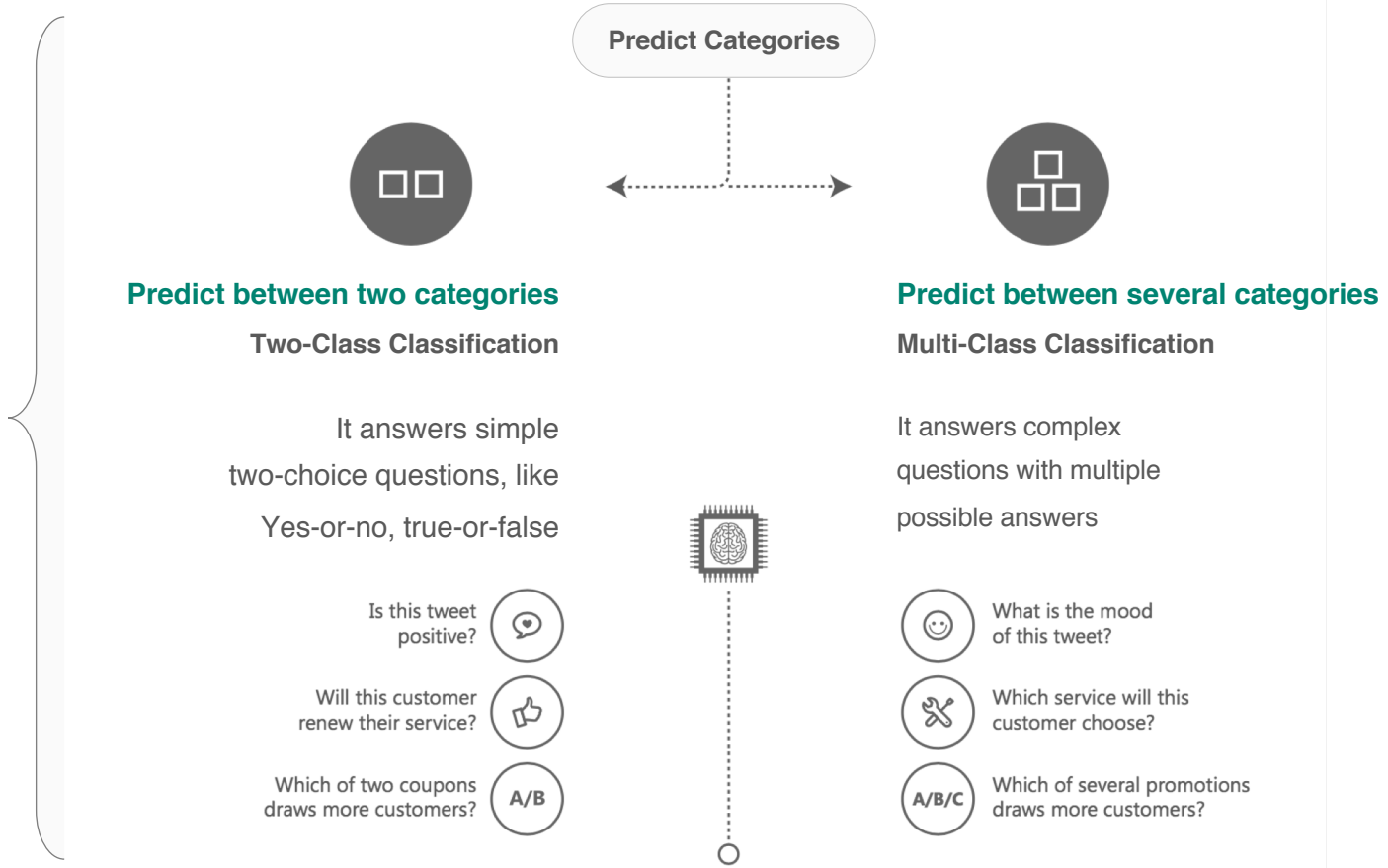


:: Let's take a look at an example for classification problem

Classification algorithm is another major types of machine learning algorithms

Classification Algorithm

Identify what category new information belongs in

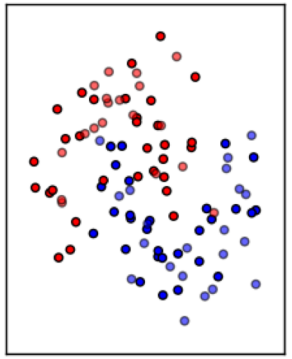


1 source: Mathworks, Introducing Machine Learning
Source of diagram: Microsoft

:: Example: how to classify the data points ?

How to classify this dataset into 2 categories ?

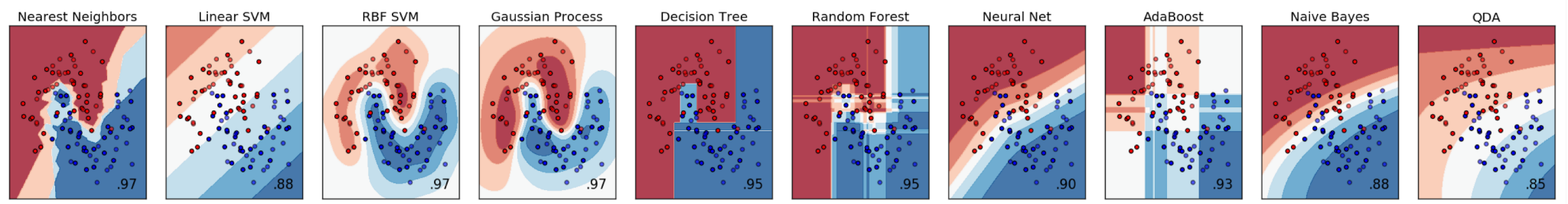
Input Data



How to classify this dataset into 2 categories ?

“red” and “blue”

Apply different algorithms on the same data set



The plots show training points in solid colors and testing points semi-transparent. The lower right shows the classification accuracy on the test set.

Source: Classifier comparison http://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html#