MODULE 1

Module 1:
Foundations Data Science, process, and tools

- Introduction to data science
- Properties of data, Asking interesting questions
- Classification of data science
- Data science process
- Collecting, cleaning and visualizing data,
- Languages, and models for data science

Introduction to data science

- Data science is a collection of techniques used to extract value from data.

- It has become an essential tool for any organization that collects, stores, and processes data as part of its operations.

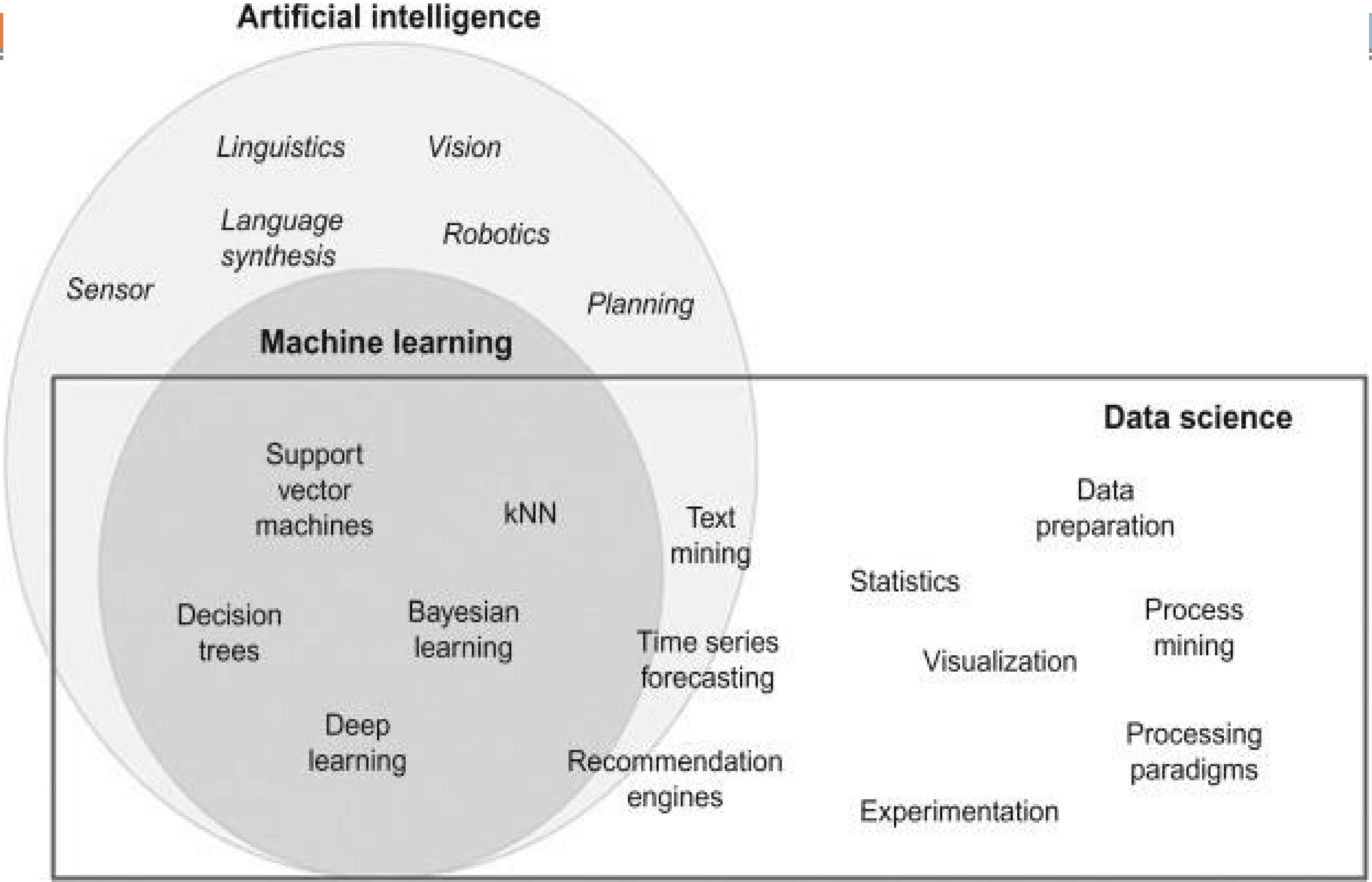
- Data science techniques rely on finding useful patterns, connections, and relationships within data.

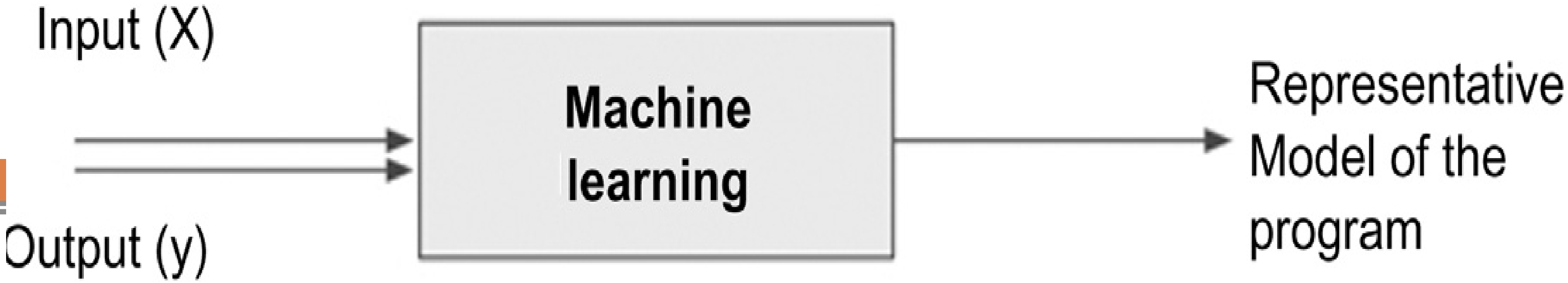
Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

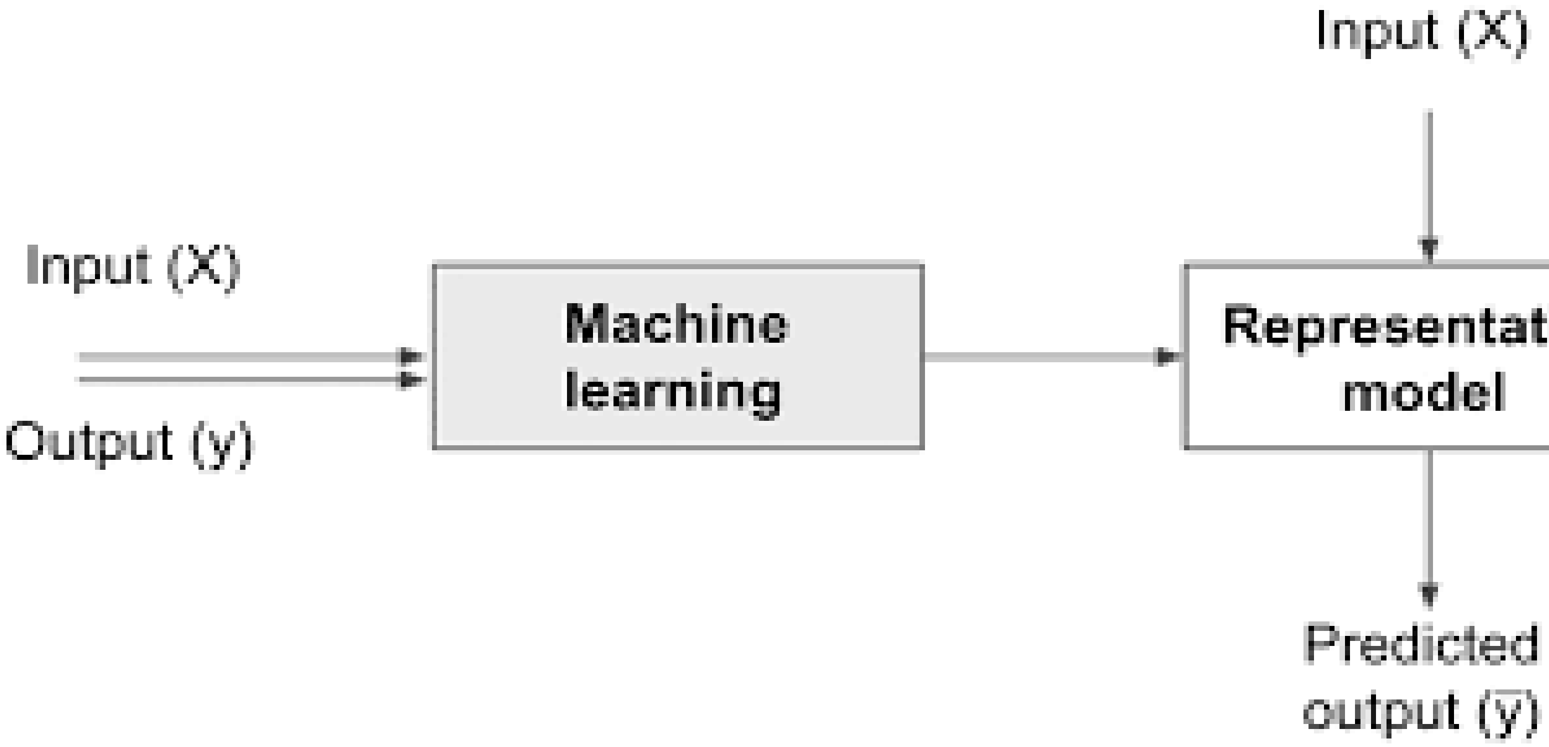
- Data science is also commonly referred to as
  - knowledge discovery,
  - machine learning,
  - predictive analytics,
  - data mining

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

AI, ML and DS


Artificial intelligence
Linguistics
Vision
Language synthesis
Robotics
Sensor
Planning
Machine learning
Data science
Support vector machines
kNN
Text mining
Data preparation
Statistics
Decision trees
Bayesian learning
Process mining
Time series forecasting
Visualization
Deep learning
Recommendation engines
Processing paradigms
Experimentation





Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Computer scientists, by nature, don't respect data

- Examples of the cultural differences between computer science and real science include:
  - Data vs. method centrism
  - Concern about results
  - Robustness
  - Precision

Reference: Skiena, S. S. (2017). The data science design manual., Springer.

Data vs. method centrism

- Scientists are data driven, while computer scientists are algorithm driven.

- Real scientists spend enormous amounts of effort collecting data to answer their question of interest

- computer scientists obsess about methods:
  - which algorithm is better than which other algorithm
  - which programming language is best for a job
  - which program is better than which other program.

Reference: Skiena, S. S. (2017). The data science design manual., Springer.

Concern about results

- Real scientists care about answers.
- They analyze data to discover something about how the world works
- Bad computer scientists worry about producing plausible looking numbers.
  - They are personally less invested in what can be learned from a computation, as opposed to getting it done quickly and efficiently

Reference: Skiena, S. S.
(2017). The data science
design manual., Springer.

Robustness

- Real scientists are comfortable with the idea that data has errors.
  - computer scientists are not

- Scientists think a lot about possible sources of bias or error in their data, and how these possible problems can effect the conclusions derived from them.

- Computer scientists chant “garbage in, garbage out"

Reference: Skiena, S. S. (2017). The data science design manual., Springer.

Precision

- Nothing is ever completely true or false in science
- Every- thing is either true or false in computer science or mathematics.
- Computer scientists care what a number is, while real scientists care what it means

Reference: Skiena, S. S. (2017). The data science design manual., Springer.

Asking Interesting Questions from Data

Asking Interesting Questions from Data

- What things might you be able to learn from a given data set?
- What do you/your people really want to know about the world?
- What will it mean to you once you find out?

Reference: Skiena, S. S. (2017). The data science design manual., Springer.

- The Baseball Encyclopedia

- The Internet Movie Database (IMDb)

- Google Ngrams
- New York Taxi Records
  - Prepare new questions from datasets (minimum 3)

Properties of Data

Properties of Data

- Structured vs. Unstructured Data
- Quantitative vs. Categorical Data
- Big Data vs. Little Data

Reference: Skiena, S. S. (2017). The data science design manual., Springer.

Structured vs. Unstructured Data

- □ Structured data

- □ Data sets are nicely structured, like the tables in a database or spreadsheet program

- □ Data is often represented by a matrix, where
  - the rows of the matrix represent distinct items or records
  - the columns represent distinct properties of these items.
  - For example, a data set about U.S. cities might contain one row for each city, with columns representing features like state, population, and area.

Reference: Skiena, S. S. (2017). The data science design manual., Springer.

- Unstructured data
  - Record information about the state of the world, but in a more heterogeneous way

  - Collection of tweets from Twitter
  - First step is build a matrix to structure it.
  - A bag of words model will construct a matrix with a row for each tweet, and a column for each frequently used vocabulary word.
  - Matrix entry M[i; j] then denotes the number of times tweet i contains word j.

Reference: Skiena, S. S. (2017). The data science design manual., Springer.

Quantitative vs. Categorical Data

- Quantitative data
  - Consists of numerical values, like height and weight.
  - Data can be incorporated directly into algebraic formulas and mathematical models, or displayed in conventional graphs and charts.

Reference: Skiena, S. S. (2017). The data science design manual., Springer.

- Categorical Data
  - Categorical data consists of labels describing the properties of the objects under investigation, like gender, hair color, and occupation.
  - This descriptive information can be every bit as precise and meaningful as numerical data, but it cannot be worked with using the same techniques.
  - Categorical data can usually be coded numerically. For example, gender might be represented as male = 0 or female = 1.
  - gray hair = 0, red hair = 1, and blond hair = 2.

Reference: Skiena, S. S. (2017). The data science design manual., Springer.

- We cannot really treat these values as numbers, for anything other than simple identity testing.

- Does it make any sense to talk about the maximum or minimum hair color?

- What is the interpretation of my hair color minus your hair color?

Reference: Skiena, S. S. (2017). The data science design manual., Springer.

Big Data vs. Little Data

- Big data
  - The analysis of massive data sets resulting from computer logs and sensor devices.
  - In principle, having more data is always better than having less, because you can always throw some of it away by sampling to get a smaller set if necessary
  - There are difficulties in working with large data sets.

Reference: Skiena, S. S. (2017). The data science design manual., Springer.

- The challenges of big data include:
  - The analysis cycle time slows as data size grows
    - Computational operations on data sets take longer as their volume increases.
  - Large data sets are complex to visualize
    - Plots with millions of points on them are impossible to display on computer screens or printed images, let alone conceptually understand.

Reference: Skiena, S. S. (2017). The data science design manual., Springer.

- Simple models do not require massive data to fit or evaluate
  - A typical data science task might be to make a decision (say, whether I should offer this fellow life insurance?) on the basis of a small number of variables: say age, gender, height, weight, and the presence or absence of existing medical conditions.

- Big data is sometimes called bad data
  - We might have to go to heroic efforts to make sense of something just because we have it.

Reference: Skiena, S. S. (2017). The data science design manual., Springer.

Classification of data science

Supervised Learning

- Data science problems can be broadly categorized into supervised or unsupervised learning models.

- Supervised or directed data science tries to infer a function or relationship based on labeled training data and uses this function to map new unlabeled data.

- Supervised techniques predict the value of the output variables based on a set of input variables.

- A model is developed from a training dataset where the values of input and output are previously known.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- The model generalizes the relationship between the input and output variables and uses it to predict for a dataset where only input variables are known.

- The output variable that is being predicted is also called a class label or target variable.

- Supervised data science needs a sufficient number of labeled records to learn the model from the data

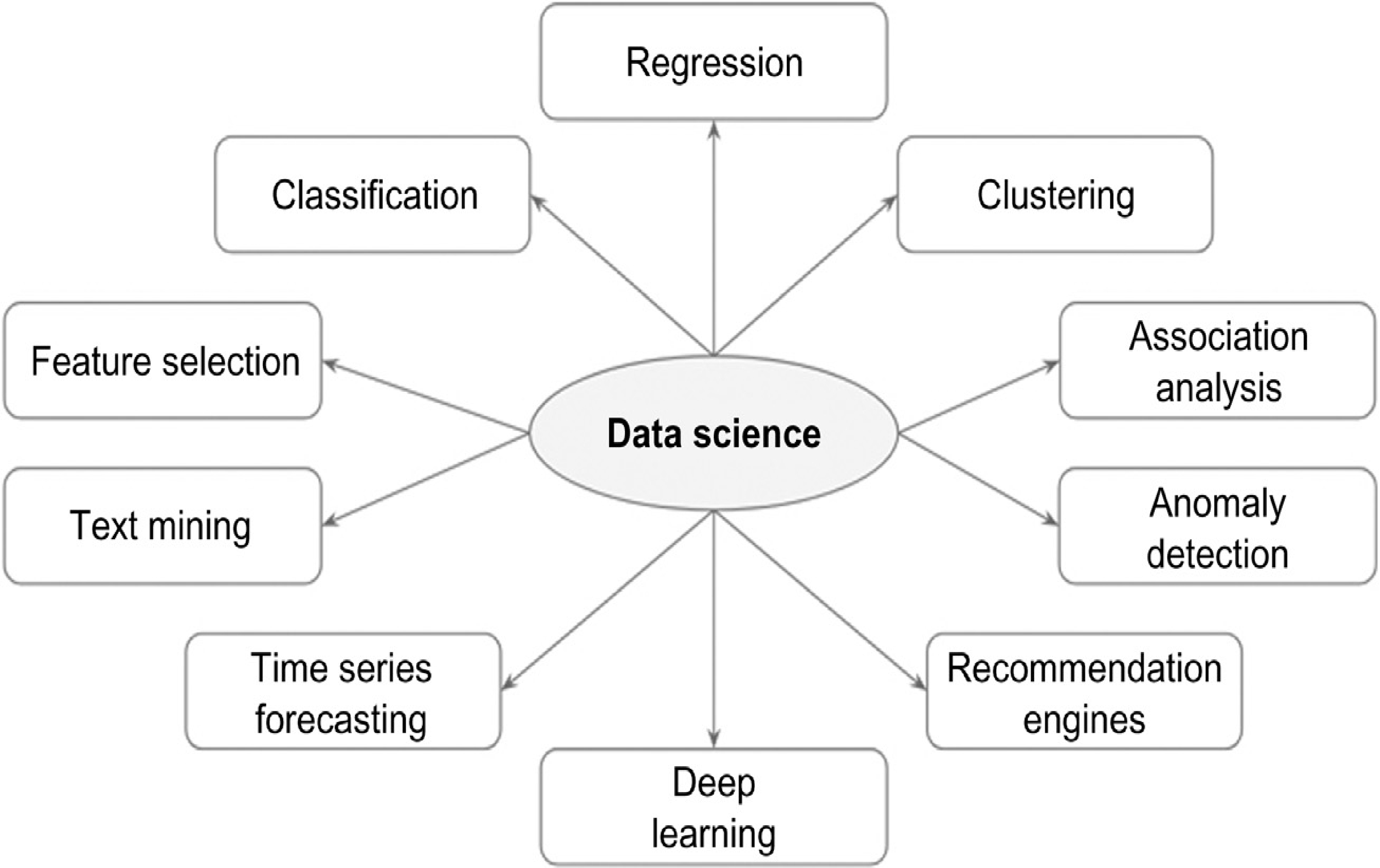
Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Unsupervised Learning

- Unsupervised or undirected data science uncovers hidden patterns in unlabeled data.
- There are no output variables to predict.
- Find patterns in data based on the relationship between data points themselves.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Data Science Tasks



Classification and Regression

- Classification and regression techniques predict a target variable based on input variables.

- The prediction is based on a generalized model built from a previously known dataset.

- In regression tasks, the output variable is numeric (e.g., the mortgage interest rate on a loan).

- Classification tasks predict output variables, which are categorical or polynomial (e.g., the yes or no decision to approve a loan)

- Deep learning is a more sophisticated artificial neural network used for classification and regression problems

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice. Morgan

Clustering

- Clustering is the process of identifying the natural groupings in a dataset
  - Generalize the uniqueness of each cluster

- Market basket analysis or Association analysis
  - Identify pairs of items that are purchased together, so that specific items can be bundled or placed next to each other.
  - commonly used in cross selling.

- Recommendation engines are the systems that recommend items to the users based on individual user preference

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Anomaly or outlier detection identifies the data points that are significantly different from other data points in a dataset
    - Credit card transaction fraud detection

- Time series forecasting is the process of predicting the future value of a variable (e.g., temperature) based on past historical values that may exhibit a trend and seasonality.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Text mining is a data science application where the input data is text,
  - which can be in the form of documents, messages, emails, or web pages
  - text file is converted to document vectors
  - standard data science tasks such as classification, clustering, etc., can be applied to vectors

- Feature selection is a process in which attributes in a dataset are reduced to a few attributes that really matter.(e.g. Height, Weight, Age, Eye color predict probability of heart disease)

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- A complete data science application can contain elements of both supervised and unsupervised technique

  - In marketing analytics, clustering can be used to find the natural clusters in customer records.
  - Each customer is assigned a cluster label at the end of the clustering process.
  - A labeled customer dataset can now be used to develop a model that assigns a cluster label for any new customer record with a supervised classification technique

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Data science tasks and examples page no. 25

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

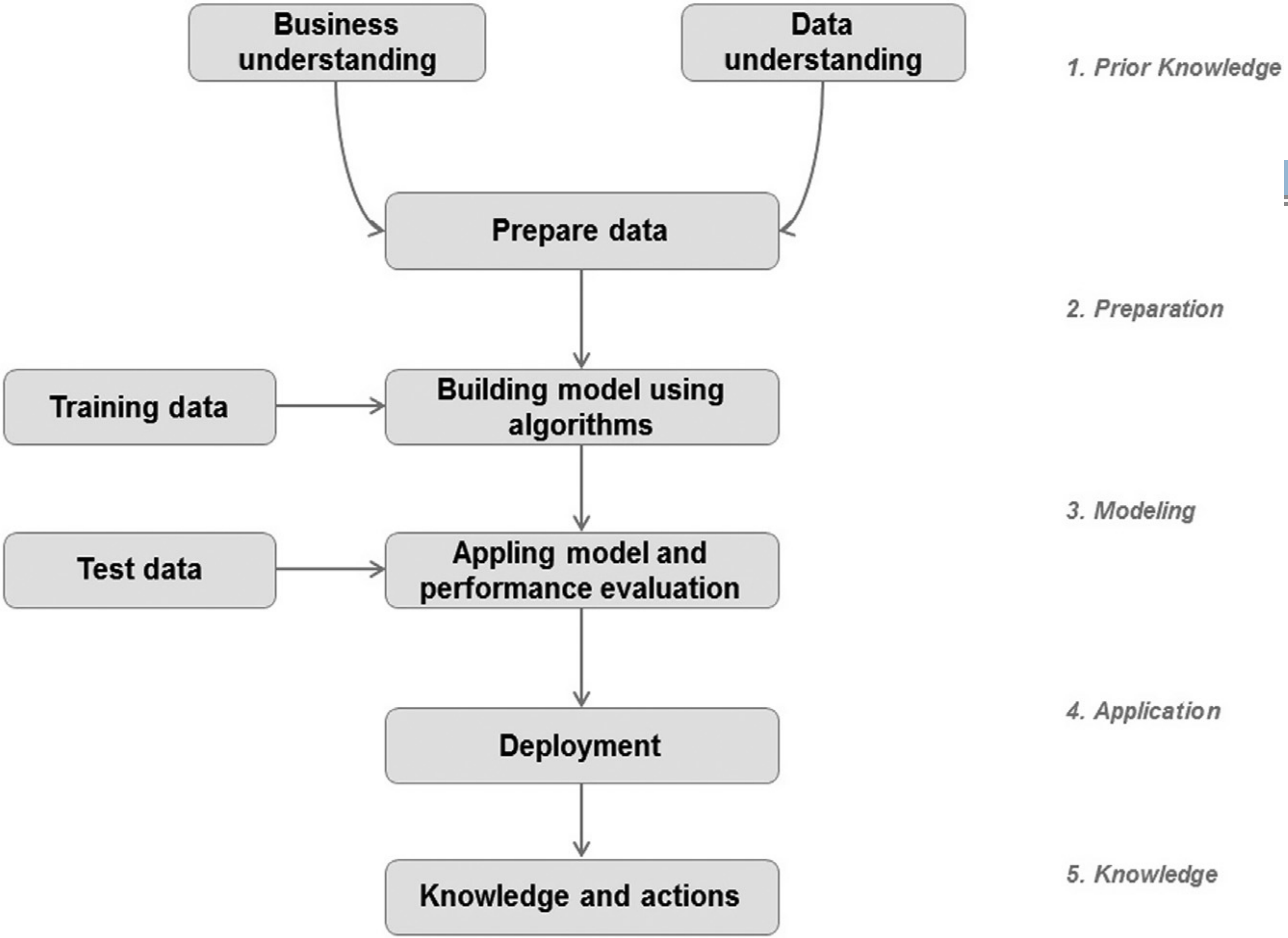
Data science process

Data science process

- The methodical discovery of useful relationships and patterns in data is enabled by a set of iterative activities collectively known as the data science process.

- The standard data science process involves
  - (1) understanding the problem
  - (2) preparing the data samples
  - (3) developing the model
  - (4) applying the model on a dataset to see how the model may work in the real world
  - (5) deploying and maintaining the models

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan



science: Concepts and practice., Morgan

Modeling

- It is the process of building representative models that can be inferred from the sample dataset which can be used for
  - Either predicting (predictive modeling) or
  - Describing the underlying pattern in the data (descriptive or explanatory modeling).

- There are many data science tools, that can automate the model building.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Most time-consuming part of the overall data science process is the preparation of data, followed by data and business understanding.

- Crucial to the success of the data science process
  - Asking the right business question
  - Gaining in-depth business understanding
  - Sourcing and preparing the data for the data science task
  - Mitigating implementation considerations
  - Integrating the model into the business process
  - Gaining knowledge from the dataset

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Business Understanding

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

PRIOR KNOWLEDGE

- Prior knowledge refers to information that is already known about a subject

- The data science problem doesn't emerge in isolation
  - it always develops on top of existing subject matter and contextual information that is already known.
  - The prior knowledge step in the data science process helps to
    - **define what problem is being solved**
    - **how it fits in the business context, and**
    - **What data is needed in order to solve the problem**

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Objective

- The data science process starts with a need for analysis, a question, or a business objective

- Without a well-defined statement of the problem, it is impossible to come up with the right dataset and pick the right data science algorithm

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Consumer loan business
  - a loan is provisioned for individuals against the collateral of assets like a home or car, that is, a mortgage or an auto loan
  - an important component of the loan, is the interest rate at which the borrower repays the loan on top of the principal.
  - The interest rate on a loan depends on a gamut of variables like
    - the current federal funds rate as determined by the central bank,
    - borrower’s credit score, income level,
    - home value, initial down payment amount,
    - current assets and liabilities of the borrower, etc.

- The business objective of this hypothetical case is:
  - If the interest rate of past borrowers with a range of credit scores is known, can the interest rate for a new borrower be predicted?

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Subject Area

- The process of data science uncovers hidden patterns in the dataset by exposing relationships between attributes

- It is up to the practitioner to sift through the exposed patterns and accept the ones that are valid and relevant to the answer of the objective question

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- The lending business
  - The objective is to predict the lending interest rate
    - Important to know how the lending business works,
    - why the prediction matters,
    - what happens after the rate is predicted,
    - what data points can be collected from borrowers,
    - what data points cannot be collected because of the external regulations and the internal policies,
    - what other external factors can affect the interest rate,
    - how to verify the validity of the outcome

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Data

- Prior knowledge in the data can also be gathered.

- Understanding how the data is collected, stored, transformed, reported, and used is essential to the data science process

- Surveys all the data available to answer the business question and narrows down the new data that need to be sourced.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- There are quite a range of factors to consider:
  - quality of the data, quantity of data,
  - availability of data, gaps in data,
  - does lack of data compel the practitioner to change the business question

- The objective of this step is to come up with a dataset to answer the business question through the data science process.

- It is critical to recognize that <u>an inferred model is only as good as the data used to create it</u>.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Dataset

- A dataset (example set) is a collection of data with a defined structure
- This structure is also sometimes referred to as a “data frame”.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Dataset Example

Borrower ID	Credit Score	Interest Rate (%)
01	500	7.31
02	600	6.70
03	700	5.95
04	700	6.40
05	800	5.40
06	800	5.70
07	750	5.90
08	550	7.00
09	650	6.50
10	825	5.70

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Data point

□ A data point (record, object or example) is a single instance in the dataset.

□ Each row in dataset table is a data point.

□ Each instance contains the same structure as the dataset.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Attribute

- An attribute (feature, input, dimension, variable, or predictor) is a single property of the dataset.
- Each column is an attribute.
- Attributes can be numeric, categorical, date-time, text, or Boolean data types.
  - Both the credit score and the interest rate are numeric attributes

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Label

- □ A label (class label, output, prediction, target, or response) is the special attribute to be predicted based on all the input attributes.

  - The interest rate is the output variable

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Identifiers

- Identifiers are special attributes that are used for locating or providing context to individual records.
  - names, account numbers, and employee ID numbers
- Identifiers are often used as lookup keys to join multiple datasets.
- They bear no information that is suitable for building data science models and should, thus, be excluded for the actual modeling
  - Borrower ID is an identifier

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

DATA PREPARATION

- Preparing the dataset to suit a data science task is the most time-consuming part of the process

- Data to be structured in a tabular format with records in the rows and attributes in the columns.

- If the data is in any other format, the data would need to be transformed by applying pivot, type conversion, join, or transpose functions, etc., to condition the data into the required structure.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Data Exploration

- Also known as exploratory data analysis, provides a set of simple tools to achieve basic understanding of the data.

- Data exploration approaches involve computing descriptive statistics and visualization of data.

- They can expose the structure of the data, the distribution of the values, the presence of extreme values, and highlight the inter-relationships within the dataset.

- Descriptive statistics like mean, median, mode, standard deviation, and range for each attribute provide an easily readable summary of the key characteristics of the distribution of data.

- A visual plot of data points provides an instant grasp of all the data points condensed into one chart

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Data Quality

- Errors in data will impact the representativeness of the model.

- Organizations use data alerts, cleansing, and transformation techniques to improve and manage the quality of the data and store them in companywide repositories called data warehouses.

- Data sourced from well-maintained data warehouses have higher quality, as there are proper controls in place to ensure a level of data accuracy for new and existing data.

- The data cleansing practices include elimination of duplicate records, quarantining outlier records that exceed the bounds, standardization of attribute values, substitution of missing values, etc

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Missing Values

- Understand the reason behind why the values are missing.

- Tracking the data lineage (provenance) of the data source can lead to the identification of systemic issues during data capture or errors in data transformation.

- Knowing the source of a missing value will often guide which mitigation methodology to use.

- The missing value can be substituted with a range of artificial data so that the issue can be managed with marginal impact on the later steps in the data science process

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- This method is useful if the missing values occur randomly and the frequency of occurrence is quite rare.

- Alternatively, to build the representative model, all the data records with missing values or records with poor data quality can be ignored.

- This method reduces the size of the dataset

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Some data science algorithms are good at handling records with missing values, while others expect the data preparation step to handle it before the model is inferred.

  - k-nearest neighbor (k-NN) algorithm for classification tasks are often robust with missing values.

  - Neural network models for classification tasks do not perform well with missing attributes, and thus, the data preparation step is essential for developing neural network models.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Data Types and Conversion

- The attributes in a dataset can be of different types
  - Continuous numeric (interest rate)
  - Integer numeric (credit score)
  - Categorical (poor, good, excellent– credit score)

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- In case of linear regression models, the input attributes have to be numeric.

- If the available data are categorical, they must be converted to numeric attribute.

- A specific numeric score can be encoded for each category value, such as poor (5400), good (5600), excellent (5700), etc.

- Numeric values can be converted to categorical data types by a technique called binning
  - A range of values are specified for each category, for example, a score between 400 and 500 can be encoded as “low”

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Transformation

- □ Input attributes are expected to be numeric and normalized for algorithms like KNN

- □ The algorithm compares the values of different attributes and calculates distance between the data points.

- □ Normalization prevents one attribute dominating the distance results because of large values.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- For example, consider income (in thousands) and credit score (in hundreds).

- The distance calculation will always be dominated by slight variations in income.

- One solution is to convert the range of income and credit score to a more uniform scale from 0 to 1 by normalization.

- This way, <u>a consistent comparison can be made between the two different attributes with different units</u>

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Outliers

- Outliers are anomalies in a given dataset.

- Outliers may occur because of
  - correct data capture or
    - (few people with income in tens of millions)
  - erroneous data capture
    - (human height as 1.73 cm instead of 1.73 m)

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- The presence of outliers needs to be understood and will require special treatments.

- The purpose of creating a representative model is to generalize a pattern or a relationship within a dataset and the presence of outliers skews the representativeness of the inferred model.

- Detecting outliers may be the primary purpose of some data science applications, like fraud or intrusion detection.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Feature Selection

- Many data science problems involve a dataset with hundreds to thousands of attributes

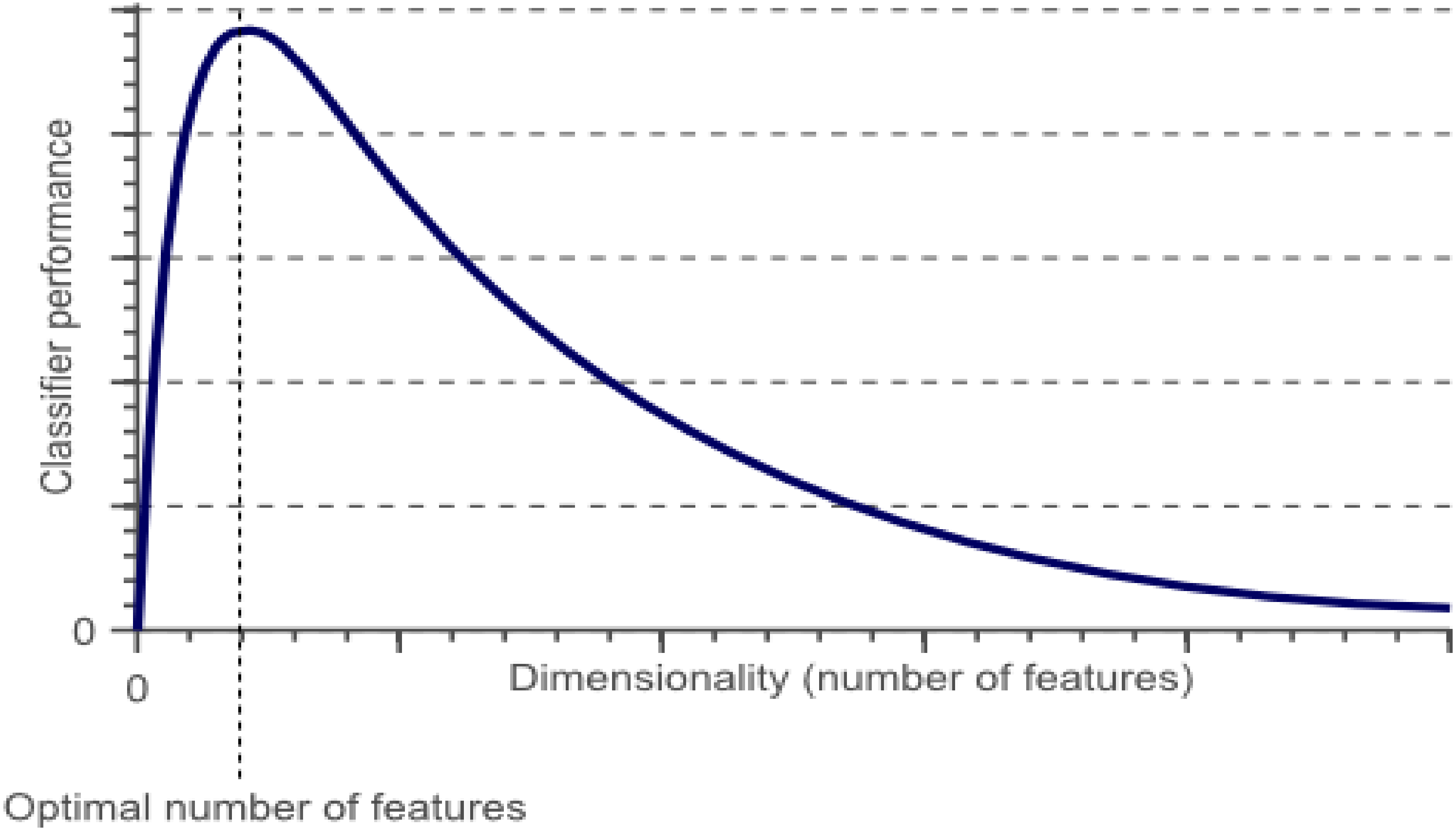
- A large number of attributes in the dataset significantly increases the complexity of a model and may degrade the performance of the model due to the curse of dimensionality.

- Reducing the number of attributes, without significant loss in the performance of the model, is called feature selection.

- It leads to a more simplified model

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Curse of Dimensionality



Data Sampling

- Sampling is a process of selecting a subset of records as a representation of the original dataset for use in data analysis or modeling.

- The sample data serve as a representative of the original dataset with similar properties, such as a similar mean.

- Sampling reduces the amount of data that need to be processed and speeds up the build process of the modeling

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and
practice., Morgan

- Stratified sampling is a process of sampling where each class is equally represented in the sample; this allows the model to focus on the difference between the patterns of each class that is, normal and outlier records.

- In classification applications, sampling is used create multiple base models, each developed using a different set of sampled training datasets.

- These base models are used to build one meta model, called the ensemble model, where the error rate is improved when compared to that of the base models.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

MODELING

- A model is the abstract representation of the data and the relationships in a given dataset.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Modelling steps

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Training and Testing Datasets

- The modeling step creates a representative model inferred from the data.
- The Dataset used to create the model, with known attributes and target, is called the training dataset.
- The validity of the created model will also need to be checked with another known dataset called the test dataset or validation dataset.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- To facilitate this process, the overall known dataset can be split into a training dataset and a test dataset.

- A standard rule of thumb is two-thirds of the data are to be used as training and one-third as a test dataset

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Learning Algorithms

- The business question and the availability of data will dictate what data science task (association, classification, regression, etc.,) can to be used
  - Interest rate prediction is a regression problem.
  - A simple linear regression technique will be used to model and generalize the relationship between credit score and interest rate.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Evaluation of the Model

- A model should not memorize and output the same values that are in the training records.
- The phenomenon of a model memorizing the training data is called overfitting.
- An overfitted model just memorizes the training records and will underperform on real unlabeled new data.
- The model should generalize or learn the relationship between credit score and interest rate.
- To evaluate this relationship, the validation or test dataset, which was not previously used in building the model, is used for evaluation

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- The actual value of the interest rate can be compared against the predicted value using the model, and thus, the prediction error can be calculated.

- As long as the error is acceptable, this model is ready for deployment.

- The error rate can be used to compare this model with other models developed using different algorithms like neural networks or Bayesian models, etc.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Ensemble Modeling

- Ensemble modeling is a process where multiple diverse base models are used to predict an outcome.
- The motivation for using ensemble models is to reduce the generalization error of the prediction.
- As long as the base models are diverse and independent, the prediction error decreases when the ensemble approach is used.
- The approach seeks the wisdom of crowds in making a prediction.
- Even though the ensemble model has multiple base models within the model, it acts and performs as a single model.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- At the end of the modeling stage of the data science process, one has
  - (1) analyzed the business question;
  - (2) sourced the data relevant to answer the question;
  - (3) selected a data science technique to answer the question;
  - (4) picked a data science algorithm and prepared the data to suit the algorithm;
  - (5) split the data into training and test datasets;
  - (6) built a generalized model from the training dataset; and
  - (7) validated the model against the test dataset

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

APPLICATION

- Deployment is the stage at which the model becomes production ready or live.

- In business applications, the results of the data science process have to be assimilated into the business process—usually in software applications.

- The model deployment stage has to deal with:
  - assessing model readiness,
  - technical integration,
  - response time,
  - model maintenance, and
  - assimilation.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Collecting, cleaning and visualizing data

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Collecting Data

- The most critical issue in any data science or modeling project is finding the right data set.
  - Who might actually have the data I need?
  - Why might they decide to make it available to me?
  - How can I get my hands on it?

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Hunting

- Scrapping

- Logging

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Hunting

- Who has the data, and how can you get it?

- Companies and Proprietary Data Sources
- Government Data Sources
- Academic Data Sets

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Companies and Proprietary Data Sources

- Large companies like Facebook, Google, Amazon, American Express, and Blue Cross have amazing amounts of exciting data about users and transactions

- Companies are reluctant to share data for two good reasons:
  - Business issues, and the fear of helping their competition.
  - Privacy issues, and the fear of offending their customers

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Many responsible companies like The New York Times, Twitter, Facebook, and Google do release certain data
  - Providing customers and third parties with data that can increase sales.
  - For example, releasing data about query frequency and ad pricing can encourage more people to place ads on a given platform.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Most organizations have internal data sets of relevance to their business.
- As an employee, you should be able to get privileged access while you work there.
- Be aware that companies have internal data access policies, so you will still be subject to certain restrictions.
- Violating the terms of these policies is an excellent way to become an ex-employee

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Government Data Sources

- City, state, and federal governments have become increasingly committed to open data, to facilitate novel applications and improve how government can fullfill its mission

- Government data differs from industrial data in that, it belongs to the People.

- The Freedom of Information Act (FOI) enables any citizen to make a formal request for any government document or data set.

- Such a request triggers a process to determine what can be released without compromising the national interest or violating privacy.

- Preserving privacy is typically the biggest issue in deciding whether a particular government data set can be released.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Academic Data Sets

- An increasing fraction of academic research involves the creation of large data sets.
- Many journals now require making source data available to other researchers prior to publication.
- Expect to be able to find vast amounts of economic, medical, demographic, historical, and scientific data

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- The key to finding these data sets is to track down the relevant papers

- Google Scholar is the most accessible source of research publications.

- Research publications will typically provide pointers to where its associated data can be found.

- If not, contacting the author directly with a request should quickly yield the desired result

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- someone else has worked hard to analyze Published data sets before you got to them

- But bringing fresh questions to old data generally opens new possibilities.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Scraping

- Web pages often contain valuable text and numerical data

- Spidering is the process of downloading the right set of pages for analysis.

- Scraping is the fine art of stripping this content from each page to prepare it for computational analysis.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Scraping programs were site-specific scripts hacked up to look for particular HTML patterns flanking the content of interest

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- The most advanced form of spidering is web crawling
  - where you systematically traverse all outgoing links from a given root page
  - continuing recursively until you have visited every page on the target website.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Every major website contains a terms of service document that restricts what you can legally do with any associated data

- Aaron Schwartz case

- If you are attempting a web-scraping project professionally, be sure that management understands the terms of service before you get too creative with someone else's property.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Logging

- Internal access to a web service, communications device, or laboratory instrument grants you the right and responsibility to log all activity for downstream analysis.

- Amazing things can be done with ambient data collection from weblogs and sensing devices

- The accelerometers in cell phones can be used to measure the strength of earthquakes
  - Filter out people driving on bumpy roads or leaving their phones in a clothes dryer

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Monitoring the GPS data of a fleet of taxi cabs tracks traffic congestion on city streets.

- Computational analysis of image and video streams opens the door to countless applications.

- Another cool idea is to use cameras as weather instruments, by looking at the color of the sky in the background of the millions of photographs uploaded to photo sites daily.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Logging

- The important considerations in designing any logging system are:
  - Build it to endure with limited maintenance. Set it and forget it, by provisioning it with enough storage for unlimited expansion, and a backup.
  - Store all fields of possible value, without going crazy.
  - Use a human-readable format or transactions database, so you can understand exactly what is in there when the time comes, months or years later, to sit down and analyze your data

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Cleaning Data

- Garbage in, garbage out

- Processing before we do our real analysis, to make sure that the garbage never gets in in the first place

- Errors vs. Artifacts

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Errors vs. Artifacts

- Under ancient Jewish law, if a suspect on trial was unanimously found guilty by all judges, then this suspect would be acquitted.
- Unanimous agreement often indicates the presence of a systemic error in the judicial process.
- When something seems too good to be true, a mistake has likely been made somewhere.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Data errors represent information that is fundamentally lost in acquisition

- Gaussian noise blurring the resolution of our sensors represents error, precision which has been permanently lost.

- The two hours of missing logs because the server crashed represents data error: it is information which cannot be reconstructed again.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Artifacts are systematic problems arising from processing done to the raw information it was constructed from

- Processing artifacts can be corrected, so long as the original raw data set remains available.

- These artifacts must be detected before they can be corrected.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- The key to detecting processing artifacts is the “sniff test,“

- Something bad is usually something unexpected or surprising.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Models for data science

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Modeling

- The process of encapsulating information into a tool which can forecast and make predictions
- Predictive models are structured around some idea of what causes future events to happen.
- Extrapolating from recent trends and observations assumes a world view that the future will be like the past.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Occam's Razor

- Occam's razor is the philosophical principle that the "simplest explanation is the best explanation".

- Given two models or theories which do an equally accurate job of making predictions, we should opt for the simpler one as sounder and more robust.

- It is more likely to be making the right decision for the right reasons.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Occam’s notion of simpler generally refers to reducing the number of assumptions employed in developing the model.

- Minimize the parameter count of a model.
  - Overfitting occurs when a model tries too hard to achieve accurate performance on its training data.
  - This happens when there are so many parameters that the model can essentially memorize its training set, instead of generalizing appropriately to minimize the effects of error and outliers.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Accuracy is not the best metric to use in judging the quality of a model.

- Simpler models tend to be more robust and understandable than complicated alternatives

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Bias-Variance Trade-Offs

- The bias error is an error from erroneous assumptions in the learning algorithm.
  - High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).
- The variance is an error from sensitivity to small fluctuations in the training set.If our training set contains sampling or measurement error, this noise introduces variance into the resulting model.
  - High variance may result from an algorithm modeling the random noise in the training data (overfitting).

- Errors of bias produce underfit models.
  - They do not fitt the training data as tightly as possible

- Errors of variance result in overfit models:
  - their quest for accuracy causes them to mistake noise for signal

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

A Taxonomy of Models

- Linear vs. Non-Linear Models
- Blackbox vs. Descriptive Models
- First-Principle vs. Data-Driven Models
- Stochastic vs. Deterministic Models
- Flat vs. Hierarchical Models

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Linear vs. Non-Linear Models

- Linear models are governed by equations that weigh each feature variable by a coefficient reflecting its importance, and sum up these values to produce a score.

- Powerful machine learning techniques, such as linear regression, can be used to identify the best possible coefficients to fit training data

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- World is not linear.
- Richer mathematical descriptions include higher-order polynomials, logarithms, and exponentials.

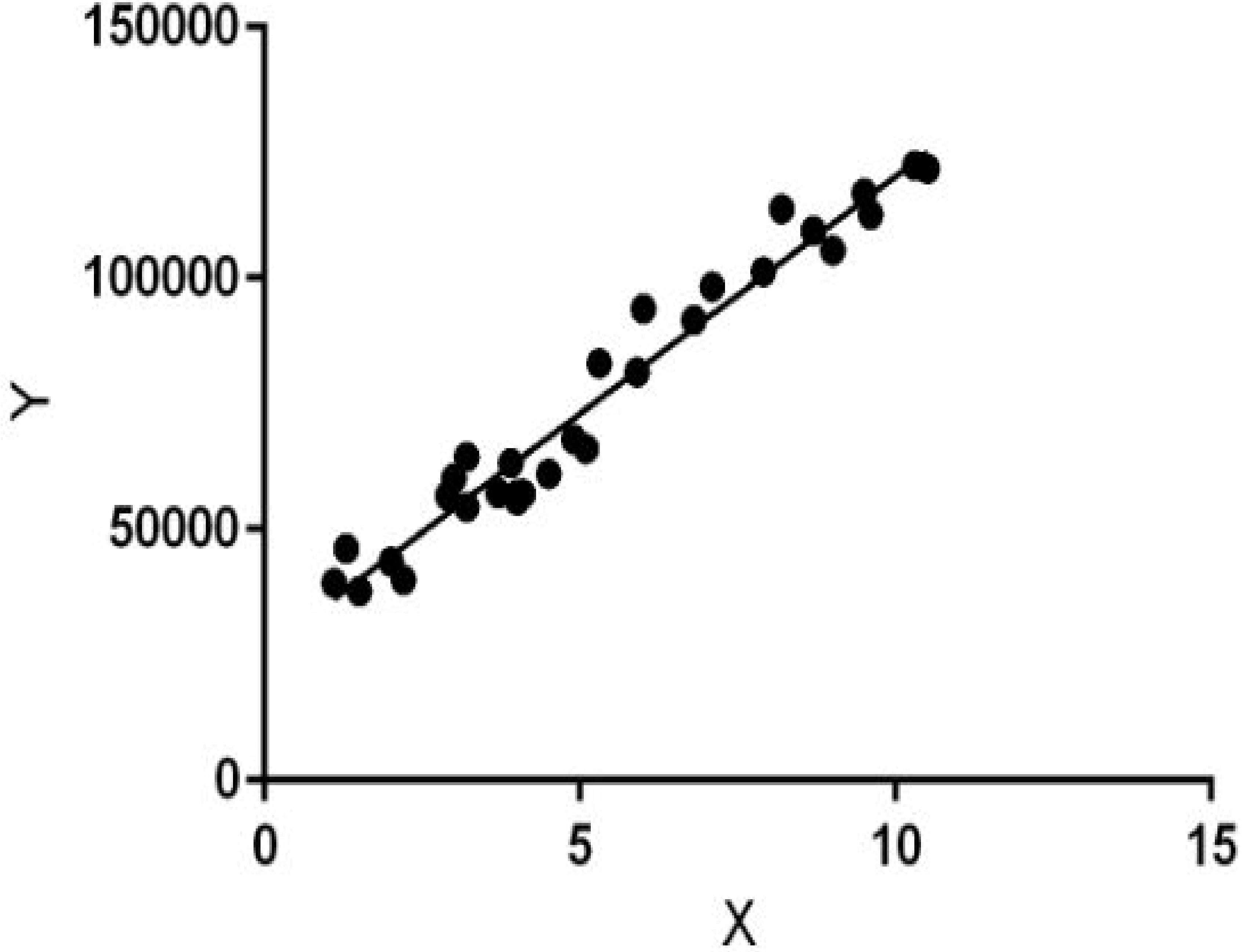
- These permit models that fit training data much more tightly than linear functions can.

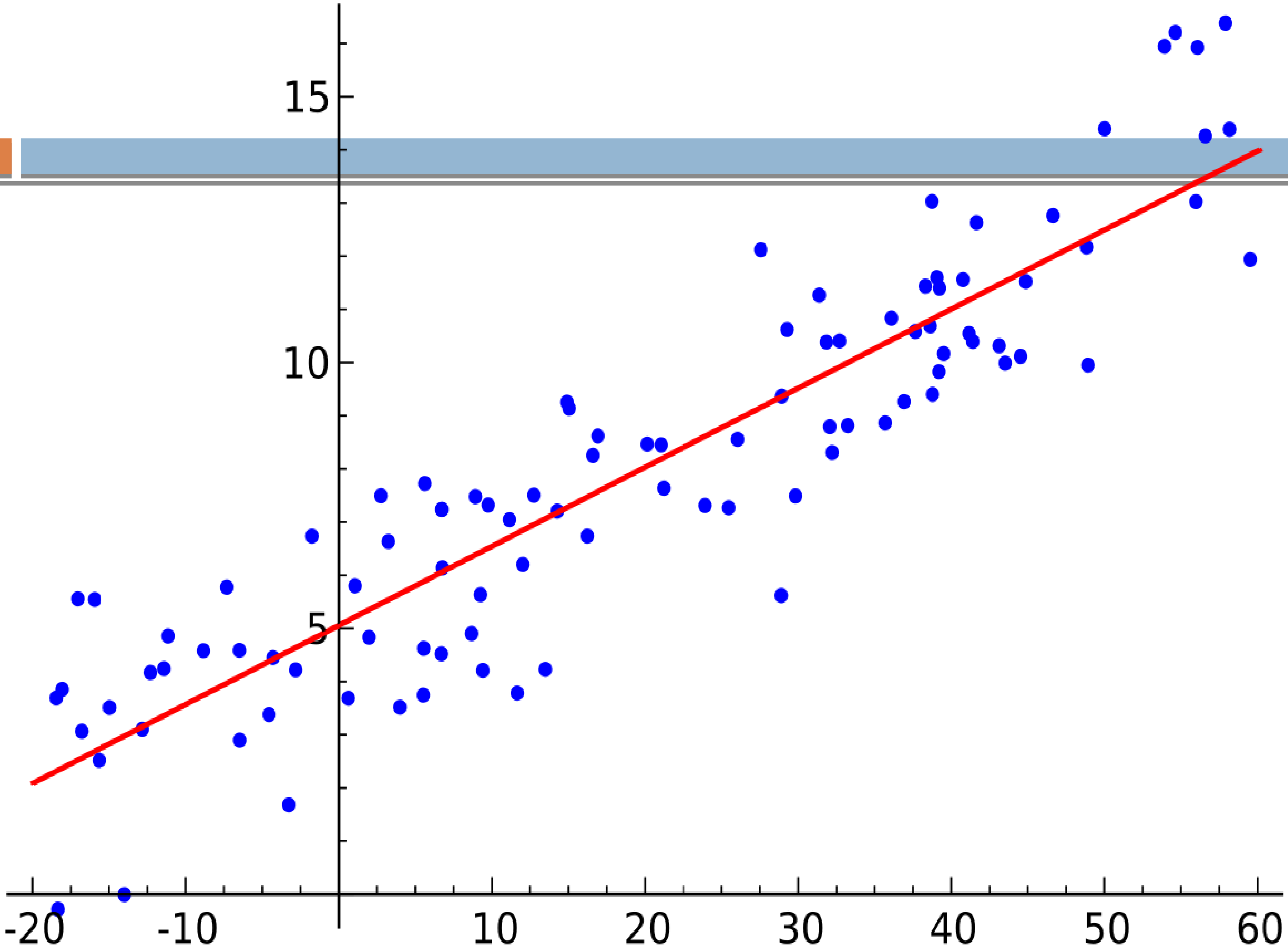
- <u>It is much harder to find the best possible coefficients to fit non-linear models</u>

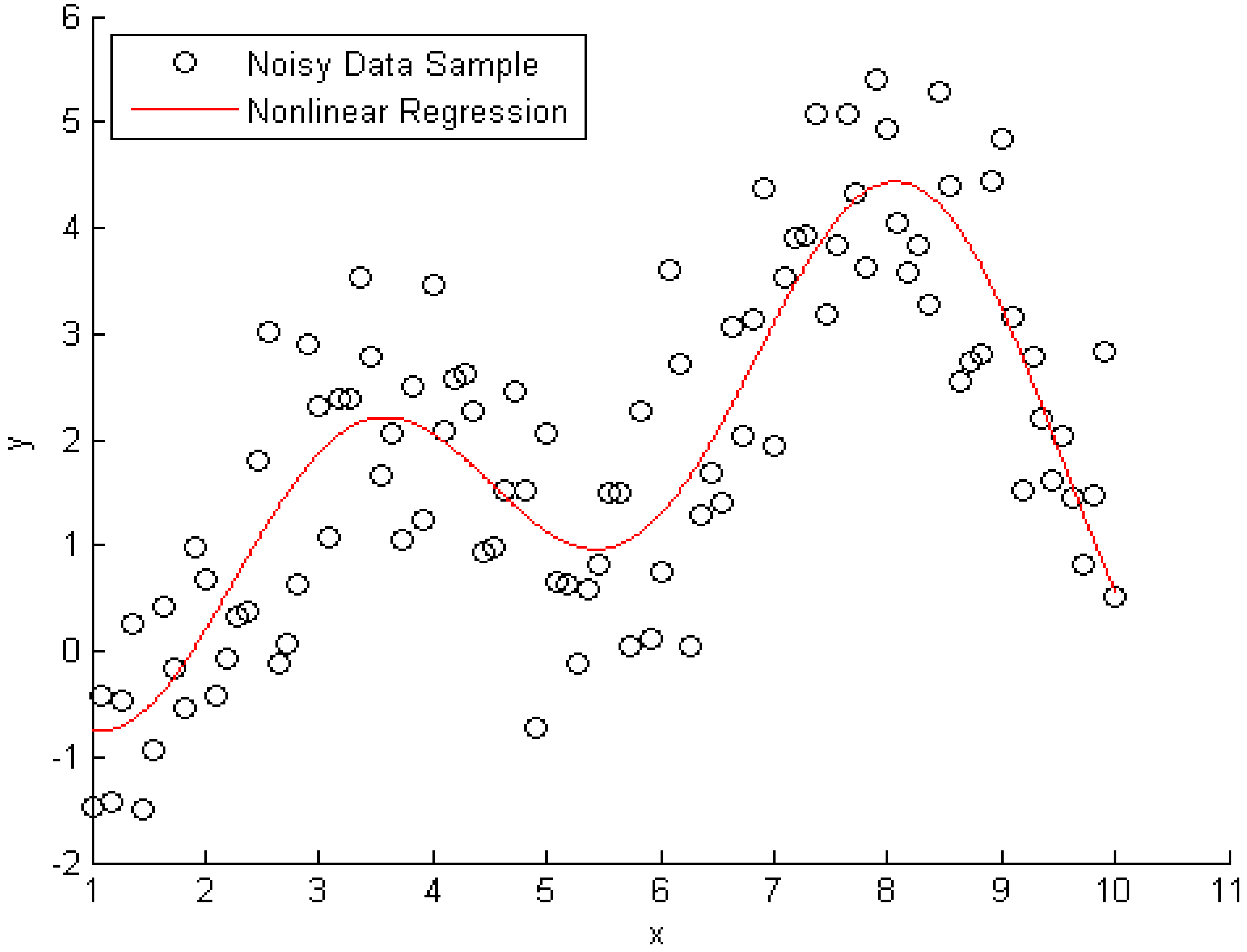
Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- □ But linear models offer substantial benefits.
  - They are readily understandable,
  - generally defensible, easy to build, and
  - avoid overfitting on modest-sized data sets.
- □ Occam's razor tells us that the simplest explanation is the best explanation.
- □ A robust linear model, yielding an accuracy of x%, better than a complex non-linear beast only a few percentage points better on limited testing data.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan


150000
100000
50000
0
Y
0
5
10
15
X


15
10
5
-20
-10
10
20
30
40
50
60



○ Noisy Data Sample
Nonlinear Regression
y
x
-2
-1
0
1
2
3
4
5
6
1
2
3
4
5
6
7
8
9
10
11

Blackbox vs. Descriptive Models

- Black boxes are devices that do their job, but in some unknown manner
- Descriptive models provide some insight into why they are making their decisions
- Theory-driven models are generally descriptive

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

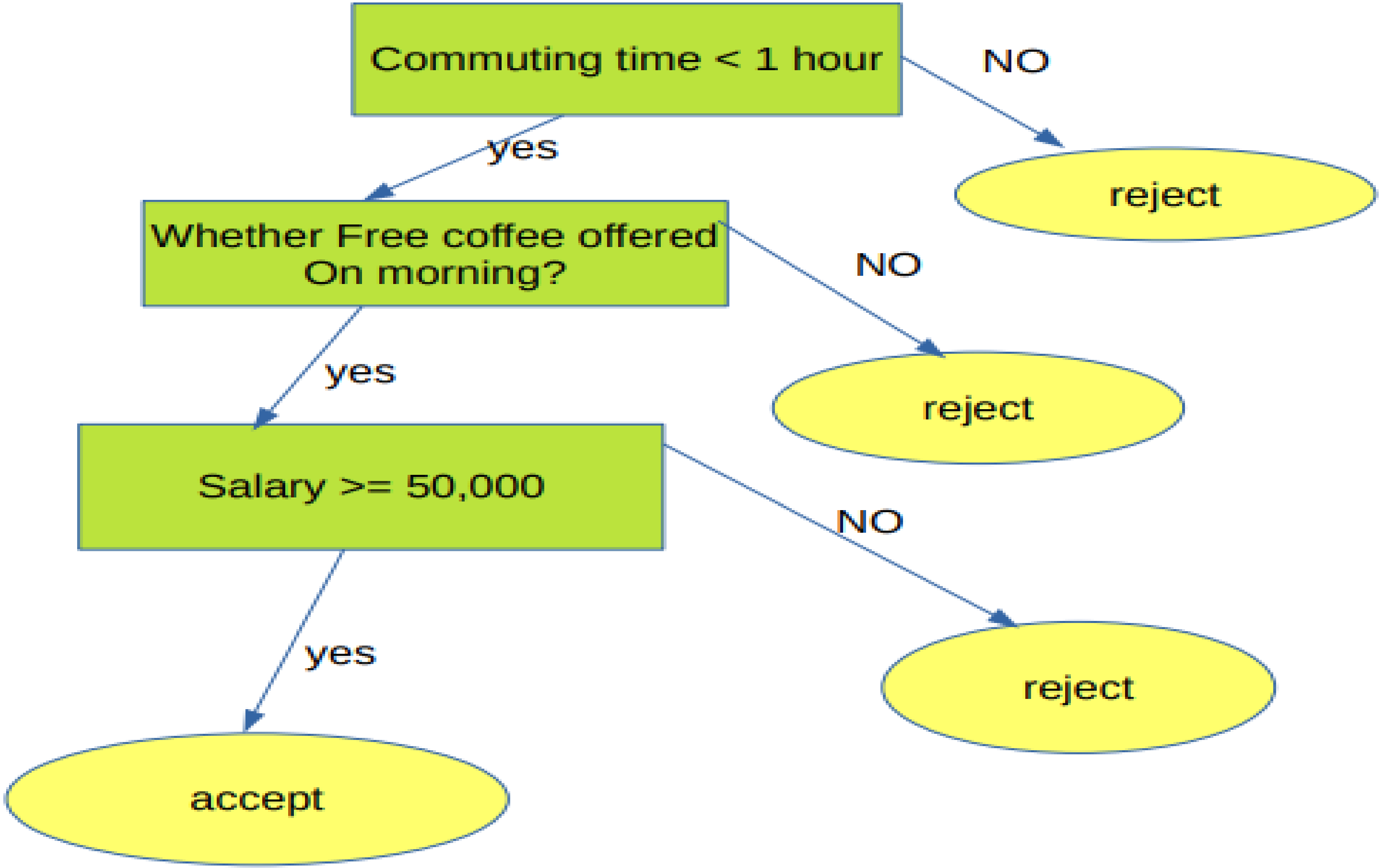
Descriptive Models

- Linear regression models are descriptive, because
  - one can see exactly which variables receive the most weight, and
  - measure how much they contribute to the resulting prediction.

- Decision tree models enable you to follow the exact decision path used to make a classification

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Decision tree for choosing JOB



Commuting time < 1 hour

NO

reject

yes

Whether Free coffee offered On morning?

NO

reject

yes

Salary >= 50,000

NO

reject

yes

accept

Blackbox Models

- Blackbox modeling techniques such as deep learning can be extremely effective.

- Neural network models are generally completely opaque as to why they do what they do.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- A system built for the military to distinguish images of cars from trucks.

- It performed well in training, but disastrously in the field.

- Only later was it realized that the training images for cars were shot on a sunny day and those of trucks on a cloudy day, so the system had learned to link the sky in the background with the class of the vehicle

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

First-Principle vs. Data-Driven Models

- First-principle models are based on a belief of how the system under investigation really works.

- It might be a theoretical explanation, like Newton's laws of motion.

- Such models can employ the full weight of classical mathematics: calculus, algebra, geometry, and more.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- voters are unhappy if the economy is bad, therefore variables which measure the state of the economy should help us predict who will win the election.

- Data-driven models are based on observed correlations between input parameters and outcome variables

- The same basic model might be used to predict tomorrow's weather or the price of a given stock, differing only on the data it was trained on.

- Machine learning methods make it possible to build an effective model on a domain one knows nothing about, provided we are given a good enough training set.

Stochastic vs. Deterministic Models

- Stochastic is a fancy word meaning randomly determined.

- Techniques that explicitly build some notion of probability into the model include logistic regression and Monte Carlo simulation.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- □ It is important that your model observe the basic properties of probabilities, including:

- □ Each probability is a value between 0 and 1

- □ That they must sum to 1

- □ Rare events do not have probability zero:
  - Any event that is possible must have a greater than zero probability of occurrence

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Deterministic models always return the same answer helps greatly in debugging their implementation.

- This speaks to the need to optimize repeatability during model development.

- Fix the initial seed if you are using a random number generator, so you can rerun it and get the same answer.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Flat vs. Hierarchical Models

- Interesting problems often exist on several different levels, each of which may require independent submodels

- Imposing a hierarchical structure on a model permits it to be built and evaluated in a logical and transparent way, instead of as a black box.

- Hierarchical models are descriptive: one can trace a final decision back to the appropriate top-level subproblem, and report how strongly it contributed to making the observed result

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Predicting the future price for a particular stock really should involve submodels for analyzing such separate issues as
  - (a)the general state of the economy,
  - (b) the company’s balance sheet, and
  - (c) the performance of other companies in its industrial sector.

- The first step to build a hierarchical model is explicitly decomposing our problem into subproblems.

- Deep learning models can be thought of as being both at flat and hierarchical, at the same time

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Baseline Models

- The first step to assess the complexity of your task involves building baseline models: the simplest reasonable models that produce answers we can compare against.

- More sophisticated models should do better than baseline models, but verifying that they really do and, if so by how much, puts its performance into the proper context.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Evaluating Models

- But the best way to assess models involves out-of-sample predictions, <u>results on data that you never saw when you built the model</u>.

- Good performance on the data that you trained models on is very suspect, because models can easily be overfit.

Evaluating Classifiers

- Two distinct labels or classes (binary classification)

- The smaller and more interesting of the two classes as positive and the larger/other class as negative.

- In a spam classification problem, the spam would typically be positive and the ham (non-spam) would be negative

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- There are four possible results of what the classification model could do on any given instance, which defines the confusion matrix or contingency table

- True Positives (TP):
  - Here our classier labels a positive item as positive, resulting in a win for the classier.

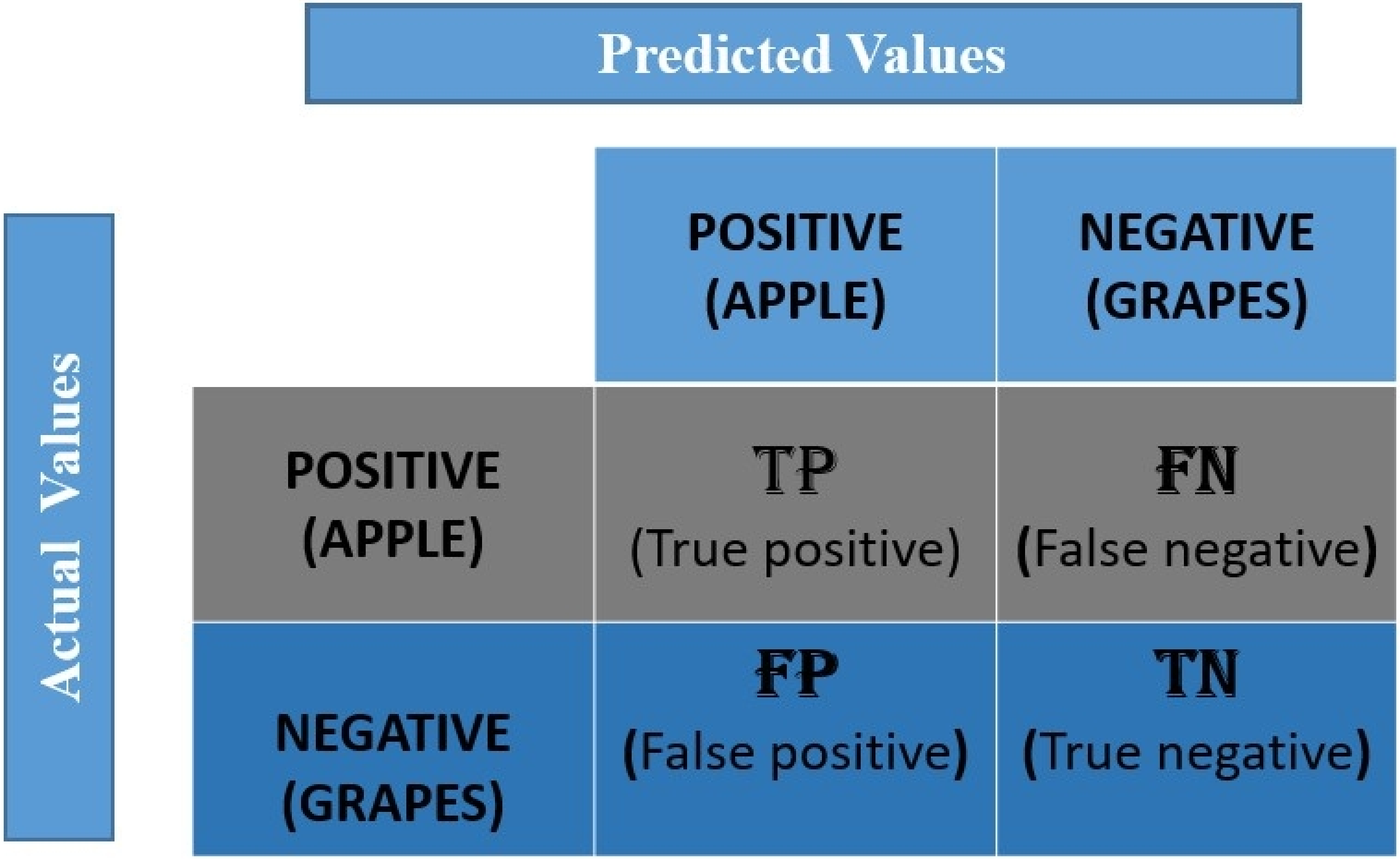
- True Negatives (TN):
  - Here the classier correctly determines that a member of the negative class deserves a negative label.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- False Positives (FP):
  - The classier mistakenly calls a negative item as a positive, resulting in a type I classification error.

- False Negatives (FN):
  - The classier mistakenly declares a positive item as negative, resulting in a type II" classification error.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan



		Predicted Values	
		POSITIVE (APPLE)	NEGATIVE (GRAPES)
Actual Values	POSITIVE (APPLE)	TP (True positive)	FN (False negative)
	NEGATIVE (GRAPES)	FP (False positive)	TN (True negative)

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Baseline Evaluators

- We must defend our classifier against two baseline opponents, the sharp and the monkey

- The sharp is the opponent who knows what evaluation system we are using, and picks the baseline model which will do best according to it.

- The sharp will try to make the evaluation statistic look bad, by achieving a high score with a useless classifier.

- That might mean declaring all items positive, or perhaps all negative.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- The monkey randomly guesses on each instance.

- To interpret our model's performance, it is important to establish by how much it beats both the sharp and the monkey.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Accuracy of the classifier

- The ratio of the number of correct predictions over total predictions

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- By multiplying such fractions by 100, we can get a percentage accuracy score.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Accuracy alone has limitations as an evaluation metric, particularly when the positive class is much smaller than the negative class

- Consider the development of a classier to diagnose whether a patient has cancer, where the positive class has the disease (i.e. tests positive) and the negative class is healthy.

- The prior distribution is that the vast majority of people are healthy ((positive)/(positive + negative))<<1/2

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- The expected accuracy of a fair-coin monkey would still be 0.5:
  - it should get an average of half of the positives and half the negatives right.
- But the sharp would declare everyone to be healthy, achieving an accuracy of 1- p.
- Suppose that only 5% of the test takers really had the disease.
- The sharp could brag about her accuracy of 95%

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Precision

- We need evaluation metrics that are more sensitive to getting the positive class right.

- Precision measures how often this classifier is correct when it dares to say positive.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- Achieving high precision is impossible for either a sharp or a monkey, because the fraction of positives ($p = 0:05$) is so low.

- If the classier issues too many positive labels, it is doomed to low precision

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Recall

- In the cancer diagnosis case, we might be more ready to tolerate false positives (errors where we scare a healthy person with a wrong diagnosis) than false negatives (errors where we kill a sick patient by misdiagnosing their illness).

- Recall measures how often you prove right on all positive instances:

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- A high recall implies that the classier has few false negatives

- The easiest way to achieve this declares that everyone has cancer, as done by a sharp always answering yes.

- This classier has high recall but low precision:

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

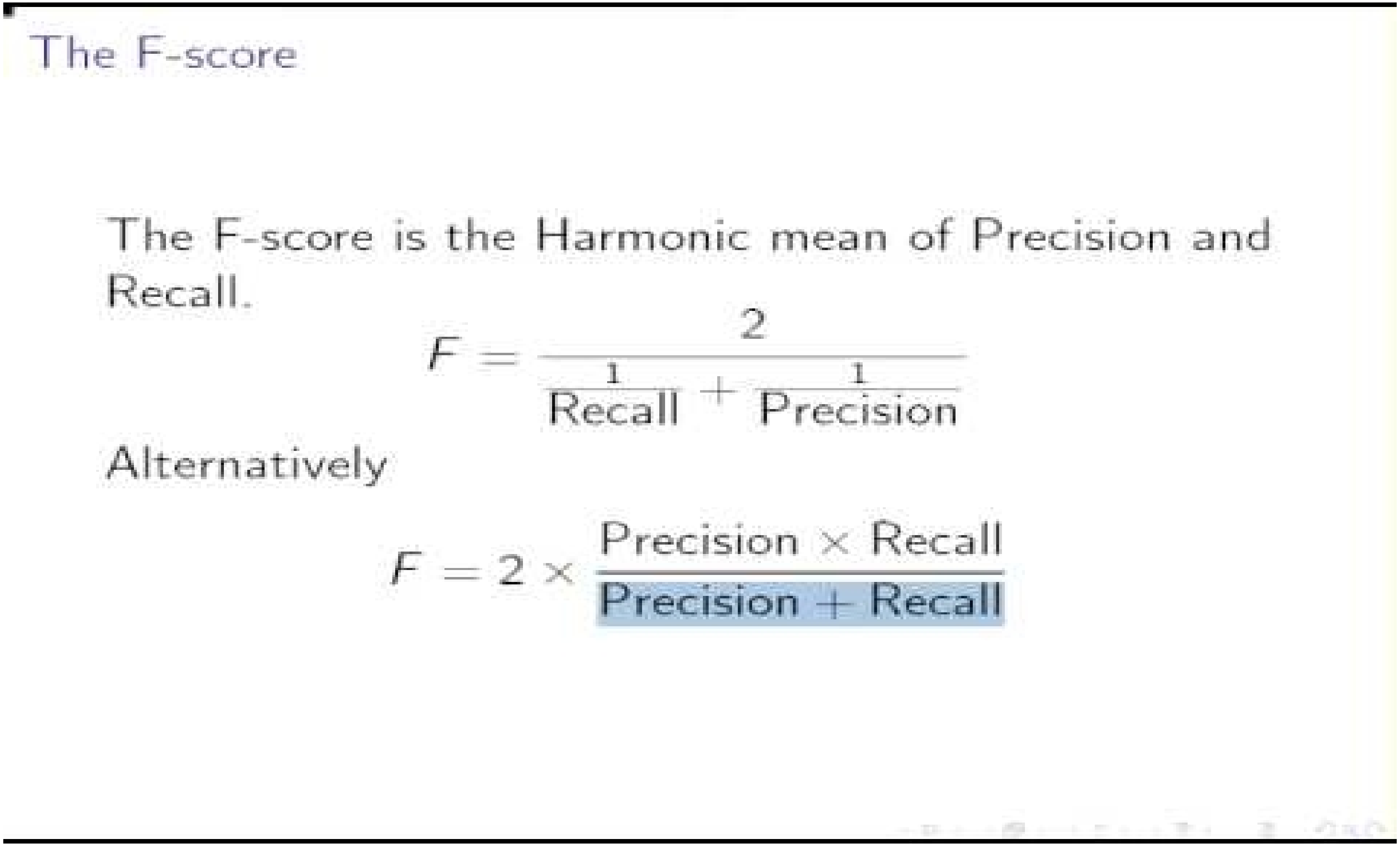
F-score (F1-score)

- Harmonic mean of precision and recall

- The harmonic mean is always less than or equal to the arithmetic mean

- Achieving a high F-score requires both high recall and high precision

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan



The F-score

The F-score is the Harmonic mean of Precision and Recall.

$$F = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$$

Alternatively

$$F = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- The higher the F-score the better the predictive power of classification procedure.
- A score 1 means classification procedure is perfect
- Lowest possible F-score is 0
- $0 \leq F \leq 1$

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Accuracy is a misleading statistic when the class sizes are substantially different

- Recall equals accuracy if and only if the classifiers are balanced

- High precision is very hard to achieve in unbalanced class sizes:

- F-score does the best job of any single statistic, but all four work together to describe the performance of a classier

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Problem 1

- Suppose a computer program for recognizing dogs in photographs identifies eight dogs in a picture containing 12 dogs and some cats.
- Of the eight dogs identified, five actually are dogs while the rest are cats.
- Compute the precision and recall of the computer program.

Problem 1

- $TP = 5$
- $FP = 3$
- $FN = 7$

- The precision P is $P = TP/(TP + FP)$

  $= 5/(5 + 3) \quad = 5/ 8$

- The recall R is R $= TP/(TP + FN)$

  $= 5/(5 + 7) \quad = 5/ 12$

Problem 2

- Let there be 10 balls (6 white and 4 red balls) in a box and let it be required to pick up the red balls from them.

- Suppose we pick up 7 balls as the red balls of which only 2 are actually red balls.

- What are the values of precision and recall in picking red ball?

Problem 2

- TP = 2
- FP = 7 – 2 = 5
- FN = 4 – 2 = 2

- The precision P is P = TP/(TP + FP)

  = 2/(2 + 5) = 2/ 7

The recall R is R = TP/(TP + FN)

= 2/(2 + 2) = 1/2

Problem 3

- A database contains 80 records on a particular topic of which 55 are relevant to a certain investigation. A search was conducted on that topic and 50 records were retrieved.
- Of the 50 records retrieved, 40 were relevant. Construct the confusion matrix for the search and calculate the precision and recall scores for the search.
- Each record may be assigned a class label “relevant" or “not relevant”.
- All the 80 records were tested for relevance. The test classified 50 records as “relevant”.
- But only 40 of them were actually relevant.

Problem 3

	Actual ‘Relevant’	Actual ‘Not Relevant’
Predicted ‘Relevant’	40	10
Predicted ‘Not Relevant’	15	15

Problem 3

- TP = 40
- FP = 10
- FN = 15

- The precision P is P = TP/(TP + FP)

  = 40/(40 + 10) = 4/ 5

The recall R is R = TP/(TP + FN)

  = 40/(40 + 15) = 40/ 55

Other measures of performance

- Using the data in the confusion matrix of a classifier of two-class dataset, several measures of performance have been defined.
- Accuracy = (TP + TN)/(TP + TN + FP + FN)
- Error rate = 1− Accuracy
- Sensitivity = TP/(TP + FN)
- Specificity = TN /(TN + FP)
- F-measure = (2 × TP)/(2 × TP + FP + FN)

Receiver Operating Characteristic (ROC)

- The acronym ROC stands for Receiver Operating Characteristic, a terminology coming from signal detection theory.

- The ROC curve was first developed by electrical engineers and radar engineers during World War II for detecting enemy objects in battlefields.

- They are now increasingly used in machine learning and data mining research.

TPR and FPR

- Let a binary classifier classify a collection of test data.

- TP = Number of true positives
- TN = Number of true negatives
- FP = Number of false positives
- FN = Number of false negatives

- TPR = True Positive Rate = TP/(TP + FN)= Fraction of positive examples correctly classified = Sensitivity

- FPR = False Positive Rate = FP /(FP + TN) = Fraction of negative examples incorrectly classified = 1 − Specificity

ROC space

- We plot the values of FPR along the horizontal axis (that is , x-axis) and the values of TPR along the vertical axis (that is, y-axis) in a plane.

- For each classifier, there is a unique point in this plane with coordinates (FPR,TPR).

- The ROC space is the part of the plane whose points correspond to (FPR,TPR).

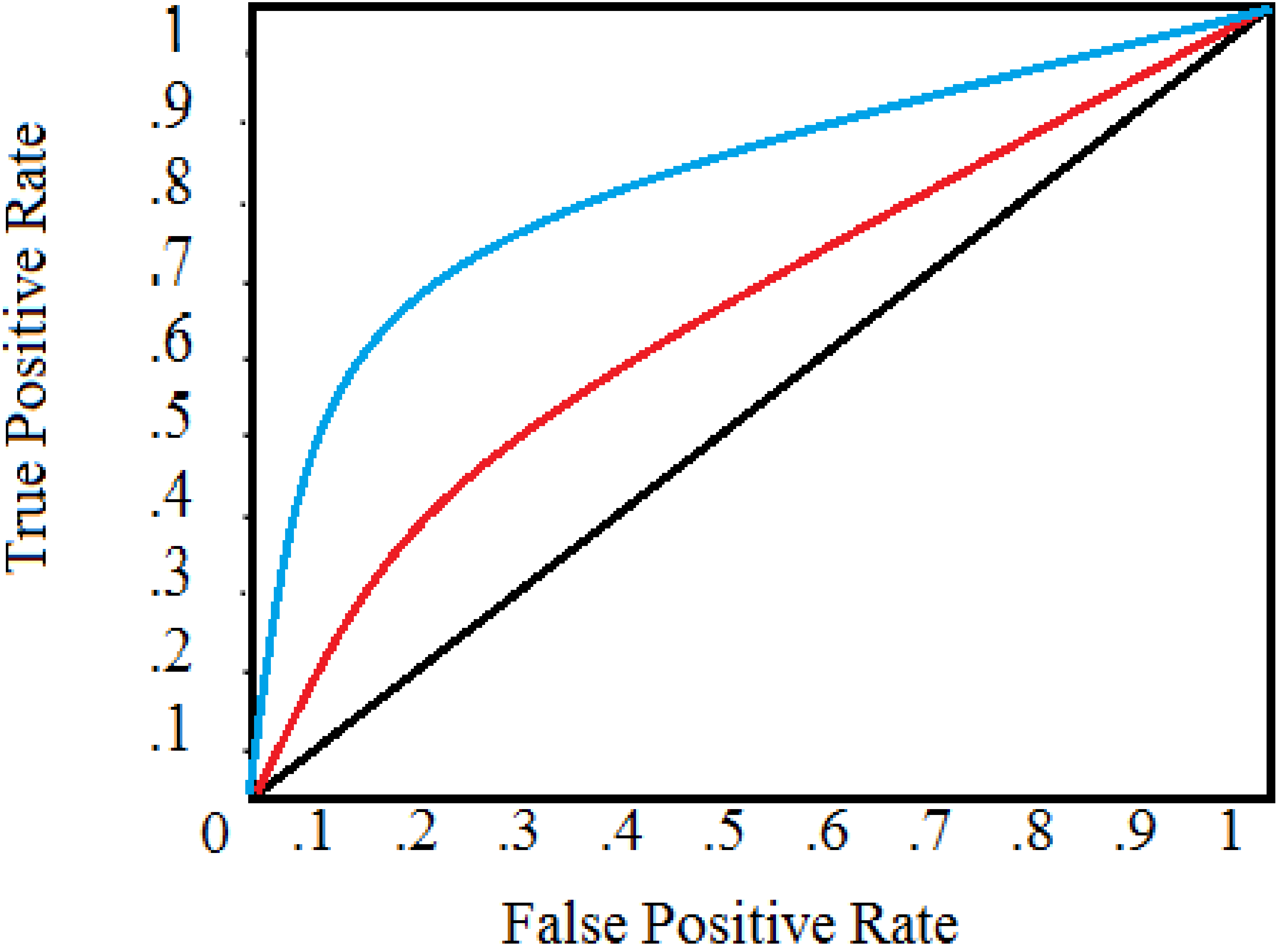
- Each prediction result or instance of a confusion matrix represents one point in the ROC space.

ROC space

- The position of the point (FPR,TPR) in the ROC space gives an indication of the performance of the classifier.

- For example, let us consider some special points in the space

- One step higher for positive examples and one step right for negative examples


True Positive Rate
1
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
False Positive Rate

Special points in ROC space

- The left bottom corner point (0, 0):
  - Always negative prediction
  - A classifier which produces this point in the ROC space never classifies an example as positive, neither rightly nor wrongly, because for this point TP = 0 and FP = 0.
  - It always makes negative predictions.
  - All positive instances are wrongly predicted and all negative instances are correctly predicted.
  - It commits no false positive errors.

Special points in ROC space

- The right top corner point (1, 1):
  - Always positive prediction
  - A classifier which produces this point in the ROC space always classifies an example as positive because for this point FN = 0 and TN = 0.
  - All positive instances are correctly predicted and all negative instances are wrongly predicted.
  - It commits no false negative errors.

Special points in ROC space

- □ The left top corner point (0, 1):
  - Perfect prediction
  - A classifier which produces this point in the ROC space may be thought as a perfect classifier.
  - It produces no false positives and no false negatives

Special points in ROC space

- Points along the diagonal:
  - Random performance
  - Consider a classifier where the class labels are randomly guessed, say by flipping a coin.
  - Then, the corresponding points in the ROC space will be lying very near the diagonal line joining the points (0, 0) and (1, 1).

ROC curve

- In the case of certain classification algorithms, the classifier may depend on a parameter.

- Different values of the parameter will give different classifiers and these in turn give different values to TPR and FPR.

- The ROC curve is the curve obtained by plotting in the ROC space the points (TPR , FPR) obtained by assigning all possible values to the parameter in the classifier

ROC curve

- The closer the ROC curve is to the top left corner (0, 1) of the ROC space, the better the accuracy of the classifier.

- Among the three classifiers A, B, C with ROC curves , the classifier C is closest to the top left corner of the ROC space.

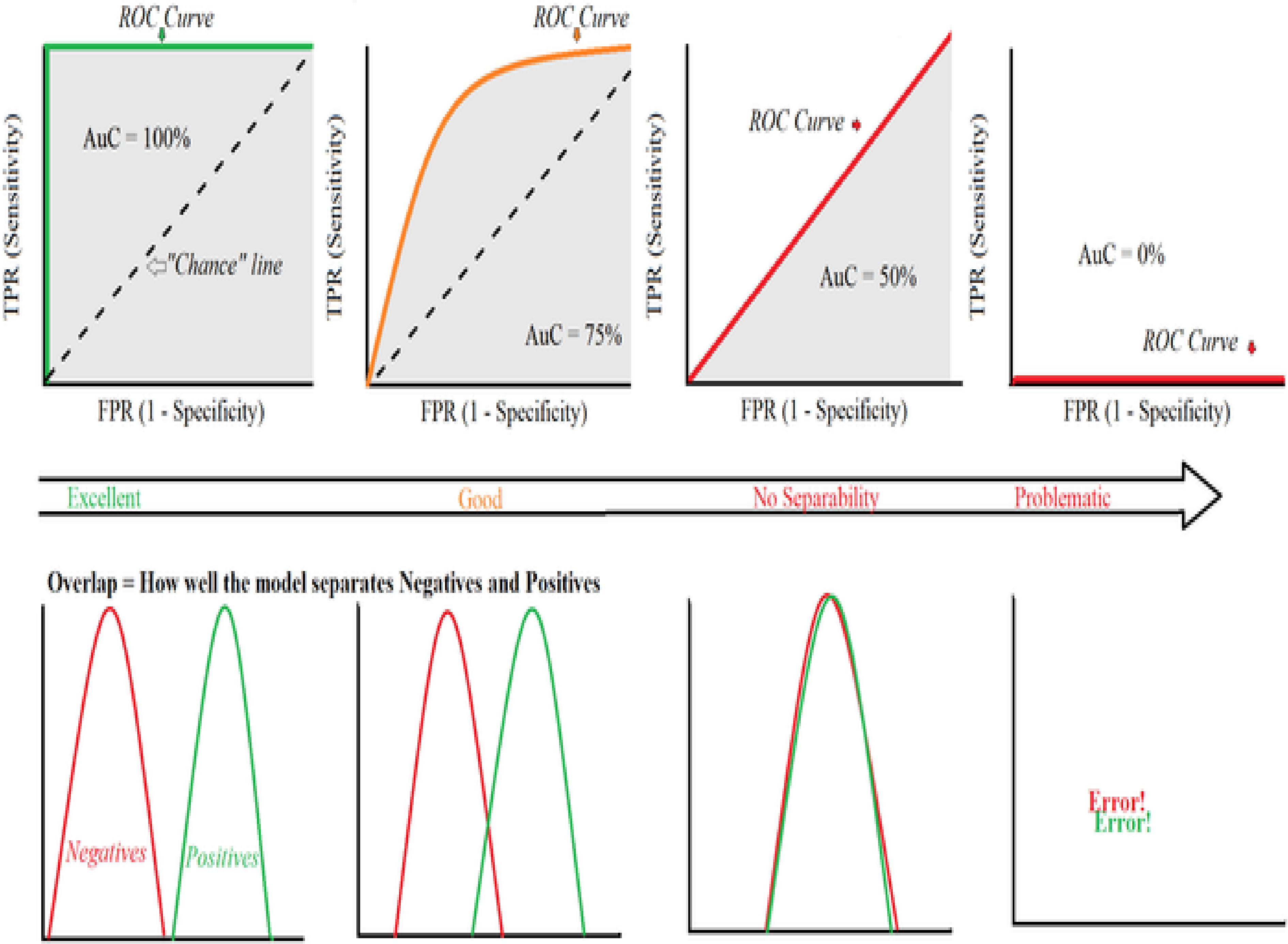
- Hence, among the three, it gives the best accuracy in predictions.

Area under the ROC curve (AUC)

- The measure of the area under the ROC curve is denoted by the acronym AUC .

- The value of AUC is a measure of the performance of a classifier.

- For the perfect classifier, AUC = 1.0


ROC Curve
AuC = 100%
"Chance" line
TPR (Sensitivity)
FPR (1 - Specificity)
ROC Curve
AuC = 75%
TPR (Sensitivity)
FPR (1 - Specificity)
ROC Curve
AuC = 50%
TPR (Sensitivity)
FPR (1 - Specificity)
AuC = 0%
ROC Curve
TPR (Sensitivity)
FPR (1 - Specificity)
Excellent
Good
No Separability
Problematic
Overlap = How well the model separates Negatives and Positives
Negatives
Positives
Error!
Error!

Evaluating Multiclass Systems

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

Evaluating Value Prediction Models

- For numerical values, error is a function of the difference between a forecast y’ = f(x) and the actual result y.

- Measuring the performance of a value prediction system involves two decisions:
  - (1) fixing the specific individual error function,
  - (2) selecting the statistic to best represent the full error distribution.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Absolute error:
  - The value Δ= y’ - y has the virtue of being simple and symmetric,
  - the sign can distinguish the case where y’ > y from y > y’

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- □ Relative error:
  - The absolute magnitude of error is meaningless without a sense of the units involved.
  - An absolute error of 1.2 in a person's predicted height is good if it is measured in millimeters, but terrible if measured in miters.

- □ Normalizing the error by the magnitude of the observation produces a unit-less quantity, which can be sensibly interpreted as a fraction or (multiplied by 100%) as a percentage:

- □ $\varepsilon = (y - y')/y$.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- Squared error:
  - The value $\Delta^2 = (y' - y)^2$ is always positive

- Large errors values contribute disproportionately to the total when squaring: Δ^2 for $\Delta = 2$ is four times larger than Δ^2 for $\Delta = 1$.

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

- A commonly-used statistic is mean squared error (MSE), which is computed it weighs each term quadratically, outliers have a disproportionate effect.

- Thus median squared error might be a more informative statistic for noisy instances

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = mean squared error

n = number of data points

Y_i = observed values

$\hat{Y}_i$ = predicted values

- Root mean squared (RMSD) error is simply the square root of mean squared error:
- The advantage of RMSD is that its magnitude is interpretable on the same scale as the original values

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

Reference: Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan