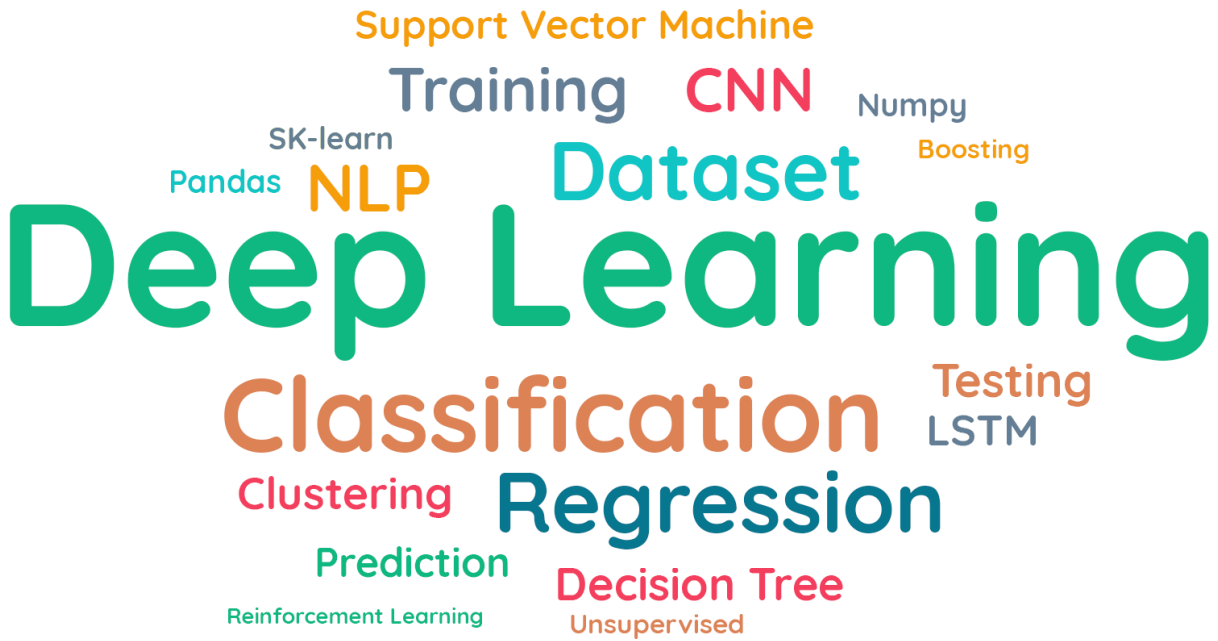
# Machine Learning Lifecycle

**Azhar Rizki Zulma**

# Overview




Apa itu Machine Learning Lifecycle?

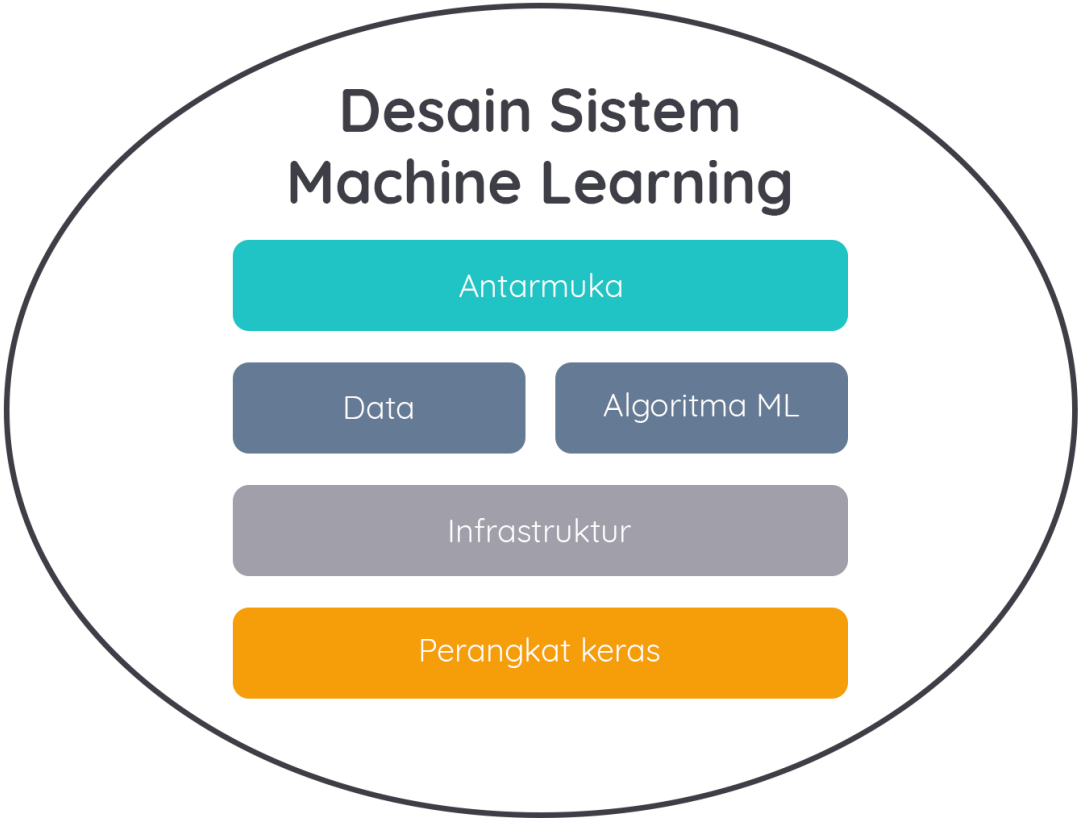
# Overview

Machine Learning Lifecycle merupakan sebuah proses siklus yang diikuti oleh proyek Data Science. Ini mendefinisikan setiap langkah yang harus diikuti organisasi untuk memanfaatkan pembelajaran dan kecerdasan buatan (AI) untuk mendapatkan nilai bisnis praktis.

# Machine Learning System Design

Sebelum memulai proyek machine learning, kita perlu melakukan proses desain sistem machine learning supaya dapat meninjau sebuah sistem secara keseluruhan. Desain sistem machine learning adalah proses menentukan antarmuka, algoritma, data, infrastruktur, dan perangkat keras untuk sistem machine learning guna memenuhi persyaratan (requirements) yang telah ditentukan sebelumnya.

# Machine Learning System Design



# Machine Learning System Design

Perbedaan Sistem Machine Learning di Dunia Riset/Akademik versus Industri berdasarkan tujuan dan data yang digunakan.

Dataset industri

Dataset akademik

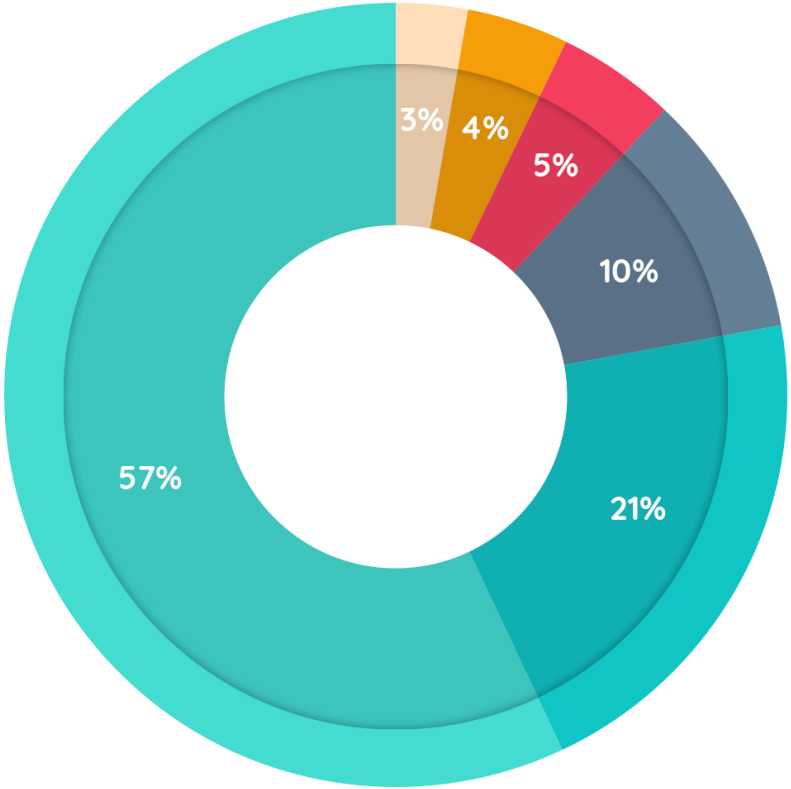
## Tujuan Sistem Machine Learning

Di dunia riset/akademik, tujuan utama sistem machine learning kebanyakan adalah untuk mendapatkan akurasi dan performa model yang tinggi, peningkatan kinerja, dan hasil tertinggi atau tahapan terbaru yang biasanya disebut sebagai state-of-the-art. Sedangkan di industri, setiap pemangku kepentingan memiliki tujuan yang berbeda sehingga diperlukan kolaborasi utuk membuat model yang bisa memenuhi semua tujuan tersebut.

## Data yang Digunakan

Sementara itu, dalam hal penggunaan data, dunia riset/akademik biasanya menggunakan data yang sudah bersih supaya dapat fokus pada pengembangan arsitektur dan proses melatih model. Saat mengembangkan sistem machine learning di industri, biasanya kita masih harus mencari dan mengumpulkan (data collecting) data mentah terlebih dahulu. Sehingga, di industri, diperlukan berbagai proses lebih lanjut seperti data cleansing dan data transformation. Selain itu, data industri biasanya merupakan data tidak terstruktur (unstructured data), tidak seimbang (imbalance data), dan berubah secara berkala.

# Machine Learning System Design




## What's the least enjoyable part of data science?

- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%

## Persyaratan Machine Learning System Design

Sistem machine learning harus mampu menangani perubahan, memudahkan dalam pemeliharaan, dan memudahkan untuk beradaptasi. Oleh karena itu, keterampilan dalam mendesain model machine learning yang dapat memenuhi berbagai persyaratan tersebut, mutlak diperlukan. Apa saja persyaratan yang dibutuhkan oleh sistem machine learning? Berikut ini adalah persyaratannya:

- ✕ Bersifat andal (reliable).
- ✕ Mampu menangani perubahan kapasitas (scalable).
- ✕ Mudah beradaptasi (adaptable).
- ✕ Mudah dalam pemeliharaan (maintainable).

## Bersifat Andal (Reliable)

Pengambilan keputusan dengan machine learning di berbagai bidang yang penuh risiko seperti kesehatan, bisnis, keuangan, serta tata kelola pemerintahan membutuhkan akuntabilitas (pertanggungjawaban) yang jelas. Terutama akuntabilitas atas keputusan yang diambil, juga bagaimana data digunakan dalam proses pengambilan keputusan tersebut. berikut adalah 2 hal penting untuk memastikan sistem machine learning bersifat andal.

# Machine Learning System Design

**Antisipasi Kegagalan:**

- Perhatikan kualitas data, kualitas data yang buruk dan tidak memadai merupakan salah satu kegagalan dalam desain sistem Machine Learning.
- Perhatikan pemilihan model, lakukan pemeriksaan asumsi model untuk memilih teknik model yang tepat agar meminimalisir kegagalan.
- Perhatikan kerapuhan dalam sebuah model *(model fragility*), Istilah “rapuh” di sini maksudnya adalah prediksi model sangat sensitif terhadap gangguan pada masukan.

# Machine Learning System Design

**Contoh Kegagalan karena kerapuhan model (*Model Fragility*):**

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

# Machine Learning System Design

**Pemeliharaan:**

Proses pemeliharaan penting untuk memastikan keandalan sistem machine learning. Dalam proses ini kita perlu mendeteksi kapan pembaruan sistem harus dilakukan dan bagaimana cara melakukannya dengan aman. Sistem machine learning adalah proses iteratif. Siklus pengembangan machine learning memerlukan umpan balik (feedback) selama proses pembelajaran yang berkelanjutan. Feedback loops adalah proses di mana perubahan atau output dari salah satu bagian sistem dikirimkan kembali ke dalam sistem sebagai input sehingga mempengaruhi tindakan atau output sistem selanjutnya. Umpan balik ini berguna untuk meningkatkan akurasi prediksi dan merupakan bagian dari pemeliharaan sistem machine learning.

## Mampu menangani perubahan kapasitas (Scalable)

Seiring dengan pertumbuhan sistem (meningkatnya volume, lalu lintas, dan kompleksitas data), kita perlu memikirkan teknik untuk menangani pertumbuhan tersebut. Dalam penerapan machine learning di dunia nyata, scalability atau kemampuan menangani perubahan kapasitas sering menjadi perhatian utama. Proses penanganan terhadap perubahan kapasitas tidak hanya berarti peningkatan sumber daya untuk menangani pertumbuhan data (scaling up), tetapi juga penanganan untuk mengurangi sumber daya yang tidak kita butuhkan (scaling down).

## Mudah beradaptasi (Adaptable)

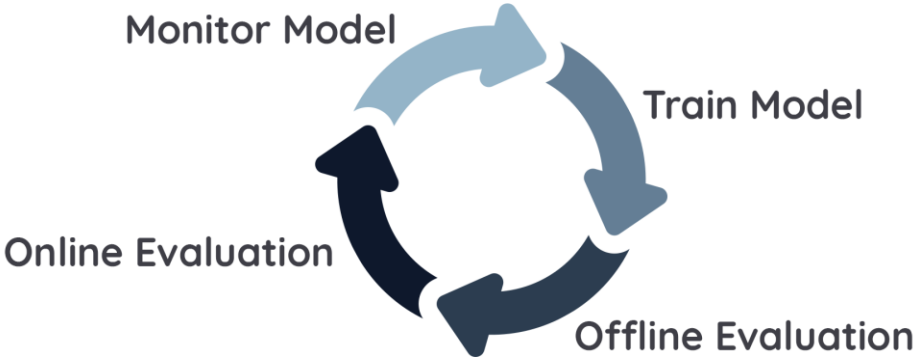
Kualitas yang membedakan sistem machine learning dengan sistem perangkat lunak tradisional adalah kemampuan untuk belajar dan beradaptasi saat sistem mengumpulkan informasi dan membuat keputusan. Dalam siklus machine learning, perubahan tidak bisa dihindari. Sistem machine learning yang efektif harus mampu beradaptasi saat terjadi perubahan. Berikut adalah beberapa cara untuk membuat sistem bersifat adaptif terhadap perubahan distribusi data:

- ✕ Memantau statistik deskriptif untuk masukan dan keluaran.
- ✕ Menggunakan arsitektur pelatihan yang dinamis.
- ✕ Melatih model Anda secara berkala

## Mudah dalam pemeliharaan (Maintanable)

Proses pengawasan dan pemeliharaan adalah sekumpulan teknik dan tata cara untuk mengelola sistem machine learning pada tahap produksi. Proses ini bertujuan untuk mendeteksi dan memperbaiki kegagalan sistem atau meningkatkan kinerja sistem secara keseluruhan. Secara umum, terdapat 4 tahapan siklus pemeliharaan, yaitu:

- Melatih model
- Evaluasi secara offline
- Evaluasi secara online
- Pengawasan model

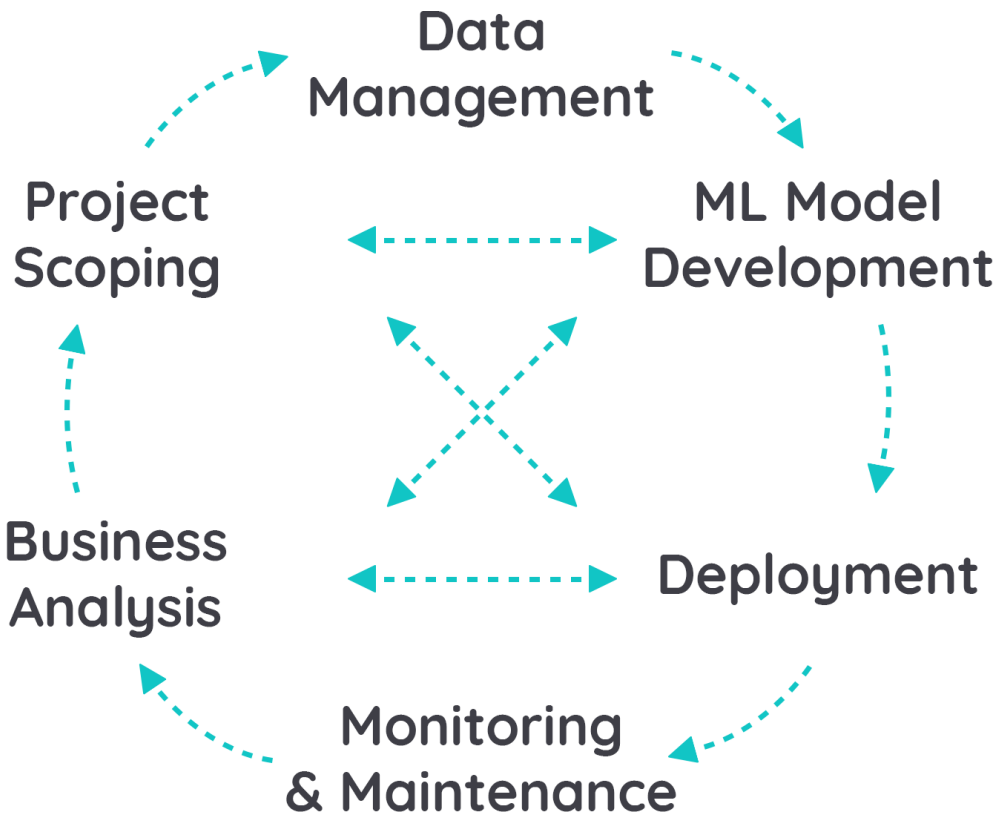


# Alur Proyek Machine Learning

Proyek machine learning di industri merupakan sebuah siklus. Secara umum, terdapat 6 tahap utama dalam siklus ini, yaitu:

- ✕ Cakupan proyek
- ✕ Manajemen data
- ✕ Pengembangan model machine learning
- ✕ Deployment
- ✕ Pengawasan dan pemeliharaan
- ✕ Analisis Bisnis

# Machine Learning System Design



# Infrastuktur Machine Learning

Kompleksitas penerapan sistem machine learning di industri berimbas pada kebutuhan infrastruktur yang masif. Kebutuhan infrastruktur dalam sistem machine learning dibagi ke dalam tiga bagian, yaitu Data, Training/Evaluasi, dan Deployment.

# Infrastuktur Machine Learning

Kompleksitas penerapan sistem machine learning di industri berimbas pada kebutuhan infrastruktur yang masif. Kebutuhan infrastruktur dalam sistem machine learning dibagi ke dalam tiga bagian, yaitu Data, Training/Evaluasi, dan Deployment.

# Data Engineering

Jika Anda masih ingat gambar tentang desain sistem machine learning dari slide sebelumnya pada Pengenalan Machine Learning System Design, data adalah bagian tersendiri, berdampingan dengan algoritma ML. Tidak bisa dipungkiri, perkembangan pesat AI dan machine learning dalam satu dekade terakhir sejalan dengan perkembangan era big data. Ketersediaan data membuat mesin “belajar” dengan lebih baik. Untuk membangun sebuah sistem yang cerdas, dibutuhkan keterampilan mengenai data engineering. Ia adalah serangkaian operasi perancangan sistem untuk mengumpulkan, menyimpan, dan menganalisis data. Bidang ini sangat luas aplikasinya di hampir setiap industri.

# Machine Learning System Design

THE DATA SCIENCE
**HIERARCHY OF NEEDS**

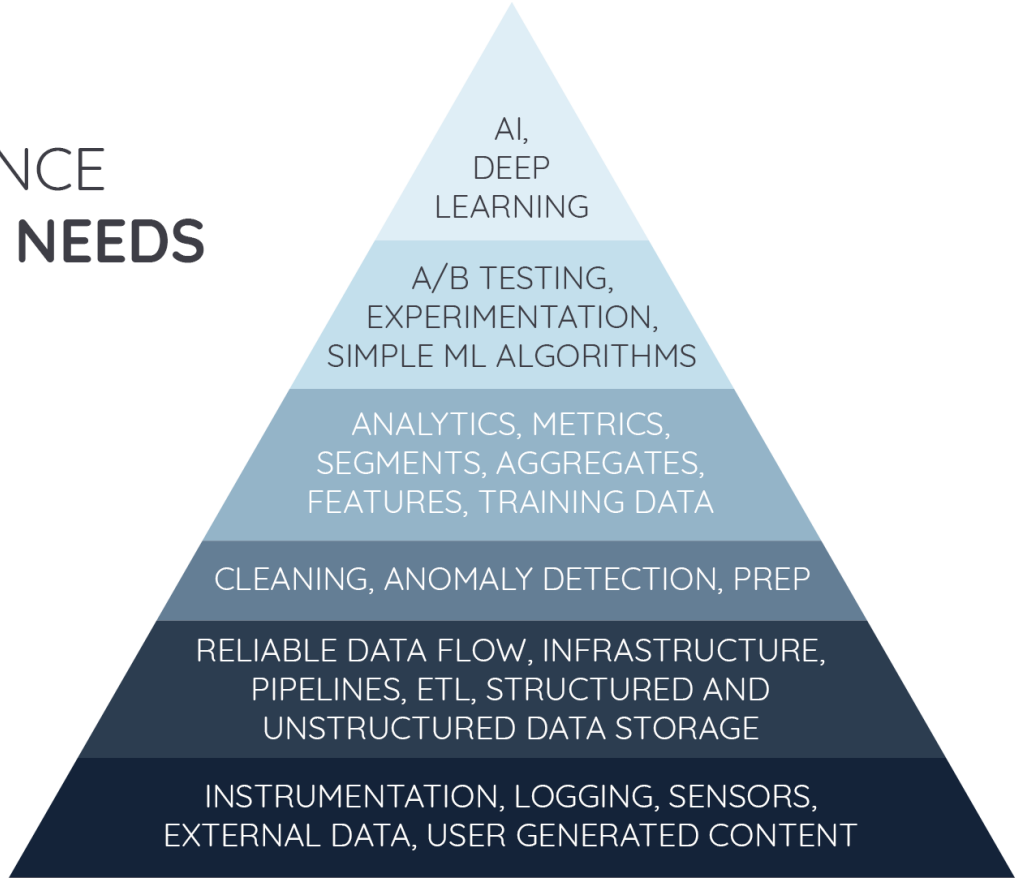
LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT



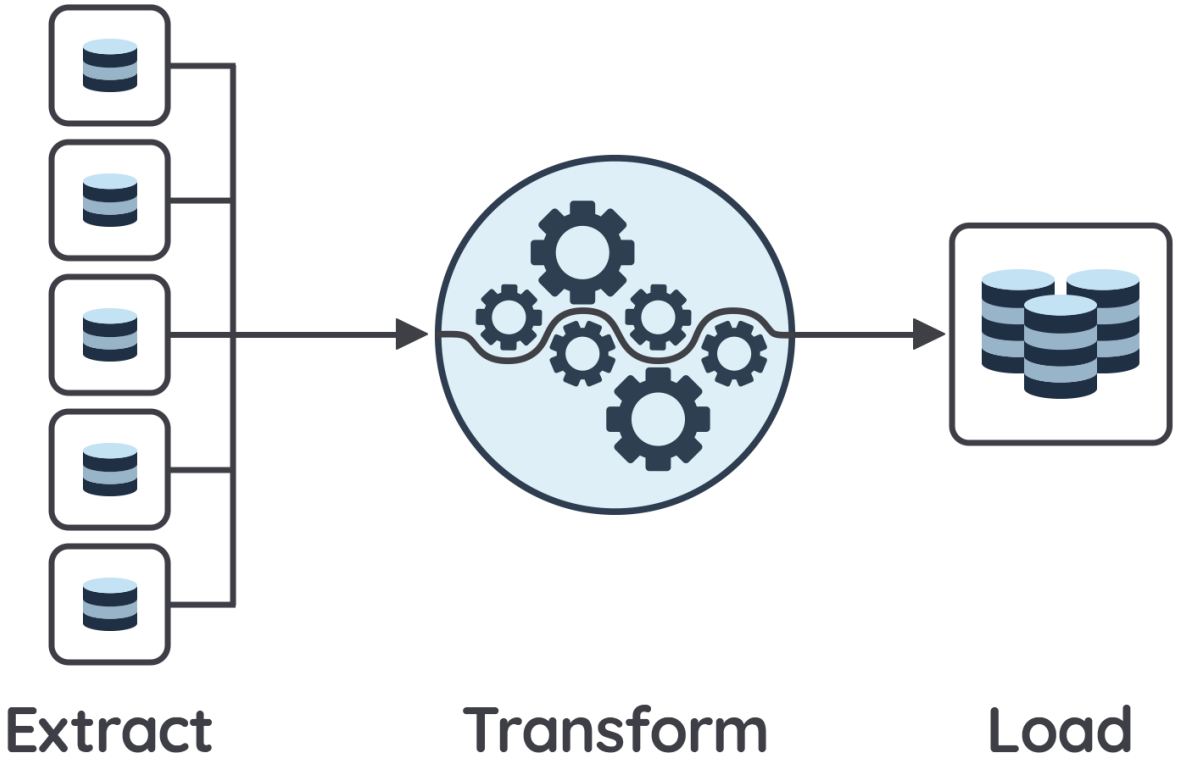
## Sumber Data

Sumber data untuk sistem machine learning berasal dari berbagai sumber. Sumber data yang pertama adalah data yang dibuat oleh pengguna (user generated data). Data dari sumber ini dihasilkan saat pengguna melakukan tindakan aktif (klik) pada website seperti memesan produk atau transaksi elektronik, mengabaikan saran, atau saat pengguna melakukan scrolling pada halaman website. Sumber data lain adalah data yang dihasilkan oleh sistem (system-generated data) atau kadang juga disebut machine-generated data. Contohnya adalah logs, metadata, dan prediksi yang dibuat oleh model. Sumber lainnya adalah data aplikasi perusahaan. Sebuah perusahaan mungkin menggunakan berbagai aplikasi untuk mengelola aset mereka seperti inventaris, hubungan pelanggan, dan pengguna.

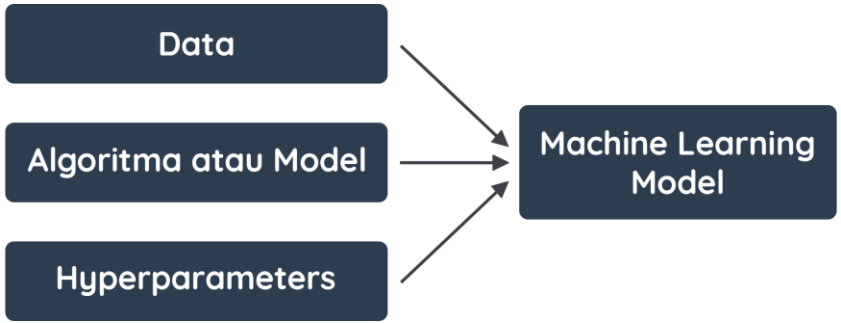
## Extract, Transform, and Loading (ETL)

Membangun data pipeline (jalur data) adalah tanggung jawab utama dalam data engineering. Data pipeline adalah serangkaian tools dan proses untuk melakukan integrasi data dari satu sistem ke sistem yang lain. Extract adalah proses ekstraksi data dari sumber-sumber database. Transform adalah sebuah proses penting untuk mengubah atau mentransformasi data agar menjadi format yang sesuai dengan kebutuhan bisnis. Sedangkan Loading adalah proses memuat atau menyimpan data ke tempat tujuan yang biasanya adalah data warehouse, sistem manajemen basis data relasional (RDBMS), atau Hadoop.

# Machine Learning System Design



# Model Development



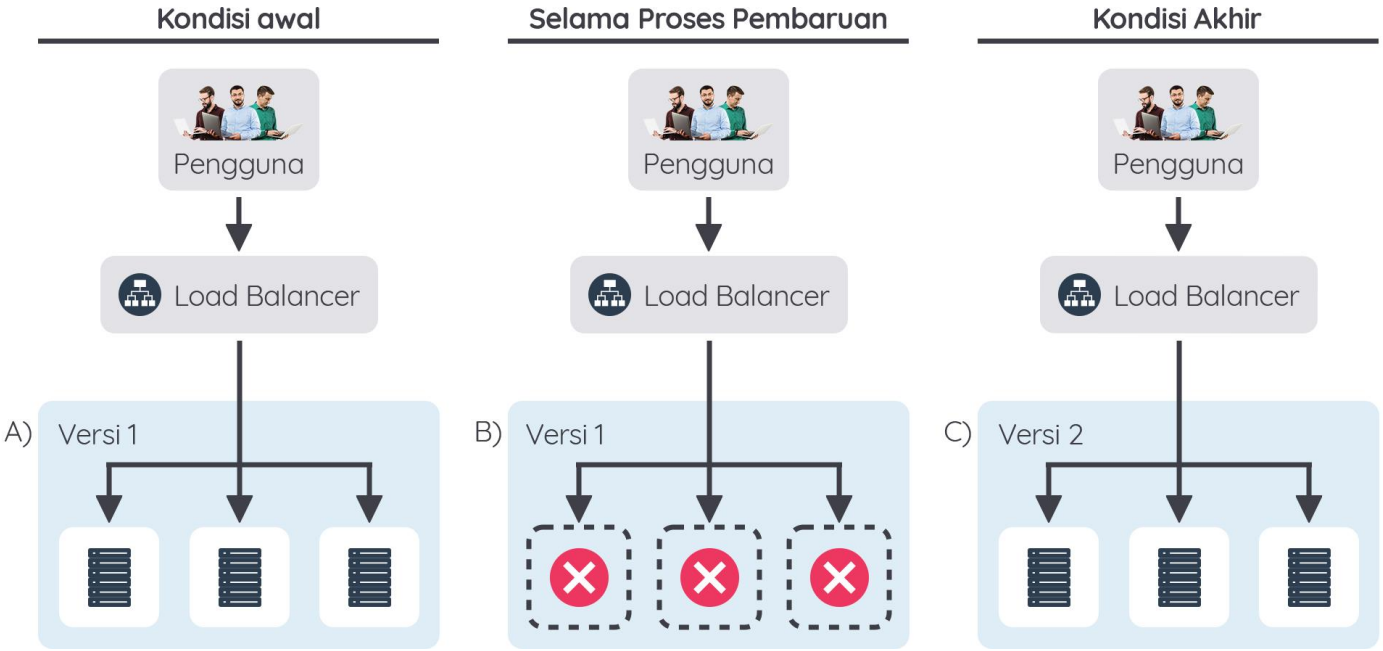

Pengembangan model adalah proses yang iteratif. Anda biasanya memulai dengan proses seleksi model (memilih satu di antara banyak model kandidat). Kemudian, Anda akan mulai mengerucutkan kategori masalah menjadi lebih spesifik, misalnya apakah ini merupakan permasalahan klasifikasi atau regresi? Tahap selanjutnya adalah membuat baseline (model dasar). Terakhir, Anda perlu melakukan pengaturan parameter untuk mendapatkan performa terbaik.

# Deployment and Monitoring

Terdapat dua kategori pada machine learning model deployment, yaitu batch scoring dan real-time scoring. Batch scoring adalah kategori deployment saat model memproses seluruh dataset yang telah dikumpulkan dari waktu ke waktu untuk menghasilkan prediksi baru. Sedangkan, pada real-time scoring data diproses segera saat diterima atau secara real-time. Sehingga, hasilnya dapat langsung ditampilkan di aplikasi pada saat itu juga.

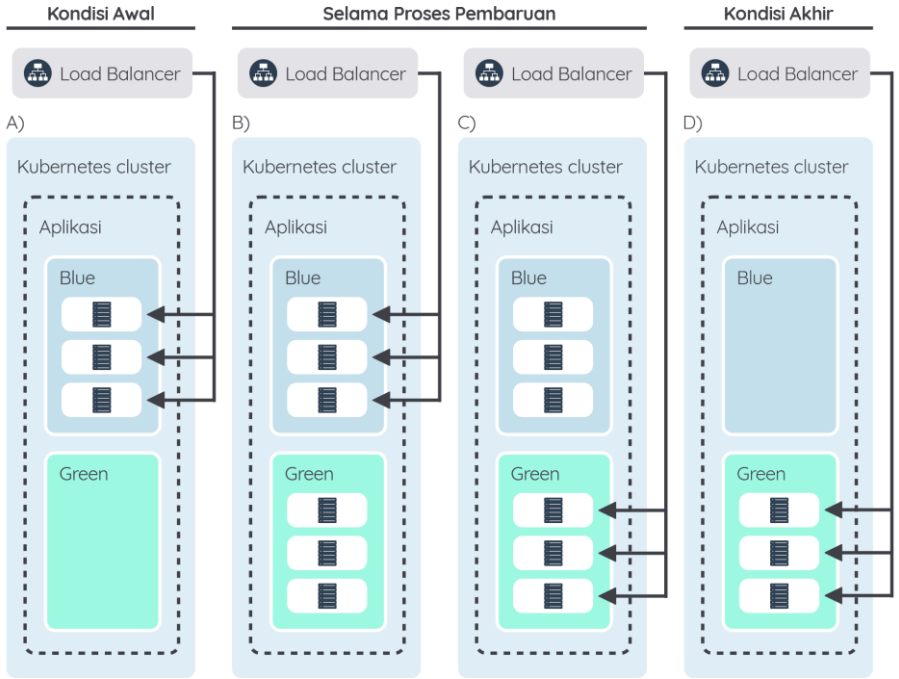
# Machine Learning System Design

**Recreate deployment**: Ia merupakan strategi penerapan ulang yang bekerja dengan sepenuhnya menurunkan versi aplikasi lama sebelum Anda meningkatkan versi aplikasi baru.



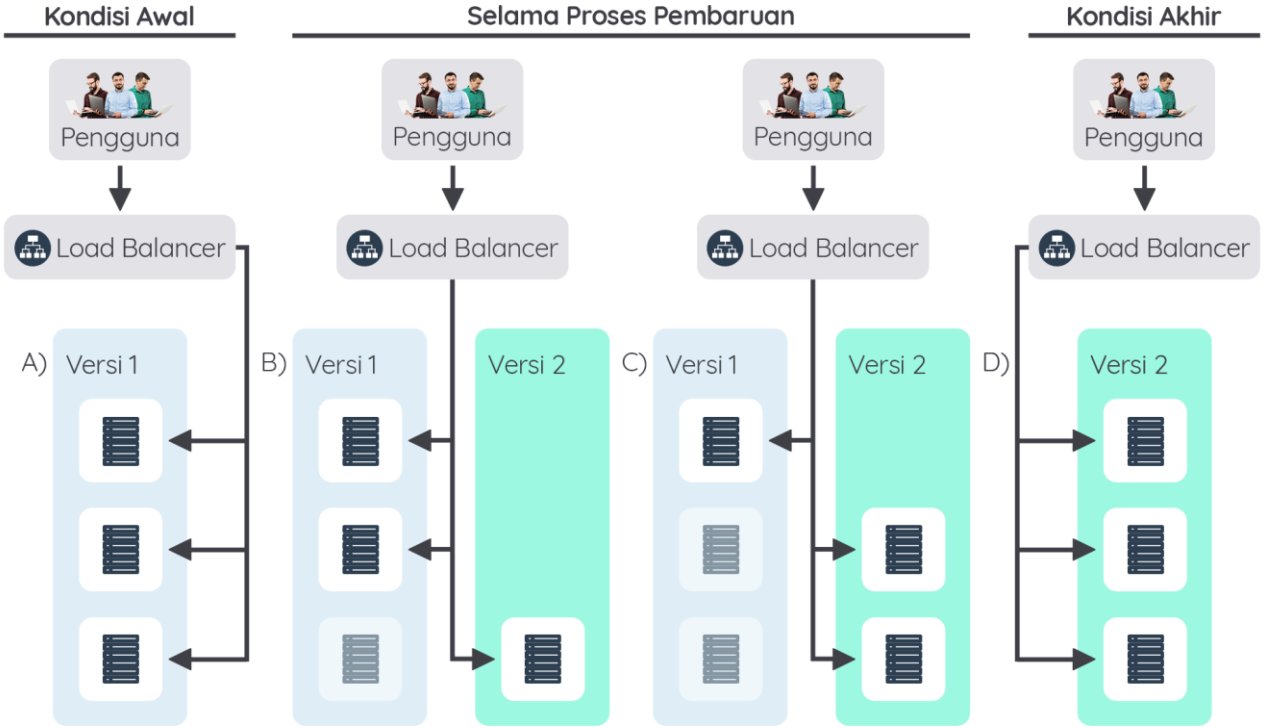
# Machine Learning System Design

**Blue-green deployment**: Ini merupakan strategi saat Anda ingin menerapkan versi baru aplikasi sekaligus memastikan bahwa layanan aplikasi tetap tersedia saat penerapan diperbarui.



# Machine Learning System Design

**Rolling Update Deployment**: Strategi ini mengalihkan lalu lintas secara bertahap ke versi baru.



# Machine Learning System Design

Setelah proses deployment, kita harus memastikan model berjalan sebagaimana mestinya. Proses ini disebut monitoring. Apa saja yang metrik yang perlu dimonitor, berikut ini adalah metrik yang perlu dimonitor:

- Performa model: Matriks ini membantu kita mendeteksi terjadinya penyimpangan dalam model
- Model input: Matriks ini digunakan untuk mengukur perubahan distribusi data masukan.
- Performa sistem: Matriks ini membantu kita menentukan bagaimana kinerja model yang ditetapkan dari sudut pandang penggunaan atau layanan.

# Menyusun Proyek Machine Learning

Portofolio machine learning merupakan aset penting untuk menunjukkan pemahaman dan keterampilan seseorang di bidang machine learning. Portofolio machine learning tidak harus terdiri dari proyek-proyek yang besar atau rumit. Proyek sederhana yang dapat menunjukkan kompetensi teknis Anda di bidang machine learning dapat dicantumkan sebagai portofolio.

# Menyusun Proyek Machine Learning

Berikut adalah alur dari pembuatan Portfolio Machine Learning:

# Menyusun Proyek Machine Learning

Berikut adalah beberapa ide portofolio machine learning:

- ✕ Proses data cleaning
- ✕ Eksplorasi dan visualisasi data
- ✕ Implementasi algoritma machine learning
- ✕ End-to-end machine learning project
- ✕ Kontribusi ke proyek open sorce

# Machine Learning Lifecycle

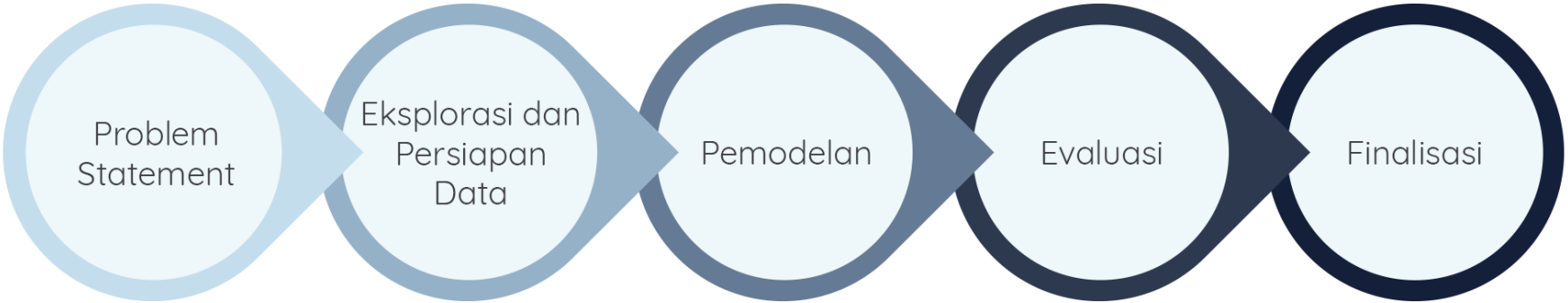
Setiap proyek machine learning itu bersifat unik. Tidak ada ketentuan khusus mengenai struktur atau kerangka proyek yang harus diikuti. Tetapi secara umum, ada template yang dapat kita ikuti saat menyusun proyek machine learning. Berikut adalah template yang umumnya digunakan dalam menyusun proyek machine learning.

# Machine Learning Lifecycle

## The Machine Learning Life Cycle


1. **Define Project Objectives**
   - ☐ Specify business problem
   - ☐ Acquire subject matter expertise
   - ☐ Define unit of analysis and prediction target
   - ☐ Prioritize modeling criteria
   - ☐ Consider risks and success criteria
   - ☐ Decide whether to continue
2. **Acquire & Explore Data**
   - ☐ Find appropriate data
   - ☐ Merge data into single table
   - ☐ Conduct exploratory data analysis
   - ☐ Find and remove any target leakage
   - ☐ Feature engineering
3. **Model Data**
   - ☐ Variable selection
   - ☐ Build candidate models
   - ☐ Model validation and selection
4. **Interpret & Communicate**
   - ☐ Interpret model
   - ☐ Communicate model insights
5. **Implement, Document & Maintain**
   - ☐ Set up batch or API prediction system
   - ☐ Document modeling process for reproducibility
   - ☐ Create model monitoring and maintenance plan

# Machine Learning Project Structure



## Problem Statements

Langkah pertama dalam proyek apa pun adalah mendefinisikan masalah yang dihadapi. Anda harus memahami situasi dan masalah yang perlu dipecahkan terlebih dahulu, kemudian memikirkan bagaimana machine learning akan menyelesaikan permasalahan tersebut secara efisien. Setelah mengetahui masalahnya dengan baik, Anda kemudian melanjutkan untuk menyelesaikannya. Pada tahapan ini, masalah didefinisikan dengan baik dan memiliki setidaknya satu solusi potensial yang relevan. Selain itu, masalah harus dapat diukur dan dapat direplikasi.

# Eksplorasi dan Persiapan Data

Berikut ini adalah tahapan dalam melakukan Eksplorasi dan Persiapan Data:

- ✕ Pengumpulan Data
- ✕ Eksplorasi Data
- ✕ Visualisasi Data
- ✕ Persiapan Data/Pre-Processing Data
- ✕ Seleksi Fitur Data

## Pemodelan

Pada tahap ini Anda dapat mulai menentukan model atau algoritma yang akan digunakan untuk menyelesaikan permasalahan. Ini adalah proses yang berulang. Pemodelan melibatkan pemilihan, pembuatan dan pelatihan model, kemudian menilai kinerjanya pada data uji Anda untuk memperkirakan kapasitas generalisasinya. Setelah mencoba strategi dengan beberapa model dan konfigurasi yang berbeda, Anda kemudian dapat memilih model terakhir dan melanjutkan ke proses selanjutnya, yaitu menetapkan model baseline. Anda dapat mulai dengan model sederhana dan secara bertahap meningkatkan level kompleksitasnya.

# Evaluasi Model

Setelah memiliki gambaran umum tentang arsitektur model yang berhasil dan pendekatan untuk permasalahan Anda, kini Anda dapat mulai fokus pada peningkatan kinerja model. Dua hal dapat kita lakukan pada model saat meningkatkan performa, antara lain:

- ✕ Meningkatkan akurasi model
- ✕ Mengurangi overfitting

# Finalisasi & Simpulan

Sampai di tahap ini, Anda telah mendapatkan model dan akurasi yang diinginkan. Langkah selanjutnya adalah melakukan finalisasi. Ada beberapa hal yang dapat Anda lakukan untuk finalisasi tergantung kebutuhan dan tujuan, misalnya, melakukan prediksi atau melanjutkan ke proses deployment. Pada bagian simpulan, Anda merangkum ide, poin penting, informasi, dan hasil analisis dari seluruh tahapan yang telah lakukan. Bagian ini menunjukkan sejauh mana tujuan telah tercapai serta bagaimana proyek Anda dapat menjawab permasalahan yang disampaikan di tahapan Problem Statement. Anda juga dapat menyampaikan keterbatasan dan tantangan yang dialami dalam proyek serta membuat rekomendasi untuk proyek selanjutnya di masa mendatang.

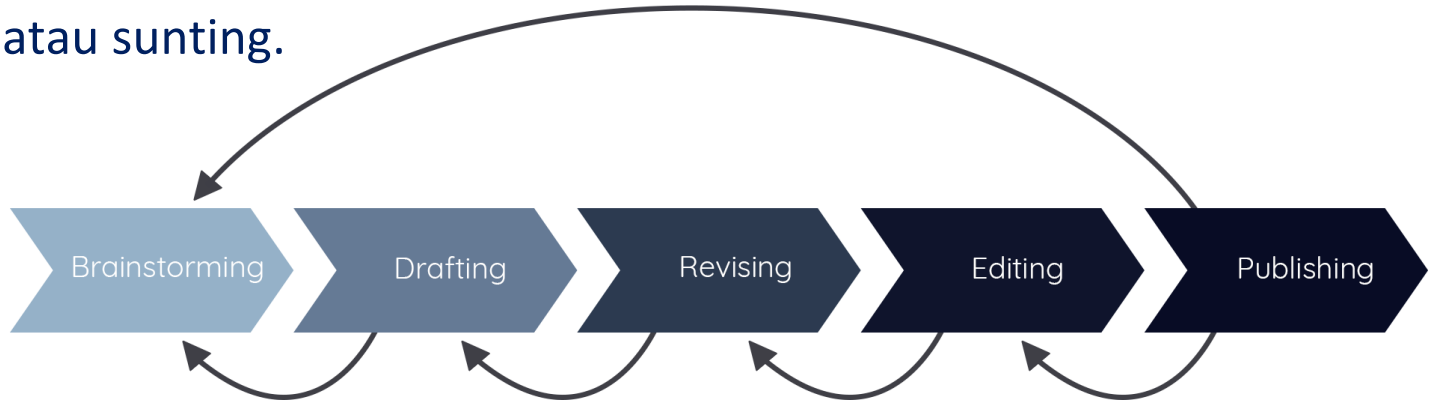
# Technical Writing

Technical writing merupakan bentuk tulisan yang ditargetkan untuk pembaca yang mencari informasi tentang topik teknis tertentu, misalnya topik tentang komputer, rekayasa perangkat lunak, keuangan, kesehatan, dan lain-lain. Technical Writing digunakan untuk menuliskan dokumentasi teknis mengenai hal-hal yang telah dilakukan dalam proyek Machine Learning yang anda buat.

# Machine Learning Lifecycle

Secara umum ada lima tahap proses penulisan dalam menulis dokumen Technical Writing, yaitu:

- Brainstorming atau proses mendapatkan ide.
- Pembuatan draf.
- Proses revisi.
- Proses edit atau sunting.
- Publikasi.



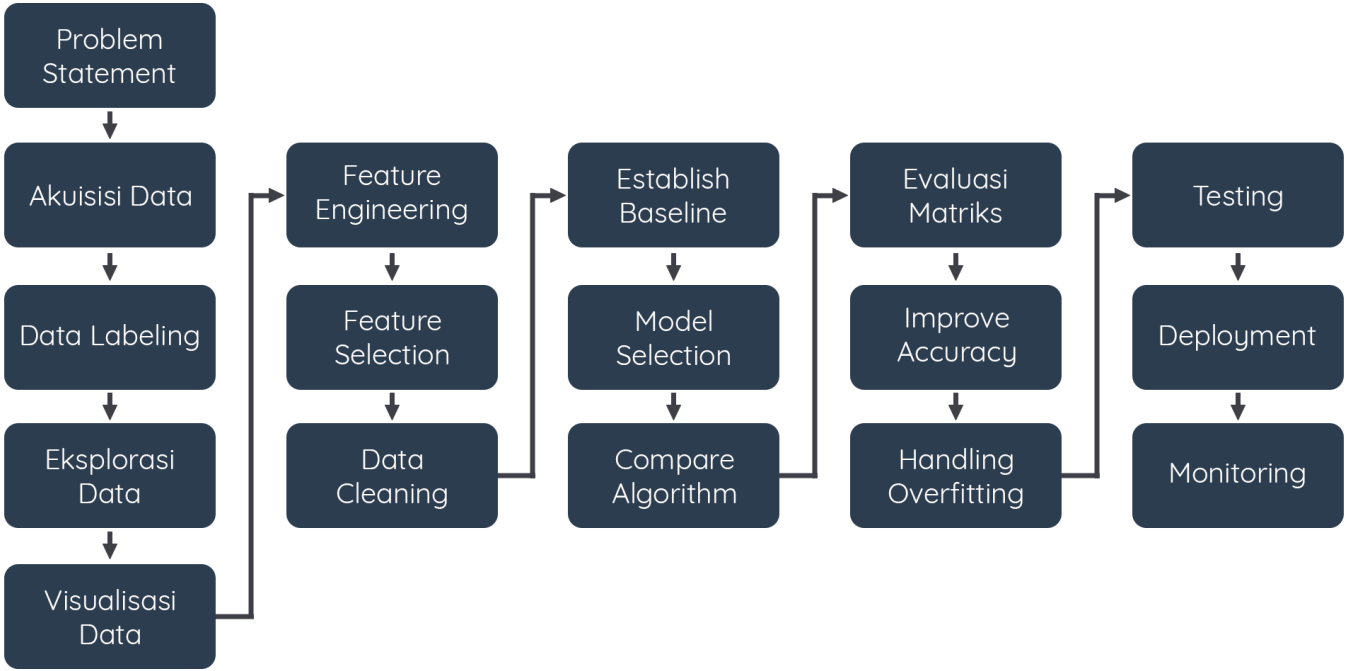
# Machine Learning Lifecycle

Perhatikan hal berikut dalam menulis dokumen Technical Writing:

- Tata Bahasa: Apa pun bahasa yang Anda gunakan, mengikuti kaidah tata bahasa merupakan hal yang penting dalam proses penulisan.
- Ilustrasi: Salah satu cara untuk membantu pembaca memahami konteks adalah dengan membuat ilustrasi atau grafik. Anda tentu sepakat, melihat ilustrasi jauh lebih menyenangkan daripada membaca teks. Bahkan dalam hal membaca materi teknis, sebagian besar dari kita lebih menyukai penyampaian informasi melalui grafik dibanding teks.

# Machine Learning Lifecycle

Ilustrasi teknis yang kompleks seperti berikut akan membuat pembaca merasa bingung. Buatlah agar ilustrasi tidak terlalu panjang atau dipersingkat.



# Machine Learning Lifecycle

Berikut ini contoh ilustrasi yang lebih mudah dipahami. Subsystem

# Publikasi Proyek Machine Learning

Publikasi Proyek dapat dilakukan di beberapa platform sesuai dengan tujuan dibuatnya proyek anda. Diantaranya ada beberapa platform yang dapat Anda gunakan untuk menampilkan konten proyek Anda, yaitu blog pribadi dan media sosial (misalnya GitHub dan Linkedin).

# Terima Kasih

**Sesi Diskusi**