MODULE 1

FOUNDATIONS DATA SCIENCE, PROCESS, AND TOOLS

SYLLABUS

- 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

IMPORTANT QUESTIONS

- What is data science? What are the different models for data science?
- What is data visualization and what are the different techniques used for visualizing data?
- Explain data science process with a neat diagram.
- Describe data science classification with a neat diagram.
- What is data cleaning? What are the different operations in data cleaning?
- Explain different types of Data Sets in Data Science.

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. Being a buzzword, there is a wide variety of definitions and criteria for what constitutes data science. Data science is also commonly referred to as knowledge discovery, machine learning, predictive analytics, and data mining.

AI, MACHINE LEARNING, AND DATA SCIENCE

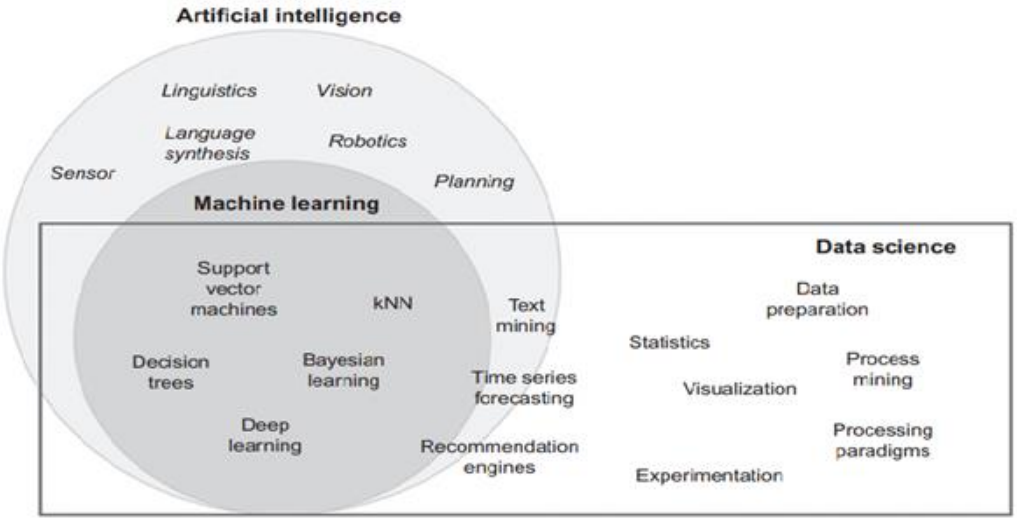


Figure 1: Relationship between AI, ML and data science

Artificial intelligence

- Artificial intelligence is about giving machines the capability of mimicking human behavior, particularly cognitive functions.
- Examples would be: facial recognition, automated driving, sorting mail based on postal code.
- There are quite a range of techniques that fall under artificial intelligence: linguistics, natural language processing, decision science, bias, vision, robotics, planning, etc. Learning is an important part of human capability.

Machine learning

- Machine learning can either be considered a sub-field or one of the tools of artificial intelligence, is providing machines with the capability of learning from experience.
- Experience for machines comes in the form of data.
- Data that is used to teach machines is called training data.
- Machine learning turns the traditional programing model upside down.
- Machine learning turns the traditional programing model upside down.

- A program, a set of instructions to a computer, transforms input signals into output signals using predetermined rules and relationships in traditional program.
- Figure 2 shows Machine learning algorithms, also called “learners”, take both the known input and output (training data) to figure out a model for the program which converts input to output.

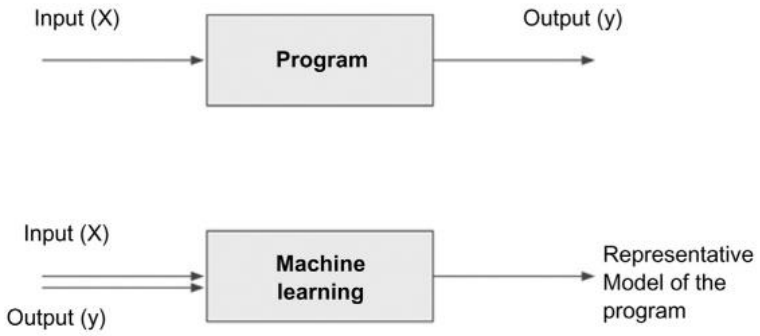


Figure 2: Traditional program and machine learning.

Data Science

Data science is the business application of machine learning, artificial intelligence, and other quantitative fields like statistics, visualization, and mathematics. It is an interdisciplinary field that extracts value from data. In the context of how data science is used today, it relies heavily on machine learning and is sometimes called data mining.

Examples of data science user cases are: recommendation engines that can recommend movies for a particular user, a fraud alert model that detects fraudulent credit card transactions, find customers who will most likely churn next month, or predict revenue for the next quarter. Data science starts with data, which can range from a simple array of a few numeric observations to a complex matrix of millions of observations with thousands of variables. Data science utilizes certain specialized computational methods in order to discover meaningful and useful structures within a dataset. The discipline of data science coexists and is closely associated with a number of related areas such as database systems, data engineering, visualization, data analysis, experimentation, and business intelligence (BI).

Data science starts with data, which can range from a simple array of a few numeric observations to a complex matrix of millions of observations with thousands of variables. Data science utilizes certain specialized computational methods in order to discover meaningful and useful structures within a dataset. The discipline of data science coexists and is closely associated with a number of related areas such as database systems, data engineering, visualization, data analysis, experimentation, and business intelligence (BI). We can further define data science by investigating some of its key features and motivations.

Extracting Meaningful Patterns

Data science involves inference and iteration of many different hypotheses. One of the key aspects of data science is the process of generalization of patterns from a dataset. The generalization should be valid, not just for the dataset used to observe the pattern, but also for new unseen data. Data science is also a process with defined steps, each with a set of tasks. The term novel indicates that data science is usually involved in finding previously unknown patterns in data. The ultimate objective of data science is to find potentially useful conclusions that can be acted upon by the users of the analysis.

Building Representative Models

In statistics, a model is the representation of a relationship between variables in a dataset. It describes how one or more variables in the data are related to other variables. Modeling is a process in which a representative abstraction is built from the observed dataset.
For example, based on credit score, income level, and requested loan amount, a model can be developed to determine the interest rate of a loan. For this task, previously known observational data including credit score, income level, loan amount, and interest rate are needed. Figure 3 shows how a model is built.

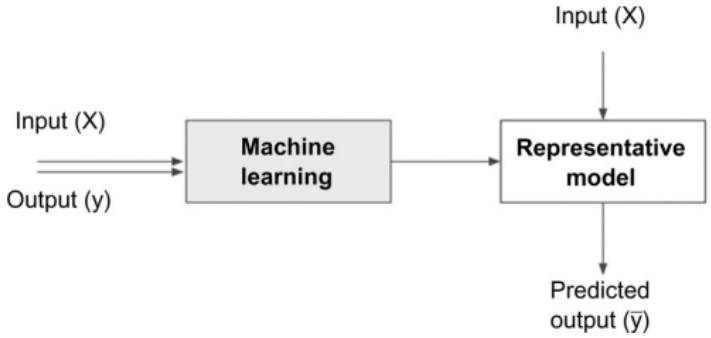


Figure 3: Building models

The various techniques used in the Steps of a Data Science Process

Descriptive statistics

- Computing mean, standard deviation, correlation, and other descriptive statistics, quantify the aggregate structure of a dataset.
- This is essential information for understanding any dataset in order to understand the structure of the data and the relationships within the dataset.
- They are used in the exploration stage of the data science process.

Exploratory visualization

- The process of expressing data in visual coordinates enables users to find patterns and relationships in the data and to comprehend large datasets.
- Similar to descriptive statistics, they are integral in the pre- and post-processing steps in data science.

Dimensional slicing:

- Online analytical processing (OLAP) applications, which are prevalent in organizations, mainly provide information on the data through dimensional slicing, filtering, and pivoting.
- OLAP analysis is enabled by a unique database schema design where the data are organized as dimensions (e.g., products, regions, dates) and quantitative facts or measures (e.g., revenue, quantity).
- With a well-defined database structure, it is easy to slice the yearly revenue by products or combination of region and products. These techniques are extremely useful and may unveil patterns in data.

Hypothesis testing:

- In confirmatory data analysis, experimental data are collected to evaluate whether a hypothesis has enough evidence to be supported or not.
- There are many types of statistical testing and they have a wide variety of business applications (e.g., A/B testing in marketing).
- In general, data science is a process where many hypotheses are generated and tested based on observational data. Since the data science algorithms are iterative, solutions can be refined in each step.

Data engineering:

- Data engineering is the process of sourcing, organizing, assembling, storing, and distributing data for effective analysis and usage. Database engineering, distributed storage, and computing frameworks (e.g., Apache Hadoop, Spark, Kafka), parallel computing, extraction transformation and loading processing, and data warehousing constitute data engineering techniques.
- Data engineering helps source and prepare for data science learning algorithms.

Business intelligence:

- Business intelligence helps organizations consume data effectively. It helps query the ad hoc data without the need to write the technical query command or use dashboards or visualizations to communicate the facts and trends.
- Business intelligence specializes in the secure delivery of information to right roles and the distribution of information at scale.
- Historical trends are usually reported, but in combination with data science, both the past and the predicted future data can be combined. BI can hold and distribute the results of data science.

DATA SCIENCE CLASSIFICATION

Data science problems can be broadly categorized into:

- supervised learning models.
- unsupervised learning models.

Supervised learning model

- 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. To do this, a model is developed from a training dataset where the values of input and output are previously known.
- 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.

Unsupervised learning model

- Unsupervised or undirected data science uncovers hidden patterns in unlabeled data.
- In unsupervised data science, there are no output variables to predict.
- The objective of this class of data science techniques, is to find patterns in data based on the relationship between data points themselves.

DATA SCIENCE TASKS

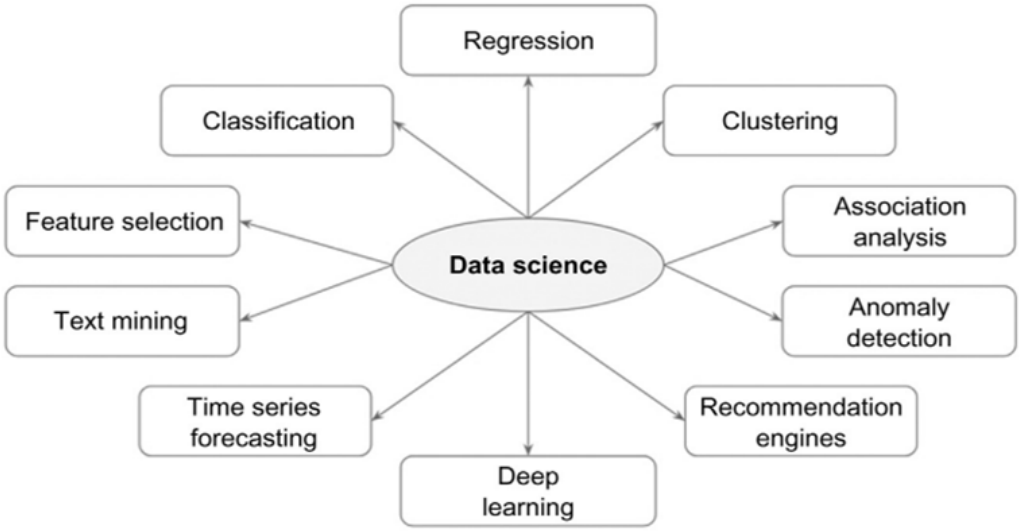


Figure 4: Data Science tasks

- 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 that is increasingly used for classification and regression problems.

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

- Clustering is the process of identifying the natural groupings in a dataset. For example, clustering is helpful in finding natural clusters in customer datasets, which can be used for market segmentation. Since this is unsupervised data science, it is up to the end user to investigate why these clusters are formed in the data and generalize the uniqueness of each cluster.

- Association analysis: In retail analytics, it is common to identify pairs of items that are purchased together, so that specific items can be bundled or placed next to each other. This task is called market basket analysis or association analysis, which is commonly used in cross selling.

- Anomaly or outlier detection identifies the data points that are significantly different from other data points in a dataset. Credit card transaction fraud detection is one of the most prolific applications of anomaly 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.

- 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. To aid the data science on text data, the text files are first converted into document vectors where each unique word is an attribute. Once the text file is converted to document vectors, standard data science tasks such as classification, clustering, etc., can be applied.

- Feature selection is a process in which attributes in a dataset are reduced to a few attributes that really matter.

Table 1 Data Science Tasks and Examples

Tasks	Description	Algorithms	Examples
Classification	Predict if a data point belongs to one of the predefined classes. The prediction will be based on learning from a known dataset	Decision trees, neural networks, Bayesian models, induction rules, *k*-nearest neighbors	Assigning voters into known buckets by political parties, e.g., soccer moms Bucketing new customers into one of the known customer groups
Regression	Predict the numeric target label of a data point. The prediction will be based on learning from a known dataset	Linear regression, logistic regression	Predicting the unemployment rate for the next year Estimating insurance premium
Anomaly detection	Predict if a data point is an outlier compared to other data points in the dataset	Distance-based, density-based, LOF	Detecting fraudulent credit card transactions and network intrusion
Time series forecasting	Predict the value of the target variable for a future timeframe based on historical values	Exponential smoothing, ARIMA, regression	Sales forecasting, production forecasting, virtually any growth phenomenon that needs to be extrapolated
Clustering	Identify natural clusters within the dataset based on inherit properties within the dataset	*k*-Means, density-based clustering (e.g., DBSCAN)	Finding customer segments in a company based on transaction, web, and customer call data
Association analysis	Identify relationships within an item set based on transaction data	FP-growth algorithm, a priori algorithm	Finding cross-selling opportunities for a retailer based on transaction purchase history
Recommendation engines	Predict the preference of an item for a user	Collaborative filtering, content-based filtering, hybrid recommenders	Finding the top recommended movies for a user

LOF, *local outlier factor;* ARIMA, *autoregressive integrated moving average;* DBSCAN, *density-based spatial clustering of applications with noise;* FP, *frequent pattern.*

COMPUTER SCIENCE, DATA SCIENCE, AND REAL SCIENCE

Real scientists strive to understand the natural world, which is a complicated and messy place. Computer scientists tend to build their own clean and organized virtual worlds and live comfortably within them. Scientists obsess about discovering things, while computer scientists invent rather than discover.

Examples of the cultural differences between computer science and real science include:

- Data vs. method centrism
- Concern about results
- Robustness
- Precision

Asking Interesting Questions from Data

- Data scientists always ask questions.
- Good data scientists have wide-ranging interests.

- Software developers are not really encouraged to ask questions, but data scientists are.

DATASET

A dataset is an ordered collection of data. A collection of information obtained through observations, measurements, study, or analysis is referred to as data. It could include information such as facts, numbers, figures, names, or even basic descriptions of objects. For our study, data can be organized in the form of graphs, charts, or tables. Through data mining, data scientists assist in the analysis of gathered data. The different forms of data set are listed below.

The Baseball Encyclopedia

- Baseball has long had an outsized importance in the world of data science. This sport has been called the national pastime of the United States.
- What makes baseball important to data science is its extensive statistical record of play, dating back for well over a hundred years. Baseball is a sport of discrete events: pitchers throw balls and batters try to hit them – that naturally lends itself to informative statistics. Fans get immersed in these statistics as children, building their intuition about the strengths and limitations of quantitative analysis.
- This historical baseball record is available at http://www.baseball-reference.com. There you will find complete statistical data on the performance of every player who even stepped on the field. This includes summary statistics of each season's batting, pitching, and fielding record, plus information about teams and awards.

The Internet Movie Database (IMDB)

- The Internet Movie Database (IMDb) provides crowd sourced and curated data about all aspects of the motion picture industry, at www.imdb.com.
- IMDb currently contains data on over 3.3 million movies and TV programs. For each film, IMDb includes its title, running time, genres, date of release, and a full list of cast and crew.
- There is financial data about each production, including the budget for making the film and how well it did at the box office.
- There are extensive ratings for each film from viewers and critics. This rating data consists of scores on a zero to ten stars scale, cross-tabulated into averages by age and gender.
- Perhaps the most natural questions to ask IMDb involve identifying the extremes of movies and actors like :
    - Which actors appeared in the most films? Earned the most money? Appeared in the lowest rated films? Had the longest career or the shortest lifespan?
    - What was the highest rated film each year, or the best in each genre? Which movies lost the most money, had the highest-powered casts, or got the least favorable reviews.

- How well does movie gross correlate with viewer ratings or awards? Do customers instinctively flock to trash, or is virtue on the part of the creative team properly rewarded?

Google Ngrams

- Google undertook an effort to scan all of the world's published books. They haven't quite gotten there yet, but the 30 million books thus far digitized represent over 20% of all books ever published.
- Google uses this data to improve search results, and provide fresh access to out-of-print books.
- Google Ngrams, an amazing resource for monitoring changes in the cultural zeitgeist. It provides the frequency with which short phrases occur in books published each year. Each phrase must occur at least forty times in their scanned book corpus. This eliminates obscure words and phrases, but leaves over two billion time series available for analysis.
- This rich data set shows how language use has changed over the past 200 years, and has been widely applied to cultural trend analysis.
- Data processing was the popular term associated with the computing field during the punched card and spinning magnetic tape era of the 1950s.

New York Taxi Records

- Taxi cabs form an important part of the urban transportation network. They roam the streets of the city looking for customers, and then drive them to their destination for a fare proportional to the length of the trip. Each cab contains a metering device to calculate the cost of the trip as a function of time. This meter serves as a record keeping device, and a mechanism to ensure that the driver charges the proper amount for each trip.
- The taxi meters currently employed in New York cabs can do many things beyond calculating fares. They act as credit card terminals, providing a way for customers to pay for rides without cash. They are integrated with global positioning systems (GPS), recording the exact location of every pickup and drop off. And finally, since they are on a wireless network, these boxes can communicate all of this data back to a central server.
- The result is a database documenting every single trip by all taxi cabs in one of the world's greatest cities.

PROPERTIES OF DATA

The different type of data associated with data science are:

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

- Big Data vs. Little Data

Structured Vs. Unstructured Data

- Structured data sets are nicely arranged, 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, and the columns represent distinct properties of these items.
- An unstructured data source record information about the state of the world, but in a more heterogeneous way. Perhaps it is a large text corpus with images and links like Wikipedia, or the complicated mix of notes and test results appearing in personal medical records.
- An unstructured data source, such as a collection of tweets from Twitter, our first step is generally to 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.
- Un structured data that has no inherent structure, which may include text documents, PDFs, images, and video.

Quantitative Vs. Categorical Data

- Quantitative data consists of numerical values, like height and weight. Such data can be incorporated directly into algebraic formulas and mathematical models, or displayed in conventional graphs and charts.
- 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.
- But things get more complicated when there are more than two characters per feature, especially when there is not an implicit order between them.

Big Data vs. Little Data

- The analysis of massive data sets resulting from computer logs and sensor devices.
- 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.

The challenges of big data

- The analysis cycle time slows as data size grows:
- Large data sets are complex to visualize
- Simple models do not require massive data to fit or evaluate

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. Figure 5 shows the different data science process.

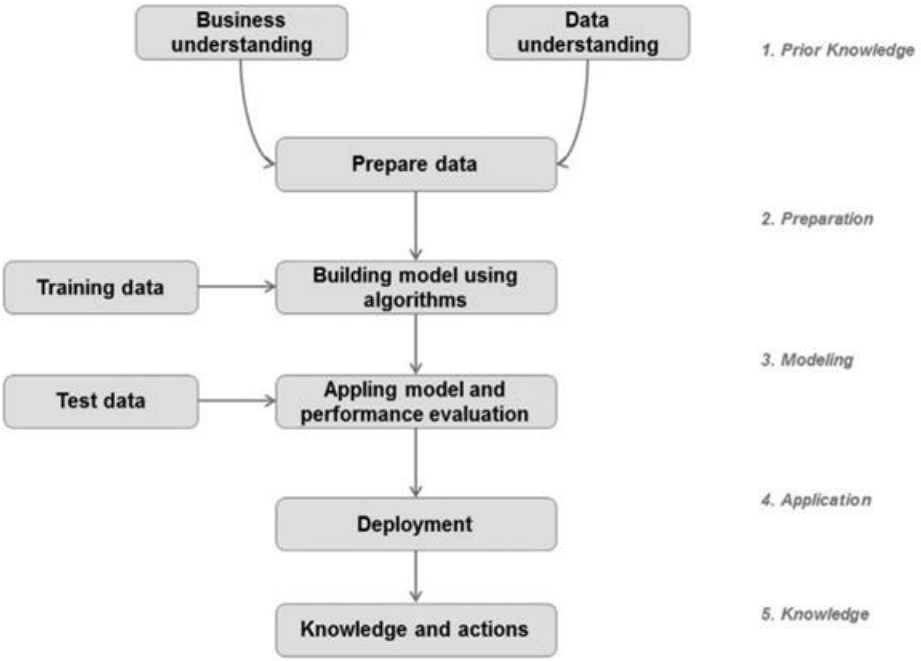


Figure 5: Data science process

The standard data science process involves:

- Understanding the problem
- Preparing the data samples
- Developing the model
- Applying the model on a dataset to see how the model may work in the real world
- Deploying and maintaining the models

The fundamental objective of any process that involves data science is to address the analysis question. The problem at hand could be a segmentation of customers, a prediction of climate

patterns, or a simple data exploration. The learning algorithm used to solve the business question could be a decision tree, an artificial neural network, or a scatterplot. The software tool to develop and implement the data science algorithm used could be custom coding, RapidMiner, R, Weka, SAS, Oracle Data Miner, Python, etc. The different data science process is:

1. **Prior Knowledge**

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

Prior knowledge includes:

- Objective
- Subject Area
- Data
- Causation Versus Correlation

1.1 Objective

- The data science process starts with a need for analysis, a question, or a business objective.
- This is the most important step in the data science process.
- 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.
- As an iterative process, it is common to go back to previous data science process steps, revise the assumptions, approach, and tactics.
- It is imperative to get the first step—the objective of the whole process—right.

1.2 Subject Area

- The process of data science uncovers hidden patterns in the dataset by exposing relationships between attributes.
- The problem is that it uncovers a lot of patterns.
- The false or spurious signals are a major problem in the data science process.
- 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.
- Hence, it is essential to know the subject matter, the context, and the business process generating the data.
- Understanding current models and business practices lays the foundation and establishes known knowledge.
- Analysis and mining the data provides the new knowledge that can be built on top of the existing knowledge.

1.3 Data

- Prior knowledge in the data should also be gathered.
- Understanding how the data is collected, stored, transformed, reported, and used is essential to the data science process.
- This part of the step surveys all the data available to answer the business question and narrows down the new data that need to be sourced.
- 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, etc.
- 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 an inferred model is only as good as the data used to create it.

1.4 Causation Versus Correlation

- The correlation between the input and output attributes doesn't guarantee causation.
- It is important to frame the data science question correctly using the existing domain and data knowledge.
- Can the credit score of the borrower be predicted based on interest rate?
- From the existing domain expertise, it is known that credit score influences the loan interest rate. Predicting credit score based on interest rate inverses the direction of the causal relationship. This question also exposes one of the key aspects of model building. The correlation between the input and output attributes doesn't guarantee causation.

2. Data Preparation

- Preparing the dataset to suit a data science task is the most time-consuming part of the process.
- It is extremely rare that datasets are available in the form required by the data science algorithms.
- Most of the data science algorithms would require 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.

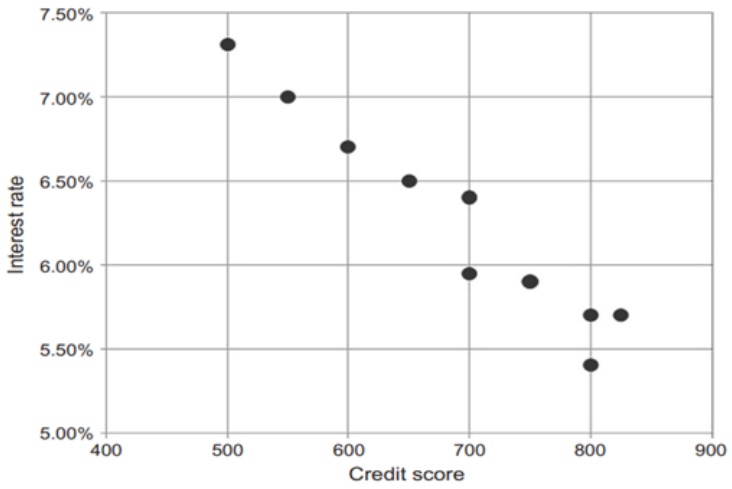
Data preparation phase includes:

- Data Exploration
- Data Quality
- Missing Values

- Data Types and Conversion
- Transformation
- Outliers
- Feature Selection
- Data Sampling

2.1 Data Exploration

- Data preparation starts with an in-depth exploration of the data and gaining a better understanding of the dataset.
- 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.
- Figure shows the scatterplot of credit score vs. loan interest rate and it can be observed that as credit score increases, interest rate decreases.



2.2 Data Quality

- Data quality is an ongoing concern wherever data is collected, processed, and stored.
- 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.
- Regardless, it is critical to check the data using data exploration techniques in addition to using prior knowledge of the data and business before building models.

2.3 Missing Values

- One of the most common data quality issues is that some records have missing attribute values.
- The first step of managing missing values is to 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.
- Example:
  Missing credit score values can be replaced with a credit score derived from the dataset (mean, minimum, or maximum value, depending on the characteristics of the attribute).
- This method is useful if the missing values occur randomly and the frequency of occurrence is quite rare.
- 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. 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.
- For example, 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.

2.4 Data Types and Conversion

- The attributes in a dataset can be of different types, such as continuous numeric (interest rate), integer numeric (credit score), or categorical.

- For example, the credit score can be expressed as categorical values (poor, good, excellent) or numeric score.
- Different data science algorithms impose different restrictions on the attribute data types.
- In case of linear regression models, the input attributes have to be numeric. If the available data are categorical, they must be converted to continuous numeric attribute.
- A specific numeric score can be encoded for each category value, such as poor =400, good = 600, excellent =700, etc.
- Similarly, numeric values can be converted to categorical data types by a technique called binning, where a range of values are specified for each category, for example, a score between 400 and 500 can be encoded as "low" and so on.

2.5 Transformation

- In some data science algorithms like k-NN, the input attributes are expected to be numeric and normalized, because 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.
- For example, consider income (expressed in USD, 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, a consistent comparison can be made between the two different attributes with different units.

2.6 Outliers

- Outliers are anomalies in a given dataset.
- Outliers may occur because of correct data capture (few people with income in tens of millions) or erroneous data capture (human height as 1.73 cm instead of 1.73 m).
- Regardless, 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.

2.7 Feature Selection

- Many data science problems involve a dataset with hundreds to thousands of attributes.
- In text mining applications, every distinct word in a document forms a distinct attribute in the dataset.
- Not all the attributes are equally important or useful in predicting the target.
- The presence of some attributes might be counterproductive.

- Some of the attributes may be highly correlated with each other, like annual income and taxes paid.
- 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.
- In general, the presence of more detailed information is desired in data science because discovering nuggets of a pattern in the data is one of the attractions of using data science techniques. But, as the number of dimensions in the data increase, data becomes sparse in high-dimensional space.
- This condition degrades the reliability of the models, especially in the case of clustering and classification.
- 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 and helps to synthesize a more effective explanation of the model.

2.8 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.
- In most cases, to gain insights, extract the information, and to build representative predictive models it is sufficient to work with samples.
- The error introduced by sampling impacts the relevancy of the model, but their benefits far outweigh the risks.
- Consider the example cases for predicting anomalies in a dataset (e.g., predicting fraudulent credit card transactions).
- The objective of anomaly detection is to classify the outliers in the data.
- These are rare events and often the dataset does not have enough examples of the outlier class.
- 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.

3. Modeling

- A model is the abstract representation of the data and the relationships in a given dataset.
- It is sufficient to have an overview of the learning algorithm, how it works, and determining what parameters need to be configured based on the understanding of the business and data.
- Figure shows the steps in the modeling phase of predictive data science.

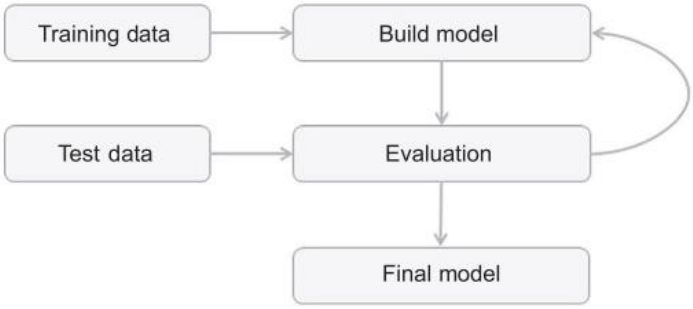


Figure 6: Steps in modelling phase

- Predictive algorithms require a prior known dataset to learn the model.
- Association analysis and clustering are descriptive data science techniques where there is no target variable to predict; hence, there is no test dataset. However, both predictive and descriptive models have an evaluation step.

Modelling phase includes:

- Training and Testing Datasets
- Learning Algorithms
- Evaluation of the Model
- Ensemble Modeling

3.1 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.
- 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.

3.2 Learning Algorithms

- The business question and the availability of data will dictate what data science task (association, classification, regression, etc.,) can to be used.
- Determines the appropriate data science algorithm within the chosen category.
- For example, within a classification task many algorithms can be chosen from: decision trees, rule induction, neural networks, Bayesian models, k-NN, etc. Likewise, within decision tree techniques, there are quite a number of variations of learning algorithms like classification and regression tree (CART), CHi-squared Automatic Interaction Detector (CHAID) etc.
- It is not uncommon to use multiple data science tasks and algorithms to solve a business question.

3.3 Evaluation of The Model

- The model generated in the form of an equation is generalized and synthesized from seven training records.
- The credit score in the equation can be substituted to see if the model estimates the interest rate for each of the seven training records.
- 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.
- 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.

3.4 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.
- Most of the practical data science applications utilize ensemble modeling techniques.
- At the end of the modeling stage of the data science process,

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
7. Validated the model against the test dataset.

4. 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
    - Assimilation.

4.1 Production Readiness

- Determines the critical qualities required for the deployment objective.
- Consider two business use cases: determining whether a consumer qualifies for a loan and determining the groupings of customers for an enterprise by marketing function.
- The critical quality in this application is the ability to find unique patterns amongst customers, not the response time of the model.
- The business application informs the choices that need to be made in the data preparation and modeling steps.

4.2 Technical Integration

- Data science automation or coding can be done using R or Python to develop models.

- Data science tools save time as they do not require the writing of custom codes to execute the algorithm.
- Allows the analyst to focus on the data, business logic, and exploring patterns from the data.
- The models created by data science tools can be ported to production applications by utilizing the Predictive Model Markup Language (PMML) or by invoking data science tools in the production application.
- PMML provides a portable and consistent format of model description which can be read by most data science tools.
- This allows the flexibility for practitioners to develop the model with one tool (e.g., RapidMiner) and deploy it in another tool or application.
- Some models such as simple regression, decision trees, and induction rules for predictive analytics can be incorporated directly into business applications and business intelligence systems easily.

4.3 Response Time

- Data science algorithms, like k-NN, are easy to build, but quite slow at predicting the unlabeled records.
- Algorithms such as the decision tree take time to build but are fast at prediction.
- There are trade-offs to be made between production responsiveness and modeling build time.
- The quality of prediction, accessibility of input data, and the response time of the prediction remain the critical quality factors in business application.

4.4 Model Refresh

- The key criterion for the ongoing relevance of the model is the representativeness of the dataset it is processing.
- The conditions in which the model is built change after the model is sent to deployment.
- For example, the relationship between the credit score and interest rate change frequently based on the prevailing macroeconomic conditions.
- Hence, the model will have to be refreshed frequently.
- The validity of the model can be routinely tested by using the new known test dataset and calculating the prediction error rate.
- If the error rate exceeds a particular threshold, then the model has to be refreshed and redeployed.
- Creating a maintenance schedule is a key part of a deployment plan that will sustain a relevant model.

4.5 Assimilation

- Objective is to o assimilate the knowledge gained from the data science analysis to the organization.

- For example, the objective may be finding logical clusters in the customer database so that separate marketing approaches can be developed for each customer cluster. Then the next step may be a classification task for new customers to bucket them in one of known clusters.
- The association analysis provides a solution for the market basket problem, where the task is to find which two products are purchased together most often.
- The challenge for the data science practitioner is to articulate these findings, establish relevance to the original business question, quantify the risks in the model, and quantify the business impact.

5. **Knowledge**

- To extract knowledge from these massive data assets, advanced approaches need to be employed, like data science algorithms, in addition to standard business intelligence reporting or statistical analysis.
- Though many of these algorithms can provide valuable knowledge, it is up to the practitioner to skillfully transform a business problem to a data problem and apply the right algorithm.
- Data science, like any other technology, provides various options in terms of algorithms and parameters within the algorithms. Using these options to extract the right information from data is a bit of an art and can be developed with practice.
- The data science process starts with prior knowledge and ends with posterior knowledge, which is the incremental insight gained.

COLLECTING DATA

- The most critical issue in any data science or modeling project is finding the right data set.
- Data collection includes hunting, scraping, and logging.
- Identifying viable data sources is an art, one that revolves around three basic questions:
    - 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?

Hunting

Who has the data, and how can you get it? The different sources of data are listed below.

Companies and Proprietary Data Sources

- Facebook, Google, Amazon, American Express
- most organizations have internal data sets of relevance to their business.

- 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.
- It is generally better for the company to provide well-behaved APIs than having cowboys repeatedly hammer and scrape their site.

Government Data Sources

- City, state, and federal governments have become increasingly committed to open data, to facilitate novel applications and improve how government can fulfill its mission.
- Government data differs from industrial data in that, in principle, it belongs to the People.
- If you cannot find what you need online after some snooping around, figure out which agency is likely to have it.

Academic Data Sets

- An increasing fraction of academic research involves the creation of large data sets.

Sweat Equity

- Sometimes you will have to work for your data, instead of just taking it from others.

Scraping

- Webpages 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. scraping programs were site-specific scripts hacked up to look for particular HTML patterns flanking the content of interest. This exploited the fact that large numbers of pages on specific websites are generated by programs themselves, and hence highly predictable in their format.
- Libraries in languages like Python make it easier to write robust spiders and scrapers.
- A spider/scraper for every popular website and made it available on SourceForge or Github.
- Certain spidering missions may be trivial, for example, hitting a single URL (uniform resource locator) at regular time intervals. Such patterns occur in monitoring, say, the sales rank of this book from its Amazon page. Somewhat more sophisticated approaches to spidering are based on the name regularity of the underlying URLs.
- 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. This is what Google does in indexing the web.

Logging

- Internal access to a web service, communications device, or laboratory instrument grant you the right and responsibility to log all activity for downstream analysis.
- Data collection from weblogs and sensing devices and IoT.
- 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.

CLEANING DATA

"Garbage in, garbage out" is the fundamental principle of data analysis. Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset.The road from raw data to a clean, analyzable data set can be a long one.

- This includes:
    - Errors vs. Artifacts
    - Data Compatibility
    - Dealing with Missing Values
    - Outlier Detection

Errors Vs. Artifacts

- Data errors represent information that is fundamentally lost in acquisition.
- Artifacts are generally 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.

Data Compatibility

- It makes no sense to compare weights of 123.5 against 78.9, when one is in pounds and the other is in kilograms.
- Data comparability issues arise whenever data sets are merged.
- Unit Conversions
- Numerical Representation Conversions

- Name Unification
- Time/Date Unification
- Financial Unification

Dealing with Missing values

- Not all data sets are complete.
- An important aspect of data cleaning is identifying fields for which data isn't there, and then properly compensating for them.
    - What is the year of death of a living person?
    - What should you do with a survey question left blank, or filled with an obviously outlandish value?
    - What is the relative frequency of events too rare to see in a limited-size sample?

- Numerical data sets expect a value for every element in a matrix.
- Setting missing values to zero is tempting, but generally wrong, because there is always some ambiguity as to whether these values should be interpreted as data or not.
- Is someone's salary zero because he is unemployed, or did he just not answer the question?
- The danger with using nonsense values as not-data symbols is that they can get misinterpreted as data when it comes time to build models.
- A linear regression model trained to predict salaries from age, education, and gender will have trouble with people who refused to answer the question.
- The simplest approach is to drop all records containing missing values.
- his works just fine when it leaves enough training data, provided the missing values are absent for non-systematic reasons.
- It is better to estimate or impute missing values, instead of leaving them blank.

General methods for filling in missing values are:

Heuristic-based imputation

- To make a reasonable guess for the value of certain fields. If I need to fill in a value for the year you will die, guessing birth year+80 will prove about right on average, and a lot faster than waiting for the final answer.

Mean value imputation

- Using the mean value of a variable as a proxy for missing values is generally sensible.

Random value imputation

- Select a random value from the column to replace the missing value.

Imputation by nearest neighbor

- This approach requires a distance function to identify the most similar records.

Imputation by interpolation

- Can use a method like linear regression to predict the values of the target column, given the other fields in the record. Such models can be trained over full records and then applied to those with missing values.

Outlier Detection

- Mistakes in data collection can easily produce outliers that can interfere with proper analysis.
- Outlier elements are often created by data entry mistakes.
- General sanity checking requires looking at the largest and smallest values in each variable/column to see whether they are too far out of line. This can best be done by plotting the frequency histogram and looking at the location of the extreme elements. Visual inspection can also confirm that the distribution looks the way it should, typically bell-shaped.
- They can also result from errors in scraping, say an irregularity in formatting causing a footnote number to be interpreted as a numerical value. Just because something is written down doesn't make it correct.
- It is too simple to just delete the rows containing outlier fields and move on. Outliers often point to more systematic problems that one must deal with.
- An interesting example concerns the largest dinosaur vertebra ever discovered. Measured at 1500 millimeters, it implies an individual that was 188 feet long. This is amazing, particularly because the second largest specimen ever discovered comes in at only 122 feet.
- The most likely explanation here is that this giant fossil never actually existed: it has been missing from the American Museum of Natural History for over a hundred years. Perhaps the original measurement was taken on a conventionally-sized bone and the center two digits accidentally transposed, reducing the vertebra down to 1050 millimeters.

EXPLORATORY DATA ANALYSIS

- Exploratory data analysis is the search for patterns and trends in a given data set.
- Visualization techniques play an important part in this quest.
- Looking carefully at your data is important for several reasons, including identifying mistakes in collection/processing, finding violations of statistical assumptions, and suggesting interesting hypotheses.

- Confronting a new data set

Basic steps that to do to get acquainted with any new data set:

Answer the basic questions

- Who constructed this data set, when, and why?
- How big is it?
- What do the fields mean?

Look for familiar or interpretable records

- Records are generally associated with a person, place, or thing that you already have some knowledge about, so you can put it in context and evaluate the soundness of the data you have on it. But if not, find a few records of special interest to get to know, perhaps the ones with the maximum or minimum values of the most important field.

Summary statistics

- Look at the basic statistics of each column.
- For categorical fields, like occupation, the analogous summary would be a report on how many different label types appear in the column, and what the three most popular categories are, with associated frequencies.
- Tukey's five number summary is a great start for numerical values, consisting of the extreme values (max and min), plus the median and quartile elements.

Pairwise correlations

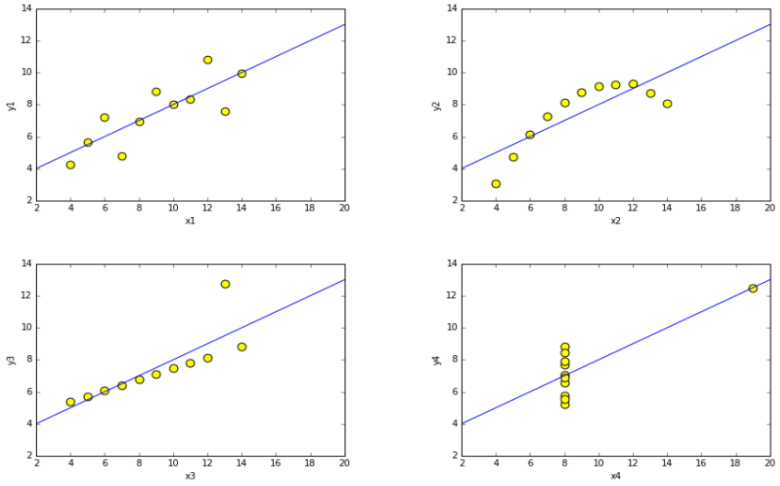
- A matrix of correlation coefficients between all pairs of columns gives an inkling of how easy it will be to build a successful model.
- There will be several features which strongly correlate with the outcome, while not strongly correlating with each other.
- Only one column from a set of perfectly correlated features has any value, because all the other features are completely defined from any single column.

Summary Statistics and Ascombe's Quartet

- Anscombe's quartet: four two-dimensional data sets, each with eleven points are shown below.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.31	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Var.	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Corr.	0.816		0.816		0.816		0.816	

- All four data sets have identical means for the x and y values, identical variances for the x and y values, and the exact same correlation between the x and y values.
- The dot plots of these data sets in they all look different, and tell substantially different stories.
- One trends linear, while a second looks almost parabolic. Two others are almost perfectly linear modulo outliers, but with wildly different slopes.
- Can instantly appreciate these differences with a glance at the scatter plot. Even simple visualizations are powerful tools for understanding what is going on in a data set.



VISUALIZATION TOOLS

An extensive collection of software tools is available to support visualization. They are listed below.

Exploratory data analysis

- Perform quick, interactive explorations of a given data set.

- Spreadsheet programs like Excel and notebook-based programming environments like iPython, R, and Mathematical are effective at building the standard plot types.
- The key here is hiding the complexity, so the plotting routines default to doing something reasonable but can be customized if necessary.

Publication/presentation quality charts

- Excel is very popular but it does not produce the best possible graphs/plots.
- The best visualizations are an interaction between scientist and software, taking full advantage of the flexibility of a tool to maximize the information content of the graphic.
- Plotting libraries like MatPlotLib or Gnuplot support a host of options enabling your graph to look exactly like you want it to.
- The statistical language R has a very extensive library of data visualizations.

Interactive visualization for external applications

- Building dashboards that facilitate user interaction with proprietary data sets is a typical task for data science-oriented software engineers.
- This builds tools that support exploratory data analysis for less technically skilled, more application-oriented personnel.
- Such systems can be readily built in programming languages like Python, using standard plotting libraries.
- There is class of third-party systems for building dashboards, like Tableau.
- These systems are programmable at a higher-level than other tools, supporting particular interaction paradigms and linked-views across distinct views of the data.

Developing a Visualization Aesthetic

Distinguishing good/bad visualizations requires developing a design aesthetic, and a vocabulary to talk about data representations.The following principles makes a chart or graph informative and beautiful, basing a design aesthetic.

- Maximize data-ink ratio
- Minimize the lie factor
- Minimize chartjunk
- Use proper scales and clear labeling
- Make effective use of color
- Exploit the power of repetition

Maximizing data-ink ratio

- The visualization is supposed to show off your data.
- In any graphic, some of the ink is used to represent the actual underlying data, while the rest is employed on graphic effects.

- Visualizations should focus on showing the data itself.
- Maximizing the data-ink ratio lets the data talk.
- The same information could be conveyed by plotting a point of the appropriate height, and would clearly be an improvement. The ratio is given as:

$$\text{Data-Ink Ratio} = \frac{\text{Data-Ink}}{\text{Total ink used in graphic}}$$

Minimizing the Lie Factor

- A visualization seeks to tell a true story about what the data is saying.
- The baldest form of lie is to fudge your data, but it remains quite possible to report your data accurately, yet deliberately mislead your audience about what it is saying.
- Graphical integrity requires minimizing this lie factor, by avoiding the techniques which tend to mislead.
- The lie factor of a chart as:

$$\text{lie factor} = \frac{\text{(size of an effect in the graphic)}}{\text{(size of the effect in the data)}}$$

Techniques to minimize lie factor are:

- Presenting means without variance
  The data values {100, 100, 100, 100, 100} and {200, 0, 100, 200, 0} tell different stories, even though both means are 100. If you cannot plot the actual points with the mean, at least show the variance, to make clear the degree to which the mean reflects the distribution.

- Presenting interpolations without the actual data
  Regression lines and fitted curves are effective at communicating trends and simplifying large data sets. But without showing the data points it is based on, it is impossible to ascertain the quality of the fit.

- Distortions of scale
  The aspect ratio of a figure can have a huge effect on how we interpret what we are seeing.

- Eliminating tick labels from numerical axes
  Even the worst scale distortions can be completely concealed by not printing numerical reference labels on axes. Only with the numerical scale markings can the actual data values be reconstructed from the plot.

- Hide the origin point from plot
  The implicit assumption in most graphs is that the range of values on the y-axis goes from zero to y max.

Minimizing chart junk

- Extraneous visual elements distract from the message the data is trying to tell.
- Interpretation of chartjunk is that it is any design element that is counterproductive, actually detracting from a data visualization rather than adding value to it.
- A bar plot, a perfectly sound way to represent time series data, and is drawn using conventional default, options using a plotting package.
- We can simplify the plot by removing elements to make the data stand better.
- The different ways to simplify the chart are:
  - Jail breaks your data
  - Stop throwing shade
  - Think outside the box
  - Make missing ink work for the graph

Effective use of color and shading

- Colors are increasingly assumed as part of any graphical communication.
- Colors play two major roles in charts, namely marking class distinctions and encoding numerical values.
- Representing points of different types, clusters, or classes with different colors encodes another layer of information on a conventional dot plot.
- The classes be easily distinguishable from each other, by using bold primary colors.
- Losses should be printed in red ink, environmental causes associated with green, nations with their flag colors, and sports teams with their jersey colors. Coloring points to represent males as blue and females as red offers a subtle clue to help the viewer interpret a scatter plot.
- Much better are color scales based on a varying either brightness or saturation.
- The brightness of a color is modulated by blending the hue with a shade of gray, somewhere between white and black.
- Saturation is controlled by mixing in a fraction of gray, where 0 produces the pure hue, and 1 removes all color.
- Large areas on plots should be shown with unsaturated colors. The converse is true for small regions, which stand out better with saturated colors.

Chart Types

1. Tabular Data

- Tables of numbers can be beautiful things, and are very effective ways to present data. Although they may appear to lack the visual appeal of graphic presentations, tables have several advantages over other representations, including:
- Representation of precision
- Representation of scale
- Multivariate visualization
- Heterogeneous data
- Compactness

2. Dot and Line Plots
3. Scatter Plots
4. Bar Plots and Pie Charts
5. Histograms
6. Data Maps

Languages for Data Science

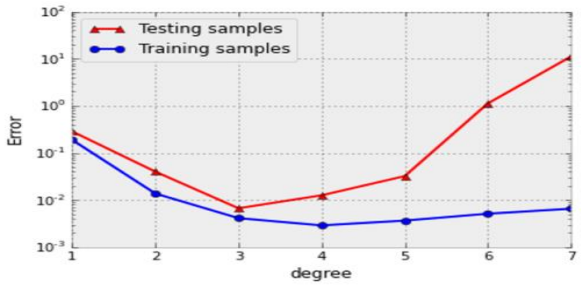
- Every sufficiently powerful programming language is capable of expressing any algorithm worth computing. But in practice, certain programming languages prove much better than others at specific tasks.
- Better here might denote easier for the programmer or perhaps more computationally efficient, depending upon the mission at hand. The primary data science programming languages are:
    - Python
    - Perl
    - R
    - Matlab
    - Java and C/C++
    - Mathematica/Wolfram Alpha
    - Excel

The Importance of Notebook Environments

- The deliverable result of every data science project should be a computable notebook tying together the code, data, computational results, and written analysis of what you have learned in the process.
- The reason this is so important is that computational results are the product of long chains of parameter selections and design decisions.
- Several problems that are solved by notebook computing environments:
    - Computations need to be reproducible
    - Computations must be tweakable
    - Data pipelines need to be documented
    - The Importance of Notebook Environments

- Figure Presents an excerpt from a Jupyter/IPython notebook, showing how it integrates code, graphics, and documentation into a descriptive document which can be executed like a program.

```
In [40]: degrees = range(1, 8)
         errors = np.array([regressor3(d) for d in degrees])
         plt.plot(degrees, errors[:, 0], marker='^', c='r', label='Testing samples
         plt.plot(degrees, errors[:, 1], marker='o', c='b', label='Training sample
         plt.yscale('log')
         plt.xlabel("degree"); plt.ylabel("Error")
           = plt.legend(loc='best')
```



By sweeping the degree we discover two regions of model performance:

- **Underfitting** (degree < 3): Characterized by the fact that the testing error will get lower if we increase the model capacity.
- **Overfitting** (degree > 3): Characterized by the fact the testing will get higher if we increase the model capacity. Note, that the training error is getting lower or just staying the same!.

Standard Data Formats

- The best computational data formats have several useful properties:
  - They are easy for computers to parse
  - They are easy for people to read
  - They are widely used by other tools and systems
- The most important data formats/representations are:
  - CSV (comma separated value) files
  - XML (eXtensible Markup Language)
  - SQL (structured query language) databases
  - JSON (JavaScript Object Notation)
  - Protocol buffers

CSV (Comma Separated Value) Files

- These files provide the simplest, most popular format to exchange data between programs.
- That each line represents a single record, with fields separated by commas.
- The csv format provides ways to escape code such special characters so they are not treated as delimiters.
- A better alternative is to use a rarer delimiter character, as in tsv or tab separated value files.

- The best test of whether your csv file is properly formatted is whether Microsoft Excel or some other spreadsheet program can read it without hassle.

XML (Extensible Markup Language)

- Structured but non-tabular data are often written as text with annotations.
- The natural output of a named-entity tagger for text wraps the relevant substrings of a text in brackets denoting person, place, or thing.
- All webpages are written in HTML, the hypertext markup language which organizes documents using bracketing commands like and to enclose bold faced text.
- XML is a language for writing specifications of such markup languages. A proper XML specification enables the user to parse any document complying with the specification.

SQL (Structured Query Language) Databases

- Spreadsheets are naturally structured around single tables of data.
- In contrast, relational databases prove excellent for manipulating multiple distinct but related tables, using SQL to provide a clunky but powerful query language.
- Any reasonable database system imports and exports records as either csv or XML files, as well as an internal content dump.
- The internal representation in databases is opaque, so it really isn't accurate to describe them as a data format. Still, I emphasize them here because SQL databases generally prove a better and more powerful solution than manipulating multiple data files in an ad hoc manner.

JSON (Javascript Object Notation)

- This is a format for transmitting data objects between programs.
- It is a natural way to communicate the state of variables/data structures from one system to another.
- This representation is basically a list of attribute-value pairs corresponding to variable/field names, and the associated values
- Library functions that support reading and writing JSON objects are readily available in all modern programming languages, it has become a very convenient way to store data structures for later use.
- JSON objects are human readable, but are quite cluttered-looking, representing arrays of records compared to CSV files. Use them for complex structured objects, but not simple tables of data.

```
{"employees":[
    {"firstName":"John", "lastName":"Doe"},
    {"firstName":"Anna", "lastName":"Smith"},
    {"firstName":"Peter", "lastName":"Jones"}
]}
```

Protocol Buffers

- These are a language/platform-neutral way of serializing structured data for communications and storage across applications.
- They are essentially lighter weight versions of XML (where you define the format of your structured data), designed to communicate small amounts of data across programs like JSON.
- This data format is used for much of the intermachine communication at Google. Apache Thrift is a related standard, used at Facebook.

A Taxonomy of Models

1. 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, yielding very effective models.
- Richer mathematical descriptions include higher-order polynomials, logarithms, and exponentials.
- These permit models that fit training data much more tightly than linear functions can.
- It is much harder to find the best possible coefficients to fit non-linear models. But we don't have to find the best possible fit.
- Deep learning methods, based on neural networks, offer excellent performance despite inherent difficulties in optimization.

2. Black box vs. Descriptive Models

- Black boxes are devices that do their job, but in some unknown manner. Stuff goes in and stuff comes out, but how the sausage is made is completely impenetrable to outsiders.
- descriptive, meaning they provide some insight into why they are making their decisions. Theory-driven models are generally descriptive, because they are explicit implementations of a particular well-developed theory.
- 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.
- Blackbox modeling techniques such as deep learning can be extremely effective.

3. 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. The model might be a discrete event simulation.

- 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.

4. 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.
- That 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.

5. Flat vs. Hierarchical Models

- 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. Certain subproblems lend themselves to theory-based, first-principle models, which can then be used as features in a general data-driven model. Explicitly 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.

EVALUATING MODELS

Evaluating Classifiers

- Evaluating a classifier means measuring how accurately our predicted labels match the standard labels in the evaluation set. For the common case of two distinct labels or classes (binary classification), we typically call 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.
- There are four possible results of what the classification model could do on any given instance, which defines the confusion matrix or contingency table.

		Predicted Class	
		Yes	No
Actual Class	Yes	True Positives (TP)	False Negatives (FN)
	No	False Positives (FP)	True Negatives (TN)

- True Positives (TP): Here our classifier labels a positive item as positive, resulting in a win for the classifier.
- True Negatives (TN): Here the classifier correctly determines that a member of the negative class deserves a negative label. Another win.
- False Positives (FP): The classifier mistakenly calls a negative item as a positive, resulting in a "type I" classification error.
- False Negatives (FN): The classifier mistakenly declares a positive item as negative, resulting in a "type II" classification error.

Accuracy, Precision, Recall, and F-Score

- There are several different evaluation statistics which can be computed from the true/false positive/negative counts.
- The reason we need so many statistics is that 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.
- In contrast, 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.

Accuracy

- The ratio of the number of correct predictions over total predictions.
- By multiplying such fractions by 100, we can get a percentage accuracy score.
- Accuracy is a sensible number which is relatively easy to explain, so it is worth providing in any evaluation environment.
- Accuracy alone has limitations as an evaluation metric, particularly when the positive class is much smaller than the negative class.

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

Precision

- Precision is the ratio of the correctly +ve labeled by our program to all +ve labeled.
- Precision answers the following: How many of those who we labeled as diabetic are actually diabetic?
- numerator: +ve labeled diabetic people.
- denominator: all +ve labeled by our program (whether they're diabetic or not in reality).

$$precision = \frac{TP}{(TP + FP)}$$

Recall (Aka sensitivity)

- Recall is the ratio of the correctly +ve labeled by our program to all who are diabetic in reality.
- Recall answers the following question: Of all the people who are diabetic, how many of those we correctly predict?
- numerator: +ve labeled diabetic people.
- denominator: all people who are diabetic (whether detected by our program or not)

$$recall = \frac{TP}{(TP + FN)}$$

F-score

- The F-score (or sometimes F1-score) is such a combination, returning the harmonic mean of precision and recall.
- F1 Score is best if there is some sort of balance between precision (p) & recall (r) in the system. Oppositely F1 Score isn't so high if one measure is improved at the expense of the other.
- For example, if P is 1 & R is 0, F1 score is 0.

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Notes:

- 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 classifier.

<u>Evaluating Value Prediction Models</u>

- *Absolute error*: The value $\Delta = y' - y$ has the virtue of being simple and symmetric, so the sign can distinguish the case where $y' > y$ from $y > y'$.

- *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 miles.

  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: $\epsilon = (y - y')/y$. Absolute error weighs instances with larger values of y as more important than smaller ones, a bias corrected when computing relative errors.

- *Squared error*: The value $\Delta^2 = (y' - y)^2$ is always positive, and hence these values can be meaningfully summed. Large errors values contribute disproportionately to the total when squaring: Δ^2 for $\Delta = 2$ is four times larger than Δ^2 for $\Delta = 1$. Thus outliers can easily come to dominate the error statistic in a large ensemble.

- Mean squared error (MSE)

  It weighs each term quadratically, outliers have a disproportionate effect. Thus, median squared error might be a more informative statistic for noisy instances.

$$MSE(Y, Y') = \frac{1}{n}\sum_{i=1}^{n}(y_i' - y_i)^2$$

- Root Mean squared error.

  Root mean squared (RMSD) error is simply the square root of mean squared error:

$$RMSD(\Theta) = \sqrt{MSE(Y, Y')}.$$