LECTURE 1

INTRODUCTION TO DATA SCIENCE

DR. PRAPASSORN TANTIPHANWADI
INDUSTRIAL ENGINEERING, FACULTY OF ENGINEERING AT KHAMPAENGSAEN
DECEMBER 2565

CONTENT

- What is Data Science?
- Data Science vs. AI vs. ML
- Data Types in Big Data
- Data Science Process
- Data Preparation
- Python Installation

DATA SCIENCE, AI AND ML

WHAT IS DATA SCIENCE?

- Data science stands for working with large amounts of unstructured information (otherwise known as big data).
- Meteorological data or statistics of search queries in Google are examples.
- The key takeaway words here are 'huge volume' and 'unstructuredness.'
- In fact, anything surrounding data selection, preparation, and analysis relates to data science.
- At the core of data science is big data—raw information stored in enterprise data warehouses.

- **Data Scientists** use mathematical statistics and machine learning methods to work with tons of information.
- A Data Scientist understands data insights and sees the figures. Their job is to analyze big data to make predictions.
- The result of their work is a predictive model—a software algorithm that finds the best solution to the problem.

DATA SCIENCE, AI AND ML

WHAT IS DATA SCIENCE?

https://towardsdatascience.com/

DATA SCIENCE, AI AND ML

WHAT'S ARTIFICIAL INTELLIGENCE (AI)?

- Artificial intelligence is **a system** or a machine that can mimic human behavior.
- AI systems show at least some of the following behaviors:
  - Planning / Learning / Problem solving / Perception / Motion / Social intelligence and creativity
- AI is not a format or a function; it's the ability to think and analyze data.
- Artificial intelligence is not intended to replace humans with smart robots. Its goal is actually to expand human skills and capabilities.

  For examples:
  - Chatbots use AI to analyze customer requests and answer them.
  - Smart assistants use AI to extract information from large datasets and optimize planning.
  - Autonomous systems that recommend content for viewers based on other content they have consumed.
  - Amazon used AI to create self-driving cars and robots they now use for delivery.

https://www.youtube.com/watch?v=peaKnkNX4vc&t=19s

DATA SCIENCE, AI AND ML

WHAT'S MACHINE LEARNING (ML)?

- Machine learning is one of AI's branches. At a high level, ML is about teaching a computer how to make accurate predictions when it is fed with data.
- For example,
  - such a system could detect whether an apricot or an apple is in a picture.
  - It can spot people that cross the road in front of a self-driving vehicle.
  - ML can also distinguish regular emails from spam.
  - It can even recognize speech to provide captions on YouTube.
  - Netflix takes advantage of predictive analytics to improve recommendations to its users. ML-powered algorithms analyze users' preferences and 'understand' which movies they love most.

How Does Netflix Use Artificial Intelligence (AI) and Big Data?

https://www.youtube.com/watch?v=8M5n3uhWKHE

How Netflix Uses Machine Learning?

https://www.youtube.com/watch?v=2ZhhGEiCotI

DATA SCIENCE, AI AND ML

DATA SCIENCE VS. AI VS. ML

- **Data science** is not limited to algorithms or statistical aspects; it covers the whole spectrum of data processing. Besides, Data Scientists use AI to interpret the past, present and future.
- **Artificial intelligence** works with models that make machines act like humans. The computer system, in one way or another, imitates human behavior.
  - AI works on automating business processes and encouraging machines to work like humans.
  - It is mostly associated with human-AI interaction gadgets – Google Home, Siri, Alexa.
- **Machine learning** is a subfield of AI, which enables a computer system to learn from data.
  - ML algorithms depend on data as they train on information delivered by data science.
  - Without data science, machine learning algorithms won't work as they train on datasets.

DATA SCIENCE, AI AND ML

Data science vs. AI vs. ML

Data Science

- based on strict analytical evidence
- deals with structured & unstructured data
- includes various data operations

Artificial Intelligence

- imparts human intellect to machines
- uses logic and decision trees
- includes machine learning

Machine Learning

- subset of AI
- uses statistical models
- machines improve with experience

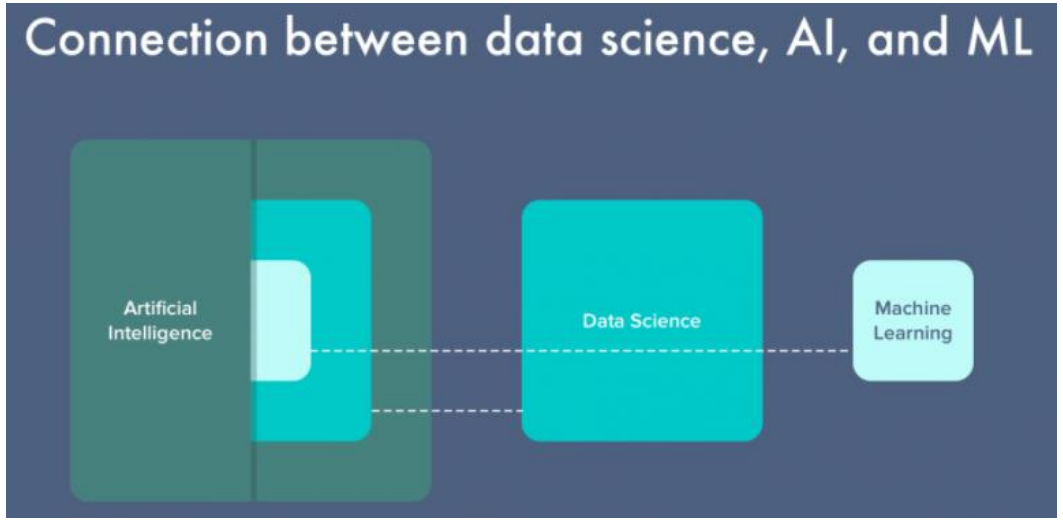
DATA SCIENCE, AI AND ML

Let's see how they could work together on an example of a self-driving car.

Machine learning: The car must recognize stop signs using cameras. We create a dataset with millions of images of street-side objects. We then train the machine learning algorithm to identify the images with stop signs.

AI: After recognizing the sign, the AI enabled car must apply the brakes—right on time, not too early nor too late.

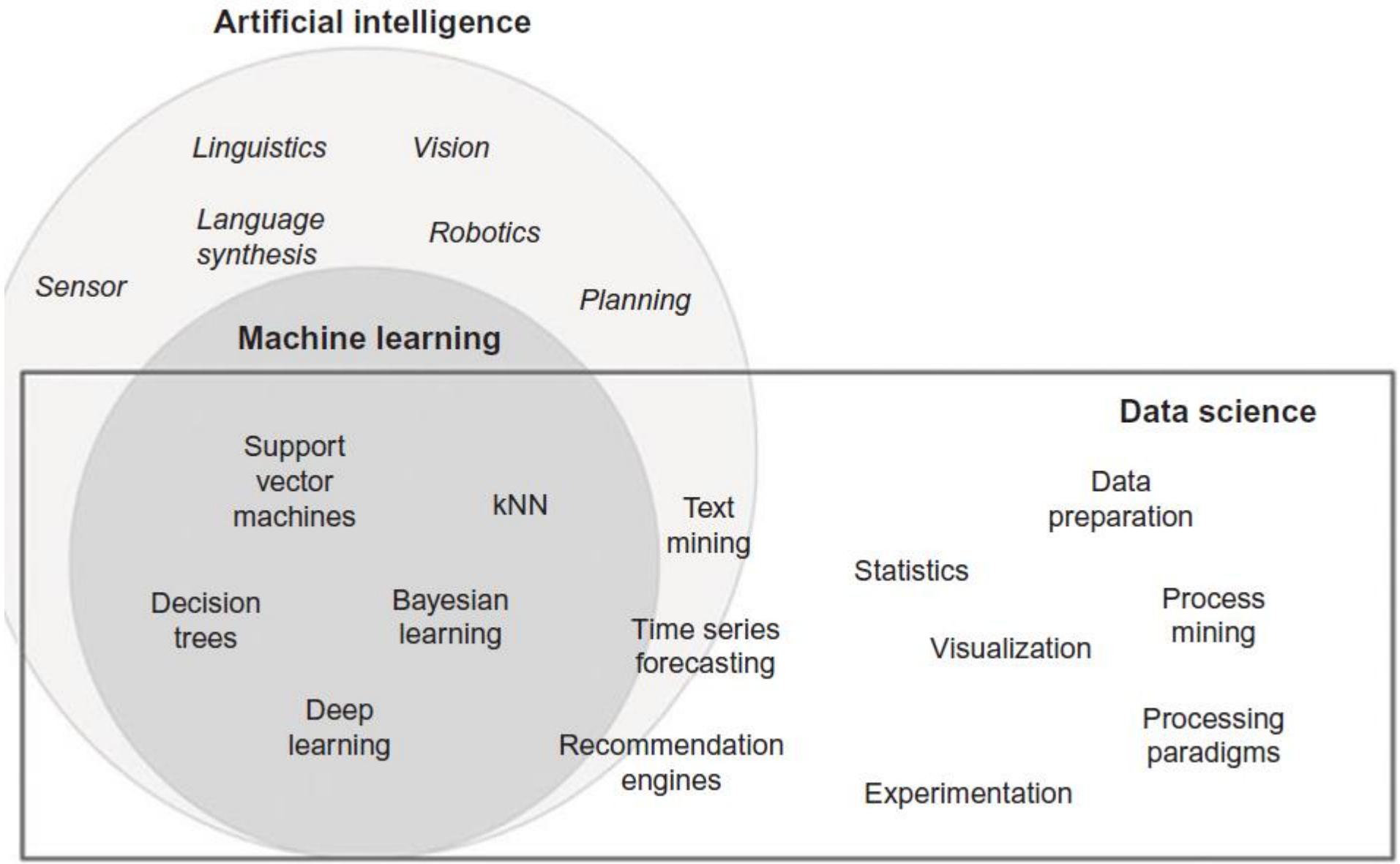
Data science: We run tests and see that in some cases the car doesn't apply brakes when it should. Once the test data is analyzed we see that there are more failed tests in the night than in the daytime. We add more nighttime images with stop signs to the dataset and get back to running tests.



Data Science vs Machine Learning – What's The Difference

https://www.youtube.com/watch?v=DP0TKM0S2-8

DATA SCIENCE, AI AND ML



DATA SCIENCE VS. BIG DATA VS. DATA ANALYTICS

WHAT IS BIG DATA?

- Is used to describe **immense volumes** of data, both **unstructured and structured**, big data can inundate a business on a day-to-day basis.
- Big data is used to analyze insights, which can lead to better decisions and strategic business moves.
- The definition of big data: "Big data is **high-volume**, and **high-velocity** or **high-variety information** assets
- The assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."

DATA TYPES IN BIG DATA

1. Structured Data
2. Unstructured Data
3. Semi-Structured Data
4. Subtypes of Data
5. Interacting with Data Through Programming

DATA SCIENCE VS. BIG DATA VS. DATA ANALYTICS

DATA TYPES IN BIG DATA

1. **Structured Data**
   - Any data that can be processed, accessed and stored as a fixed-format is named structured data.
   - Structured data in big data is the most straightforward to work with.

Example

An 'Employee' table in a database is a Structured Data Examples.

Employee_ID	Employee_Name	Gender	Department	Salary_In_ Lacs
1865	Meg Lanning	Female	HR	6,30,000
2145	Virat Kohli	Male	Finance	6,30,000
4500	Ellyse Perry	Female	HR	4,00,000
5475	Alyssa Healy	Female	HR	4,00,000
6570	Rohit Sharma	Male	Finance	5,30,000

DATA SCIENCE VS. BIG DATA VS. DATA ANALYTICS

DATA TYPES IN BIG DATA

2. Unstructured Data

- This is one of the types of big data where the data format of the relative multitude of unstructured files, for example, **image files, audio files, log files, and video files**, are incorporated.
- Any data which has an unfamiliar structure or model is arranged as unstructured data.
- Since the size is huge, unstructured data in big data has different difficulties as far as preparing for determining a value out of it.

Example The output returned by 'Google Search.'

The differences between structured and unstructured data in big data are:

- Qualitative vs Quantitative Data
- Defined vs Undefined Data
- Ease of Analysis
- Predefined Format vs Variety of Formats
- Storage in Data Houses vs Data Lakes

DATA SCIENCE VS. BIG DATA VS. DATA ANALYTICS

DATA TYPES IN BIG DATA

2. Unstructured Data

- This is one of the types of big data where the data format of the relative multitude of unstructured files, for example, **image files, audio files, log files, and video files**, are incorporated.
- Any data which has an unfamiliar structure or model is arranged as unstructured data.
- Since the size is huge, unstructured data in big data has different difficulties as far as preparing for determining a value out of it.

Example The output returned by 'Google Search.'

The differences between structured and unstructured data in big data are:

- Qualitative vs Quantitative Data
- Defined vs Undefined Data
- Ease of Analysis
- Predefined Format vs Variety of Formats
- Storage in Data Houses vs Data Lakes

DATA SCIENCE VS. BIG DATA VS. DATA ANALYTICS

DATA TYPES IN BIG DATA

2. **Unstructured Data**

Land data in a database or data lake, prepare the data, move selected data into a data warehouse, then perform reporting.

Land data in a data warehouse, analyze the data, then share data to use with other analytics and machine learning services.

DATA SCIENCE VS. BIG DATA VS. DATA ANALYTICS

DATA TYPES IN BIG DATA

3. Semi-Structured Data

- Is one of the types of big data related to the data containing both unstructured and structured data.

Example: Personal data stored in an XML file.

4. SUBTYPES OF DATA

- there are subtypes of data that hold some degree of relevance to the field of analytics.
- Frequently, these allude to the beginning of the data, for example, social media, machine, geospatial or event-triggered.

5. INTERACTING WITH DATA THROUGH PROGRAMMING

- **Scala:** a Java based-language. It was utilised to build up a few Apache items, including Spark, a significant part of the big data stages market.
- **R:** the language of decision. It is one of the top coding languages accessible for data control.
- **Python:** It is an open-source language and is viewed as one of the least complexes to learn.

DATA SCIENCE VS. BIG DATA VS. DATA ANALYTICS

APPLICATION OF BIG DATA

Big Data for Financial Services

- Banking services all use big data for their financial services.
- Common: amounts of multi-structured data living in multiple disparate systems, which big data can solve. As such, big data is used in several ways, including:
  - Customer analytics
  - Compliance analytics
  - Fraud analytics
  - Operational analytics

- Big Data in Communications
- Big data for retail

DATA SCIENCE VS. BIG DATA VS. DATA ANALYTICS

WHAT IS DATA ANALYTICS

- The term data analytics refers to the process of examining datasets to draw conclusions about the information they contain.
- Data analytic techniques enable you to take raw data and uncover patterns to extract valuable insights from it.
- Today, many data analytics techniques use specialized systems and software that integrate machine learning algorithms, automation and other capabilities.
- Data Scientists and Analysts use data analytics techniques in their research, and businesses also use it to inform their decisions.
- Data analysis can help companies better understand their customers, evaluate their ad campaigns, personalize content, create content strategies and develop products.

DATA SCIENCE VS. BIG DATA VS. DATA ANALYTICS

WHAT IS DATA ANALYTICS?

1. What insights can you gain from data analytics?

2. Data Analytics Technology

1. Artificial intelligence (AI) is the field of developing and using computer systems that can simulate human intelligence to complete tasks.
2. Machine Learning (ML) is a subset of AI that is significant for data analytics and involves algorithms that can learn on their own.
3. Data management: Before you can analyze data, you need to have procedures in place for managing the flow of data in and out of your systems and keeping your data organized.
4. Data Mining: refers to the process of sorting through large amounts of data to identify patterns and discover relationships between data points.
5. Predictive analytics: technology helps you analyze historical data to predict future outcomes and the likelihood of various outcomes occurring. These technologies typically use statistical algorithms and machine learning.

DATA SCIENCE VS. BIG DATA VS. DATA ANALYTICS

DATA ANALYTICS APPLICATIONS

Kumar shares that there **are four main types of data analysis**:

- **Descriptive analytics** reveal what happened in the past
- **Diagnostic analytics** answer why something happened
- **Predictive analytics** tell what will probably happen in the future
- **Prescriptive analytics** show what actions should be taken to make progress or avoid problems in the future

9 Exciting examples of data analytics driving change

1. Increasing the quality of medical care
2. Fighting climate change in local communities
3. Revealing trends for research institutions
4. Stopping hackers in their tracks
5. Serving customers with useful products
6. Driving marketing campaigns for businesses
7. Promoting smart energy usage for utility companies
8. Improving the insurance industry
9. Creating manufacturer warranties that make sense

DATA SCIENCE IN VARIOUS INDUSTRY

TOP TEN AREAS

1. **Finance** - Data science helps in risk assessment and monitoring, potential fraudulent behavior, payments, customer analysis, and experience, among many other utilizations.
2. **Healthcare** - By connecting pattern recognition, analytics, statistics, and deep learning algorithms, data science makes healthcare more efficient.
3. **Travel industry**
4. **Energy industry** - experiences major fluctuations in prices and higher costs of projects. Data scientists help in cutting costs, reducing risks, optimizing investments and improving equipment maintenance.
5. **Manufacturing**
6. **Gaming**
7. **Pharmaceuticals** - Connected to human health. For example,
    - a pharmaceutical company can utilize data science to ensure a more stable approach for planning clinical trials.
    - Another application can be seen before the trial even starts, by identifying suitable candidates based on their body structure such as chemical structure, medical history or other important characteristics. Data scientists read, evaluate, monitor and perform these analyses.

DATA SCIENCE IN VARIOUS INDUSTRY TOP 7 AREAS

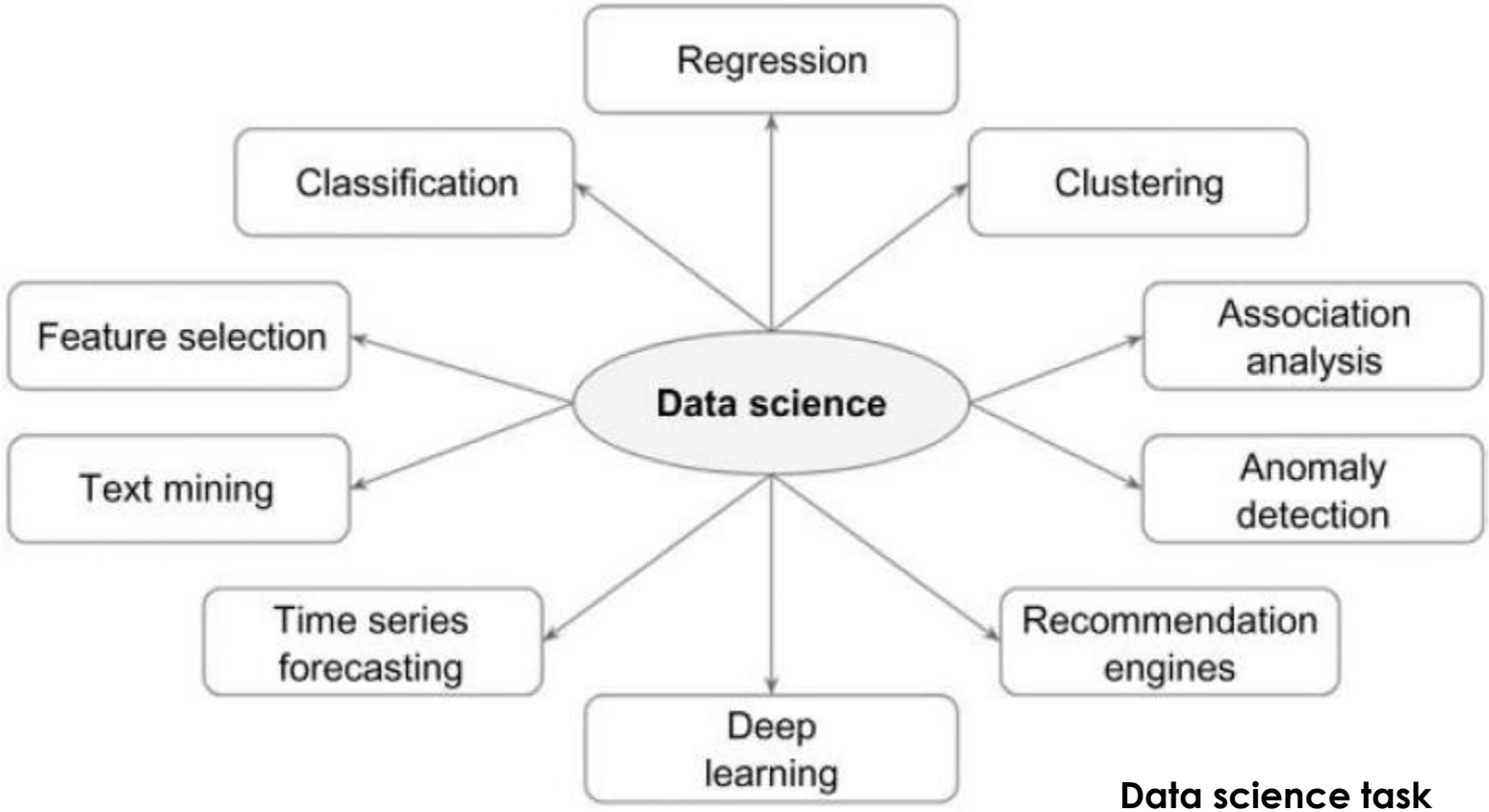
5. **Manufacturing** – Industry 4.0 (with the introduction of robotization and automation), the manufacturing industry keeps growing in need of data scientists where
    1. they can apply their knowledge of broad data management solutions through quality assurance, tracking defects, and increasing the quality of supplier relations.
    2. Similar to the energy industry, utilizing preventive maintenance to troubleshoot potential future equipment issues is another focus where data scientists can find good usage of their skills.
    3. Avoiding delays in the production process, implementing artificial intelligence and predictive analytics offers the possibility to manage frequent manufacturing issues: overproduction of products, logistics or inventory.
    4. In short, data scientists help in identifying inefficiencies and tuning the production process.

8. **Logistics & Transport** – Current applications by data-driven businesses within the industry include
    1. Reducing freight costs through delivery path optimization
    2. Dynamic price matching of supply to demand
    3. Warehouse optimization
    4. Forecasting demand
    5. Estimating total delivery times
    6. Extending the life of assets through finding patterns in usage data — identifying the need for maintenance

DATA SCIENCE PROBLEMS

Data science problems can also be classified into tasks such as:

- classification,
- regression,
- association analysis,
- clustering,
- anomaly detection,
- Recommendation engines,
- feature selection,
- time series forecasting,
- deep learning,
- and text mining



Data science task

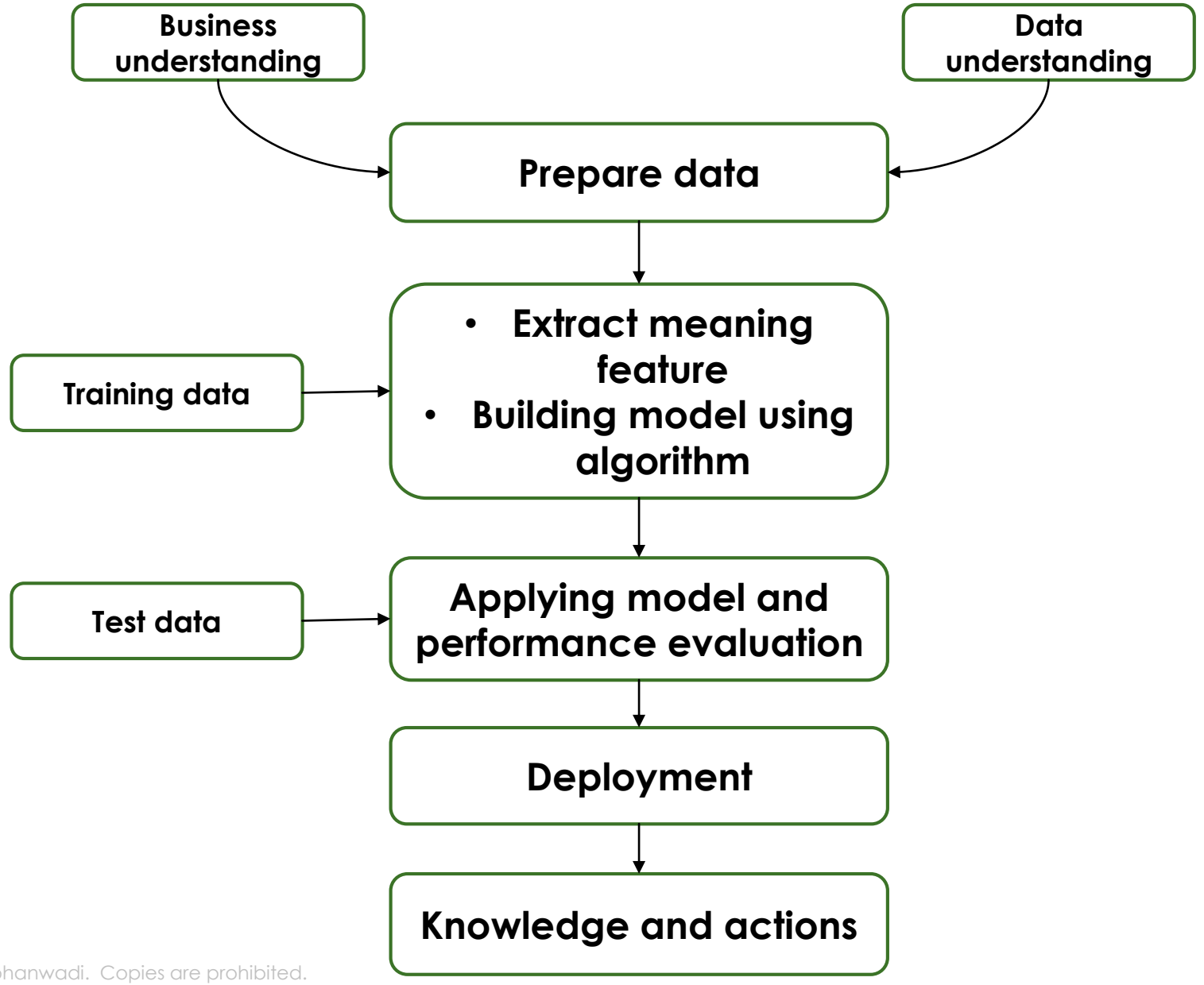
DATA SCIENCE ALGORITHMS

- An algorithm is a logical step-by-step procedure for solving a problem. In data science, it is the blueprint for how a particular data problem is solved.
- Many of the learning algorithms are recursive, where a set of steps are repeated many times until a limiting condition is met.
- Some algorithms also contain a random variable as an input and are aptly called randomized algorithms.
- A classification task can be solved using many different learning algorithms such as decision trees, artificial neural networks, k-NN, and even some regression algorithms.
- The choice of which algorithm to use depends on the type of dataset, objective, structure of the data, presence of outliers, available computational power, number of records, number of attributes, and so on.
- It is up to the data science practitioner to decide which algorithm (s) to use by evaluating the performance of multiple algorithms.
- There have been hundreds of algorithms developed in the last few decades to solve data science problems.
- Utilzing Python, Matlab, SAS, R,…

DATA SCIENCE TASK 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

DATA SCIENCE PROCESS



1. PRIOR KNOWLEDGE
2. PREPARATION
3. MODELING
4. APPLICATION
5. KNOWLEDGE

CORE ALGORITHM

- **Classification** is a model used to predict a target variable that is binary or categorical when a set of input variables are given. The model does this by learning the generalized relationship between the predicted target variable with all other input attributes.
- **Decision trees** approach the classification problem by partitioning the data into purer subsets based on the values of the input attributes.
- **Rule induction** is a data science process of deducing "if-then" rules from a dataset or from the decision trees. The decision rules explain an inherent relationship between the input attributes and the target labels.
- **Naïve Bayesian** algorithms provide a probabilistic way of building a model. This approach calculates the probability for each value of the class variable for given values of input variables.
- **k-NN** algorithm go through the trouble of extracting complex relationships from the data when the entire training dataset can be memorized and the relationship can appear to have been generalized?
- **Artificial neural networks**: Neurons are the nerve cells that connect with each other to form a
- biological neural network in our brain. The working of these interconnected nerve cells inspired the approach of some complex data problems.
- **Support vector machines (SVMs)** were developed to address optical character recognition problems: how can an algorithm be trained to detect boundaries between different patterns, and thus, identify characters?
- **Ensemble learners** are "meta" models where the model is a combination of several different individual models. If certain conditions are met, ensemble learners can gain from the wisdom of crowds and greatly reduce the generalization error in data science.

CORE ALGORITHM

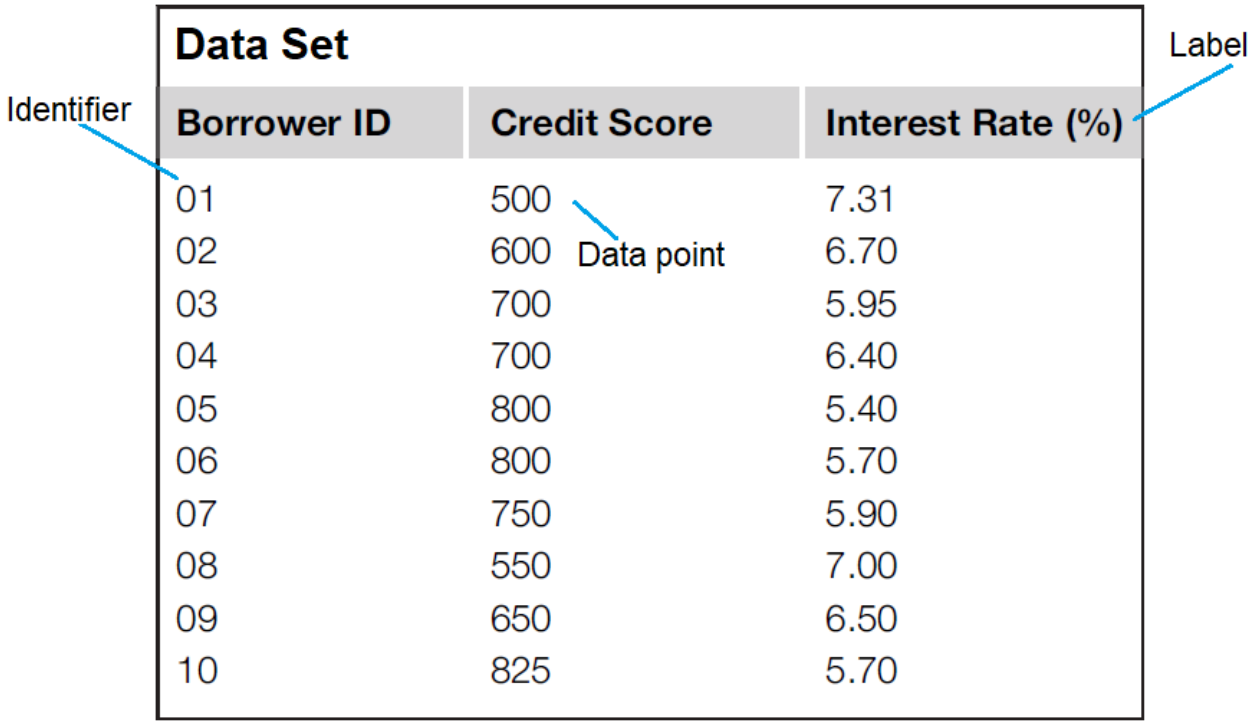
Mathematics

- **Linear regression**
- **Logistic regression**: addresses the issue of predicting a target variable thatmay be binary or binomial.
- **association analysis**: Associating an item in a transaction with another item in the transaction to determine the most frequently occurring patterns.
- **Clustering** is the data science task of identifying natural groups in the data.
- **k-means clustering** technique identifies a cluster based on a central prototype record
- **Forecasting** is a common application of time series analysis.
- **Deep Learning**, describes a set of algorithms to model high level abstractions in data

DATA

Terminology

- **A dataset** (example set) is a collection of data with a defined structure. This structure is also sometimes referred to as a "data frame".
- **A data point** (record, object or example) is a single instance in the dataset.
- **An attribute** (feature, input, dimension, variable, or predictor) is a single property of the dataset. Attributes can be numeric, categorical, date-time, text, or Boolean data types.
- **A label** (class label, output, prediction, target, or response) is the special attribute to be predicted based on all the input attributes.
- **Identifiers** are special attributes that are used for locating or providing context to individual records.
- **Causation Versus Correlation**

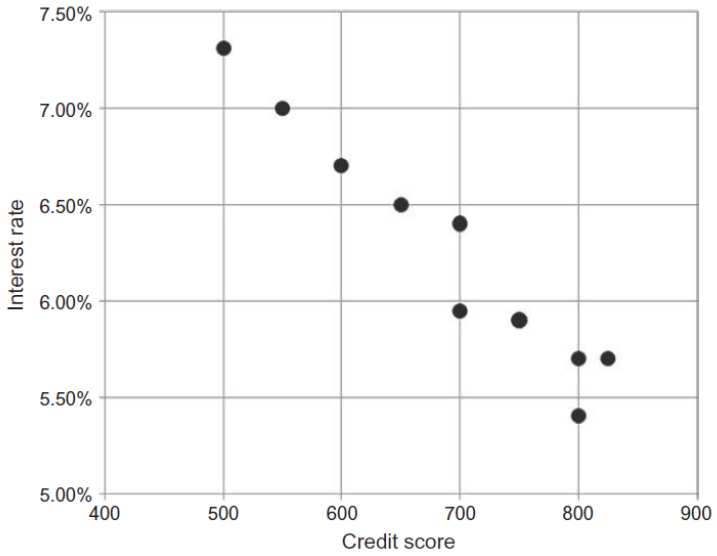


DATA

DATA PREPARATION

Preparing the dataset to suit a data science task.

- **Data Exploration** starts with an in-depth exploration of the data and gaining a better understanding of the dataset.
  - 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.
  - On the other hand, a visual plot of data points provides an insto condensed into one chart.



DATA

DATA PREPARATION

- **Missing Values:** The first step of managing missing values is to understand the reason behind why the values are missing.
  - Tracking the data lineage 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.
  - 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.
  - Some data science algorithms are good at handling records with missing values.
  - For example, k-nearest neighbor (k-NN) algorithm for classification tasks are often robust with missing values.

DATA

DATA PREPARATION

- **Data Types and Conversion:**
  - 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=5400, good=5600, excellent=5700, etc.
  - 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.

- **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.
  - One solution is to convert the range in each attribuite to a more uniform scale from 0 to 1 by normalization.

DATA PREPARATION

- **Outliers:**
  - Outliers may occur because of correct data capture or erroneous data capture.
  - Detecting outliers may be the primary purpose of some data science applications, like fraud or intrusion detection.

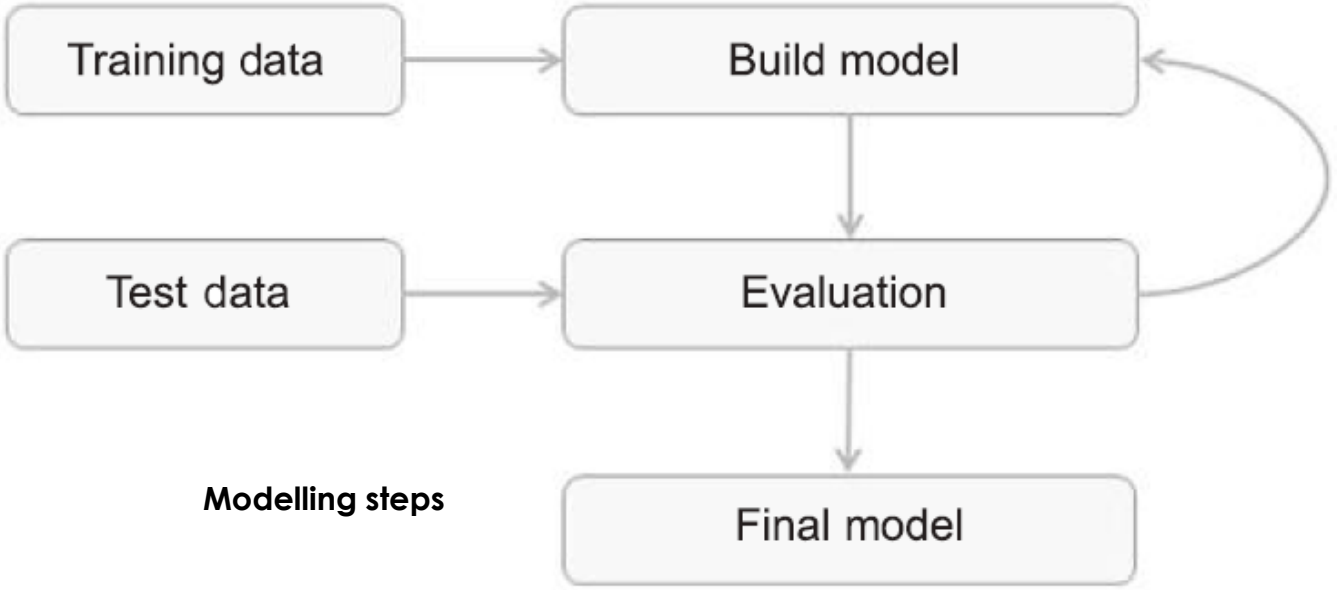
- **Feature Selection**
  - In practice, many data science problems involve a dataset with hundreds to thousands of attributes.
  - Not all the attributes are equally important or useful in predicting the target.
  - The presence of some attributes might be counterproductive. Some of the attributes may be highly correlated with each other.
  - 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 and helps to synthesize a more effective explanation of the model.

DATA PREPARATION

- **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.
  - In most cases, to gain insights, extract the information, and to build representative predictive
  - models it is sufficient to work with samples.
  - For data science applications, it is necessary to segment the datasets into training and test samples.
  - The training dataset is sampled from the original dataset using simple sampling or class label specific sampling.
    - **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.

DATA PREPARATION

- **Modelling**
  - A **model** is the abstract representation of the data and the relationships in a given dataset.
  - There are a few hundred data science algorithms in use today, derived from statistics, machine learning, pattern recognition, etc.
  - As a data science **practitioner,** it is sufficient to 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.
  - **Classification and regression** tasks are predictive techniques because they predict an outcome variable based on one or more input variables.
  - 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.



Modelling steps

DATA PREPARATION

- **Training and Testing Datasets**
  - 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.

Training data set

Borrower	Credit Score (*X*)	Interest Rate (*Y*) (%)
01	500	7.31
02	600	6.70
03	700	5.95
05	800	5.40
06	800	5.70
08	550	7.00
09	650	6.50

Test data set

Borrower	Credit Score (*X*)	Interest Rate (*Y*)
04	700	6.40
07	750	5.90
10	825	5.70

- **Evaluate model**
  - 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.

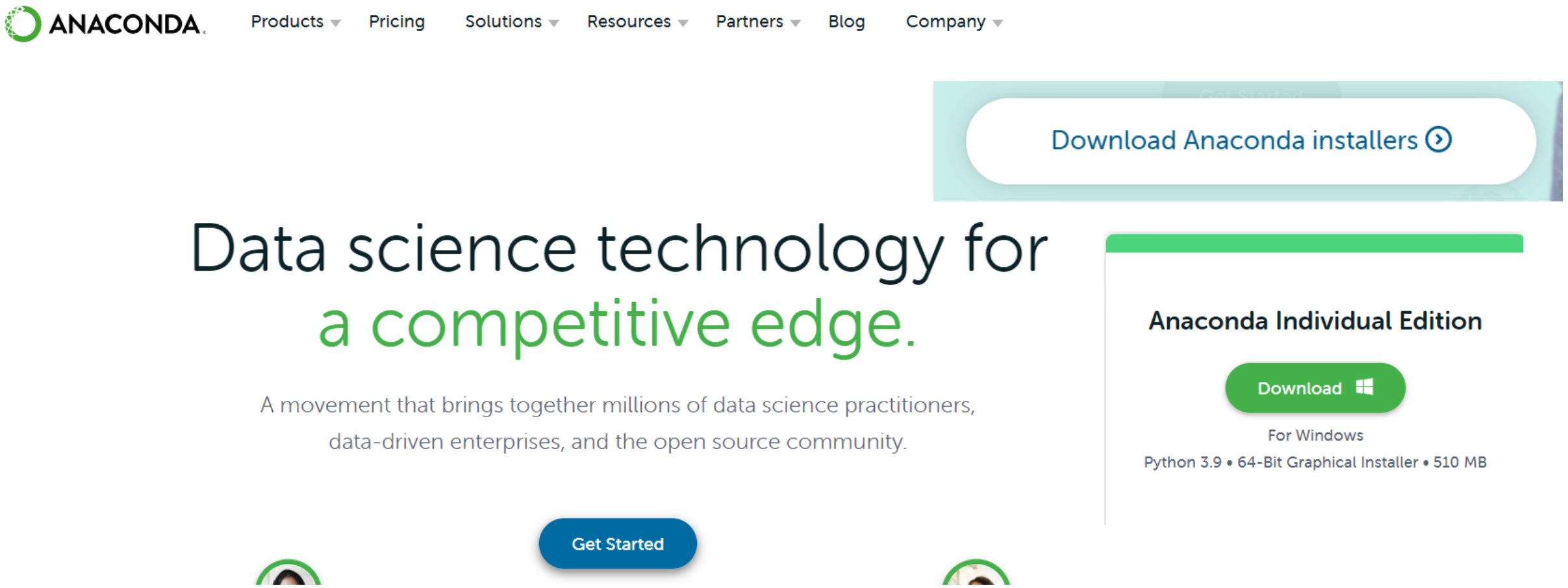
DATA PREPARATION

- **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.

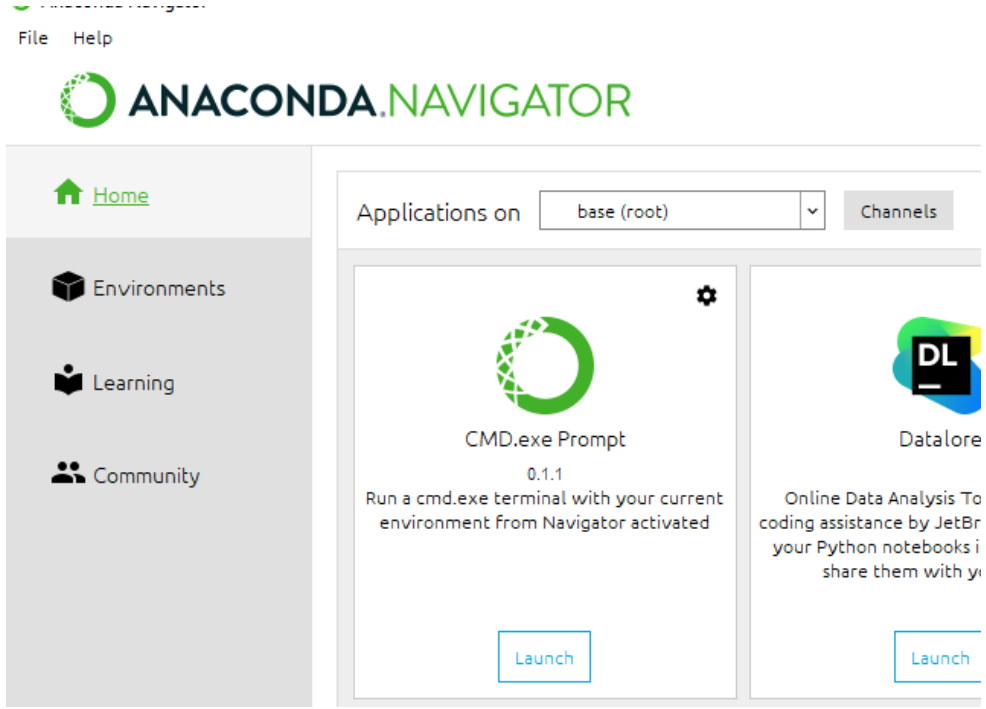
At the end of the modeling stage of the data science process, one has
(1) Analyzed the business question;
(2) sourced the data relevant to answer the question;
(3) selected a data science technique to answer the question;
(4) picked a data science algorithm and prepared the data to suit the algorithm;
(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.
This model can now be used to predict the output.

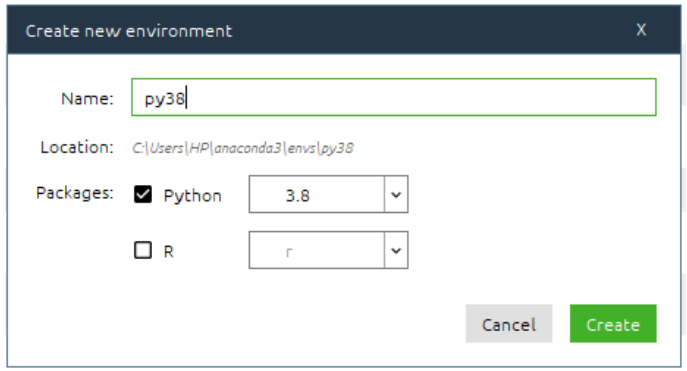
DATA SCIENCE VS. PYTHON

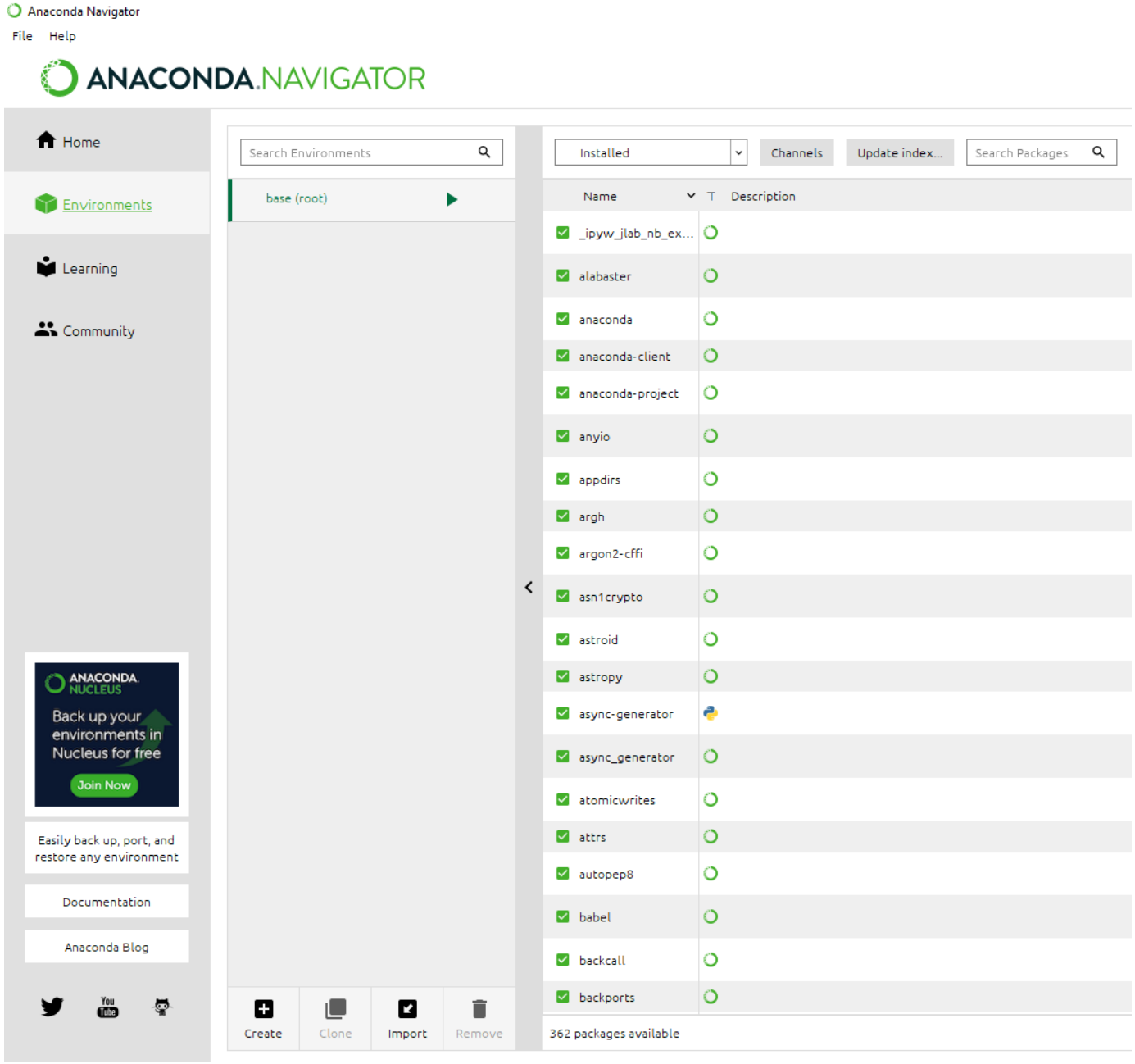
- Install anaconda https://www.anaconda.com/
- Install python, tensorflow



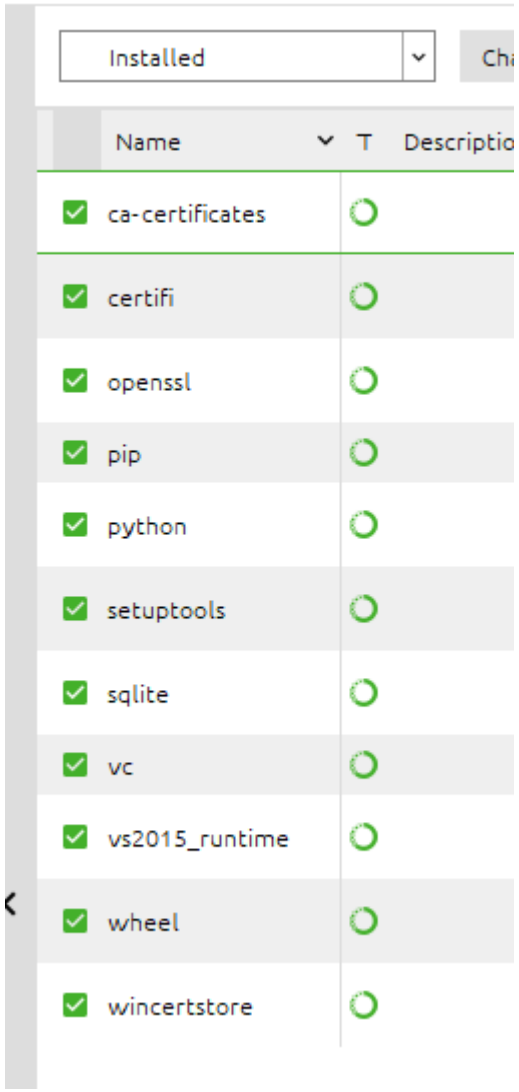
CREATE ENVIRONMENT

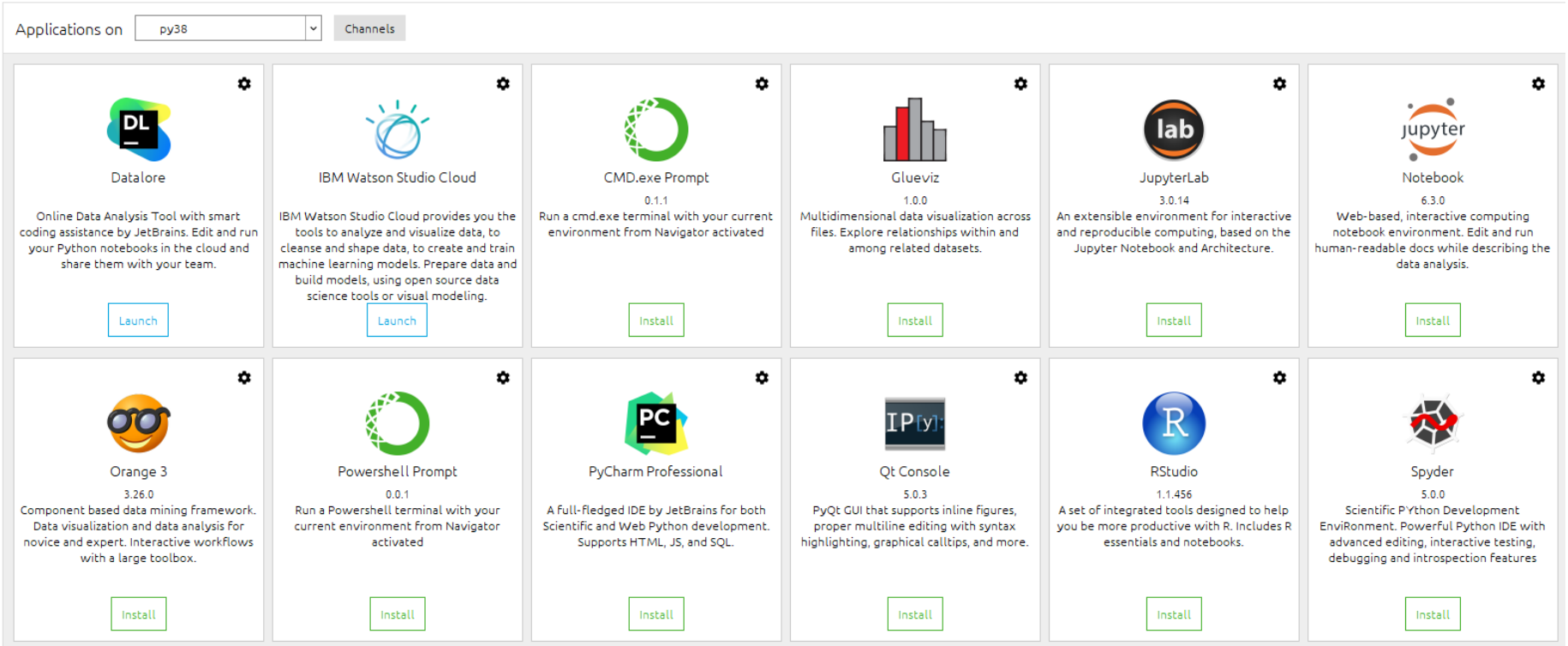




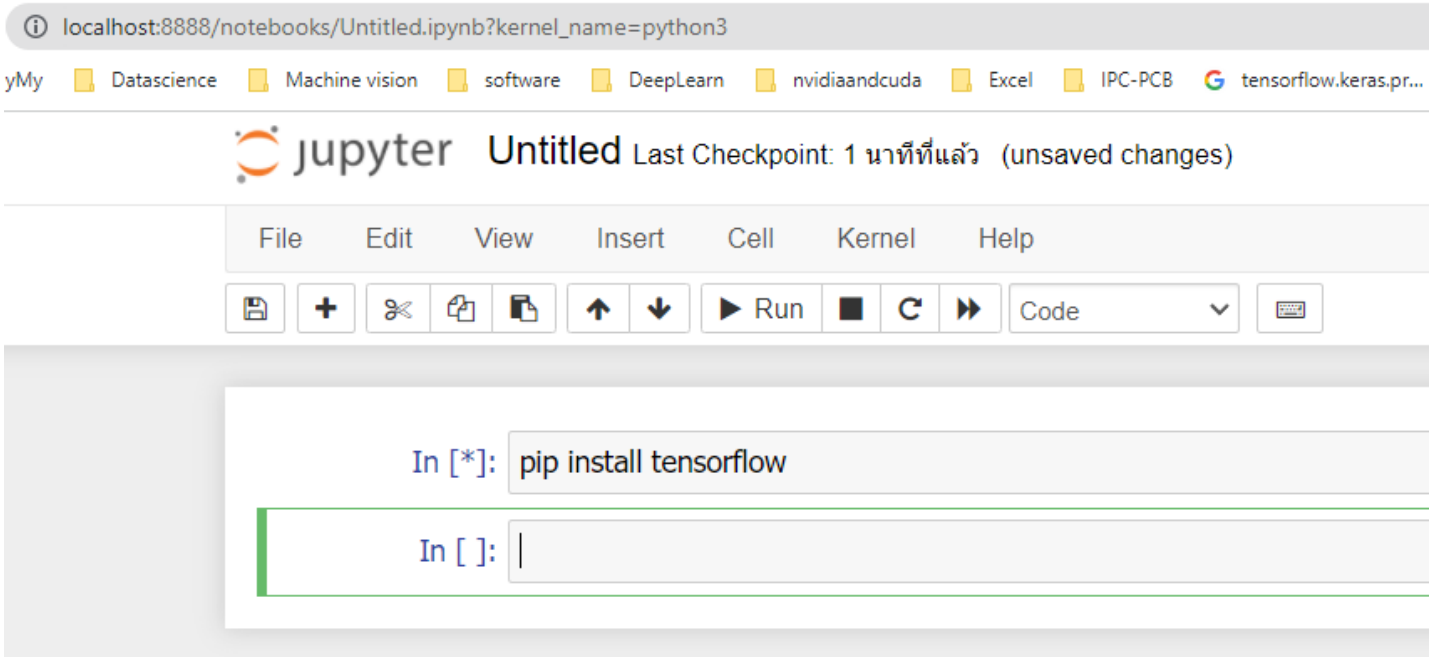


CREATE ENVIRONMENT





CREATE ENVIRONMENT



n1-modules-0.2.8 requests-oauthlib-1.3.0 rsa-4.7.2 tensorboard-2.7.0 tensorboard-data-server-0.6.1 tensorboard-plugin-wit-1.8.0 tensorflow-2.7.0 tensorflow-estimator-2.7.0 tensorflow-io-gcs-filesystem-0.21.0 termcolor-1.1.0 typing-extensions-3.10.0.2 werkzeug-2.0.2 wrapt-1.13.3
Note: you may need to restart the kernel to use updated packages.

In [2]:
```
import tensorflow as tf
```

In [3]:
```
import tensorflow.keras
```

In [ ]:

THE END