Case Study Vocabulary

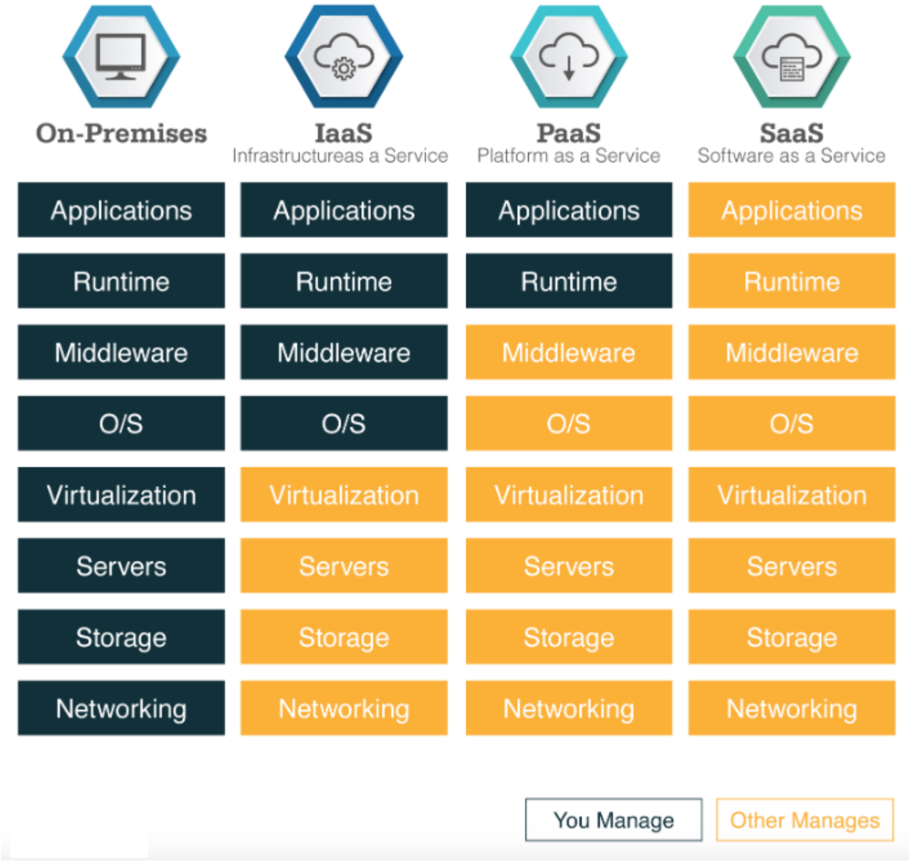
Behavioral data - is generated by the things people do online. The way that your users behave provides significant insight into your audience, website, and products and services. Organizations that leverage customer behavioral insights outperform peers by 85% in sales growth.

Cloud delivery models:

Infrastructure as a service (IaaS) - The consumer has the freedom to architect, deploy, and manage the application on the virtual resources provided by the vendor. They can buy and automate VMs and virtualized hardware resources on-demand instead of having to purchase hardware while the provider manages the actual servers, storage, and networking. Some examples of IaaS vendors include **Amazon Web Services (AWS), Microsoft Azure**, and **Google Cloud Platform**.

Platform as a service (PaaS) - Developers on the client side can create customized applications and deploy and manage them on the provider's infrastructure. The provider manages the underlying OS, middleware, and hardware virtualization. Some examples of PaaS include **Windows Azure**, **Google AppEngine**, **Heroku**, and **AWS Elastic Beanstalk**.

Software as a service (SaaS) - The cloud provider builds and manages applications for various business functions and delivers them to users via the internet. These apps are hosted on the provider's infrastructure. On the client side, they run on web browsers and don't require any downloads or installations. most well-known of all three models is SaaS. SaaS examples: ClickUp, Google Workspace apps, Microsoft 365, Netflix, Zoom. Shopify, HubSpot, Salesforce, DonorBox, Slack, Buffer, DocuSign, and many more.



On-Premises	IaaS Infrastructureas a Service	PaaS Platform as a Service	SaaS Software as a Service
Applications	Applications	Applications	Applications
Runtime	Runtime	Runtime	Runtime
Middleware	Middleware	Middleware	Middleware
O/S	O/S	O/S	O/S
Virtualization	Virtualization	Virtualization	Virtualization
Servers	Servers	Servers	Servers
Storage	Storage	Storage	Storage
Networking	Networking	Networking	Networking

Cloud deployment models - There are four cloud deployment models: **public**, **private**, **community**, and **hybrid**. Each deployment model is defined according to where the infrastructure for the environment is located. There are three main cloud service models: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service(IaaS).

Collaborative filtering - a technique that can filter out items that a user might like based on reactions by similar users. It works by searching a large group of people and finding a smaller set of users with tastes like a particular user.

Content-based filtering - works with data that the user provides, either explicitly (rating) or implicitly (clicking on a link). Based on that data, a user profile is generated, which is then used to make suggestions to the user.

Cost function - cost functions are used to estimate how badly models are performing. Put simply, a cost function is a measure of how wrong the model is in terms of its ability to estimate the relationship between X and y. This is typically expressed as a difference or distance between the predicted value and the actual value.

F-measure - The F-score (also known as the F1 score or F-measure) is a metric used to evaluate the performance of a Machine Learning model.

Hyperparameter - parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning. The prefix 'hyper_' suggests that they are 'top-level' parameters that control the learning process and the model parameters that result from it.

K-nearest neighbor (k-NN) algorithm - is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most like the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

Matrix factorization - is a class of collaborative filtering algorithms used in recommender systems. Matrix factorization algorithms work by decomposing the user-item interaction matrix into the product of two lower dimensionality rectangular matrices

Mean absolute error (MAE) - With any machine learning project, it is essential to measure the performance of the model. What we need is a metric to quantify the prediction error in a way that is easily understandable to an audience without a strong technical background. For regression problems, the Mean Absolute Error (MAE) is just such a metric.

The mean absolute error is the average difference between the observations (true values) and model output (predictions). The sign of these differences is ignored so that cancellations between positive and negative values do not occur. If we didn’t ignore the sign, the MAE calculated would likely be far lower than the true difference between model and data.

Overfitting - is a concept in data science, which occurs when a statistical model fits exactly against its training data. When this happens, the algorithm unfortunately cannot perform accurately against unseen data, defeating its purpose. Generalization of a model to new data is ultimately what allows us to use machine learning algorithms every day to make predictions and classify data.

When machine learning algorithms are constructed, they leverage a sample dataset to train the model. However, when the model trains for too long on sample data or when the model is too complex, it can start to learn the “noise,” or irrelevant information, within the dataset. When the model memorizes the noise and fits too closely to the training set, the model becomes “overfitted,” and it is unable to generalize well to new data. If a model cannot generalize well to new data, then it will not be able to perform the classification or prediction tasks that it was intended for.

Popularity bias - is a phenomenon that occurs when an algorithm produces results that are systemically prejudiced due to erroneous assumptions in the machine learning process.

Precision - can be seen as a measure of quality and **recall** as a measure of quantity. Higher precision means that an algorithm returns more relevant results than irrelevant ones, and high recall means that an algorithm returns most of the relevant results (whether irrelevant ones are also returned).

Recall – see above.

Reinforcement learning - is a feedback-based Machine learning technique in which an agent learns to behave in an environment by performing the actions and seeing the results of actions. For each good action, the agent gets positive feedback, and for each bad action, the agent gets negative feedback or penalty.

Right to anonymity – **Anonymity** has long been a core attribute of the internet, and encryption is the backbone technology that enables it. The ability to speak and act anonymously and share information online securely enables free speech and political activity, protects vulnerable and targeted groups, allows for personal development without the threat of persecution or shame and secures the digital economy**. This anonymity is seen by many as an inherent right in the digital age**. But is this right absolute? Recent revelations about the wide range of technologies being used to circumvent online privacy point to a shift in the debate.

Right to privacy - Artificial intelligence and big data present new challenges to protecting our right to privacy both online and offline. Predictive algorithms and online social networks add complex nuances to the idea of informed consent. A global, cross-disciplinary effort is underway to make predictive algorithms fair, accountable, and transparent as these algorithms come to mediate more and more the areas of civic life and public discourse. But these tools and ideas operate inside of existing socio-technical systems that need to balance conflicting goals and tensions.

Root-mean-square error (RMSE) - is one of the most commonly used measures for evaluating the quality of predictions. It shows how far predictions fall from measured true values using Euclidean distance.

Stochastic gradient descent - **Stochastic gradient descent** is an optimization algorithm often used in machine learning applications to find the model parameters that correspond to the best fit between predicted and actual outputs. It's an inexact but powerful technique.

Stochastic gradient descent is widely used in machine learning applications. Combined with backpropagation, it's dominant in neural network training applications.

Training Data - Machine learning uses algorithms to learn from data in datasets. They find patterns, develop understanding, make decisions, and evaluate those decisions.

In machine learning, datasets are split into two subsets.

The first subset is known as the **training data** - it's a portion of our actual dataset that is fed into the machine learning model to discover and learn patterns. In this way, it trains our model.

The other subset is known as the **testing data**. We'll cover more on this below.

Training data is typically larger than testing data. This is because we want to feed the model with as much data as possible to find and learn meaningful patterns. Once data from our datasets are fed to a machine learning algorithm, it learns patterns from the data and makes decisions.

Helpful YouTube Videos

https://youtu.be/z-EtmaFJieY - Machine Learning & Artificial Intelligence: Crash Course Computer Science: By Crash Course

https://youtu.be/a0_lo_GDcFw - What Is Artificial Intelligence? Crash Course AI: By Crash Course

https://youtu.be/4b5d3muPQmA - K-means clustering: By StatQuest

https://youtu.be/ZspR5PZemcs - How does Netflix recommend movies? Matrix Factorization

https://youtu.be/-4E2-0sxVUM - How computers represent images and the algorithms they use for object recognition: By Crash Course

https://youtu.be/IHZwWFHWa-w - About Gradient descent: By 3blue1brown

https://youtu.be/6PZEVNuBL0g - An in-depth look at the case study and principles within: By Mr Amith

https://youtu.be/0yCJMt9Mx9c - How does AI learn: By TED Ed

IB COMPUTER SCIENCE PAPER 3 EXAMINATION TECHNIQUES

https://youtu.be/Pw6nQC5zV8k

TERMINOLOGY - TAKEN FROM IB CASE STUDY: May I recommend the following?

Behavioural data BEHAVIORAL DATA FROM SNOWPLOW

Cloud delivery models: CLOUND DELIVERY MODELS FROM ALERTLOGIC

Infrastructure as a service (IaaS)

Platform as a service (PaaS)

Software as a service (SaaS)

Cloud deployment models DEPLOYMENT MODELS FROM GEEKS FOR GEEKS

Collaborative filtering COLLABRATIVE FILTERING FROM TOWARDS DATA SCIENCE

Content-based filtering

Cost function COST FUNCTION FROM ANALYTICSVIDHYA

F-measure F-MEASURE FROM DEEP AI

Hyperparameter HYPERPARAMETER FROM TOWARDS DATA SCIENCE

K-nearest neighbour (k-NN) algorithm K CLUSTERING FROM TOWARDS DATA SCIENCE

Matrix factorization MATRIX FACTORIZATION FROM TOWARDS DATA SCIENCE

Mean absolute error (MAE) MEANS ABSOLUTE ERROR FROM MEDIUM

Overfitting OVERFITTING FROM GEEKS FOR GEEKS

Popularity bias POPULAR BIAS FROM DLACM

Precision PRECISION FROM MEDIUM

Recall

Reinforcement learning REINFORCEMENT LEARNING FROM TECH TARGET

Right to anonymity

Right to privacy

Root-mean-square error (RMSE)

Stochastic gradient descent STOCHASTIC GRADIENT DESCENT FROM TOWARDS DATA SCIENCE

Training data TRAINING DATA FROM OBVIOUSLY