LEARNING PROGRESS REVIEW
WEEK 1

Data Science

Batch 11

By

OMICRON

Data Science – Batch 11

OMICRON

Anggota Kelompok 3:

Anugrah Yazid Ghani

Fajar Achmad

Muhammad Fikri Fadila

Edo Mohammad Hadad Gibran

DigitalSkola

TOPIK

INTRODUCTION TO DATA SCIENCE

DATA SCIENCE METHODOLOGY

INTRODUCTION TO DATA SCIENCE

WEEK 1 – SESSION 2

DS OMICRON – BATCH 11

DigitalSkola

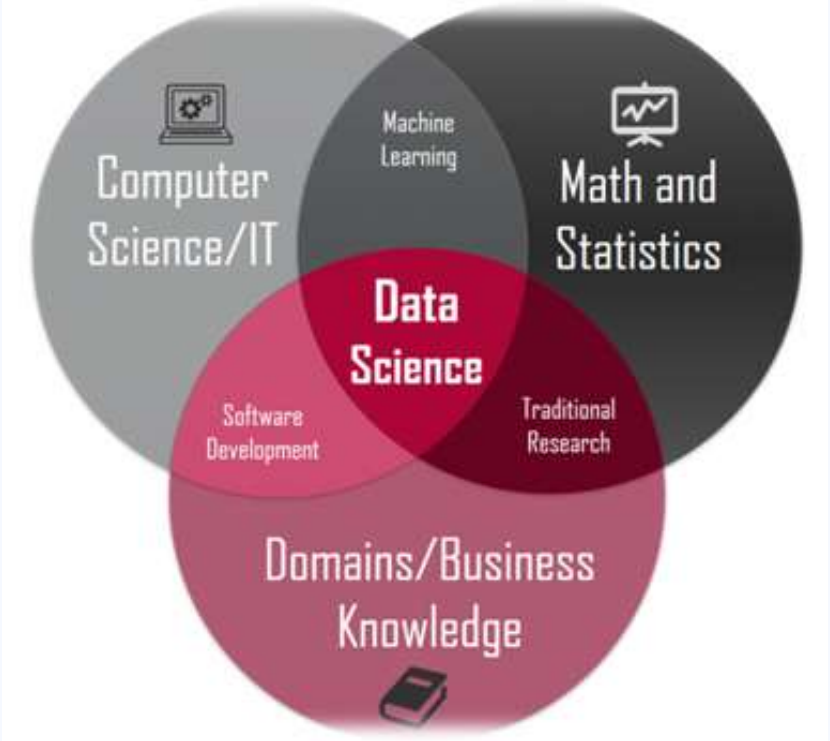
OUTLINE

1. Apa itu Data Science?
2. Kemampuan yang diperlukan untuk menjadi Data Scientist
3. Tujuan dari Data Science
4. Manfaat dan Keunggulan Data Science
5. Library
6. Data Science's Subset
   - Artificial Intelligence (AI)
   - Machine Learning (ML)
   - Deep Learning (DL)
7. Data Science in Business
8. Exploratory Data Analysis (EDA)

DigitalSkola

APA ITU **DATA SCIENCE?**

- **Data**: Fakta dan statistik terkumpul yang dapat dijadikan untuk referensi atau analisis
- **Science**: Aktivitas intelektual dan praktis yang mencakup studi sistematis tentang struktur dan perilaku dunia fisik dan alam melalui pengamatan dan eksperimen.

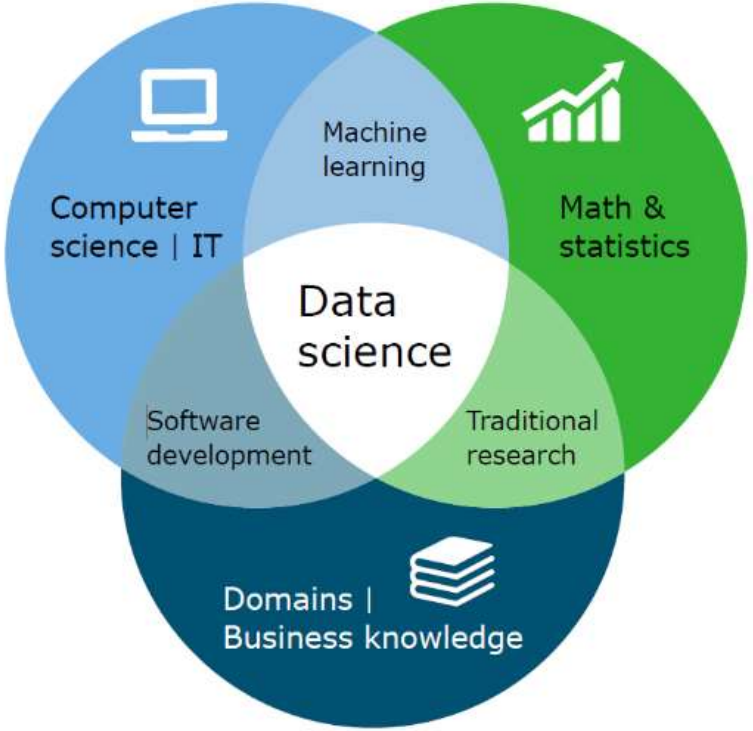


sumber: https://medium.com/@jrendz/pengenalan-data-science-b49a52eeef9c

JADI, DATA SCIENCE ADALAH ILMU YANG MENGGABUNGKAN **ILMU KOMPUTER**, **PEMROGRAMAN**, **MATEMATIKA**, **STATISTIKA**, DAN **ILMU BISNIS**.

KEMAMPUAN YANG DIPERLUKAN UNTUK MENJADI **DATA SCIENTIST**

- Kemampuan tentang **Business Logic** di bidang yang digeluti oleh perusahaan yang sedang ditangani.
- Kemampuan **Statistika** dan **Matematika** yang cukup untuk mengetahui pola-pola data beserta algoritmanya.
- Kemampuan **IT / Computure Science** untuk menggunakan tools dari sistem Big Data dan Data Science yang membantunya dalam mengolah dan menganalisis data.



sumber: https://www.wur.nl/en/Value-Creation-Cooperation/Collaborating-with-WUR-1/WDCC/Data-Education/Data-Scientist.htm

TUJUAN DARI DATA SCIENCE

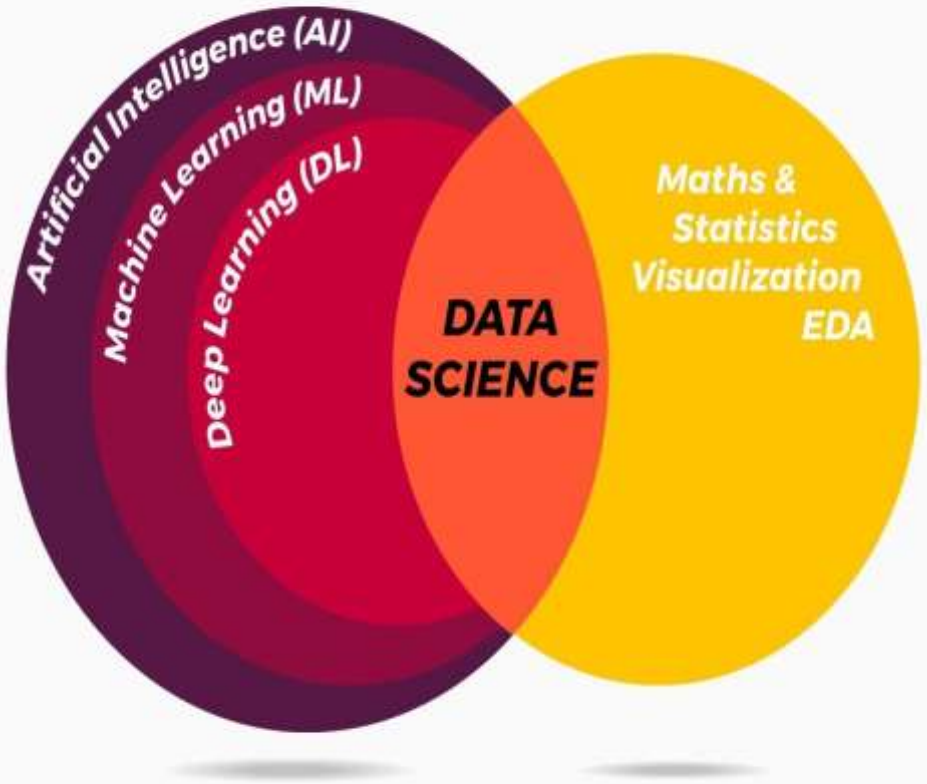
- Mengolah dan mengekstrak pengetahuan atau informasi dari data untuk mendapatkan data yang benar.
- Membantu sebuah perusahaan untuk membuat keputusan yang lebih cepat dan lebih baik.
- Memprediksi kecurangan dan menggunakannya demi membuat peringatan yang dapat membantu memastikan tindakan cepat ketika data tersebut tidak bisa diakui.

MANFAAT DAN KEUNGGULAN DATA SCIENCE

- Hasil analisis lebih efisien dan lebih akurat.
- Membantu tim sales dan marketing memahami customer mereka secara rinci untuk tercapainya kepuasan customer.
- Mengurangi resiko penipuan.
- Mengembangkan dan menghasilkan produk relevan.

LIBRARY

- Pustaka atau Library adalah sekumpulan kode / program yang memiliki fungsi-fungsi tertentu dan dapat dipanggil kedalam program lain.
- Library dibuat untuk mempermudah dalam membangun sebuah aplikasi. Dengam library programer tidak harus membangun kode dari awal untuk suatu fungsi tertentu.
- Pustaka yang disediakan bergantung pada bahasa pemrograman yang digunakan, untuk mengetahui pustaka apa saja yang ada perlu dilakukan eksplorasi terhadap bahasa pemrograman yang digunakan.



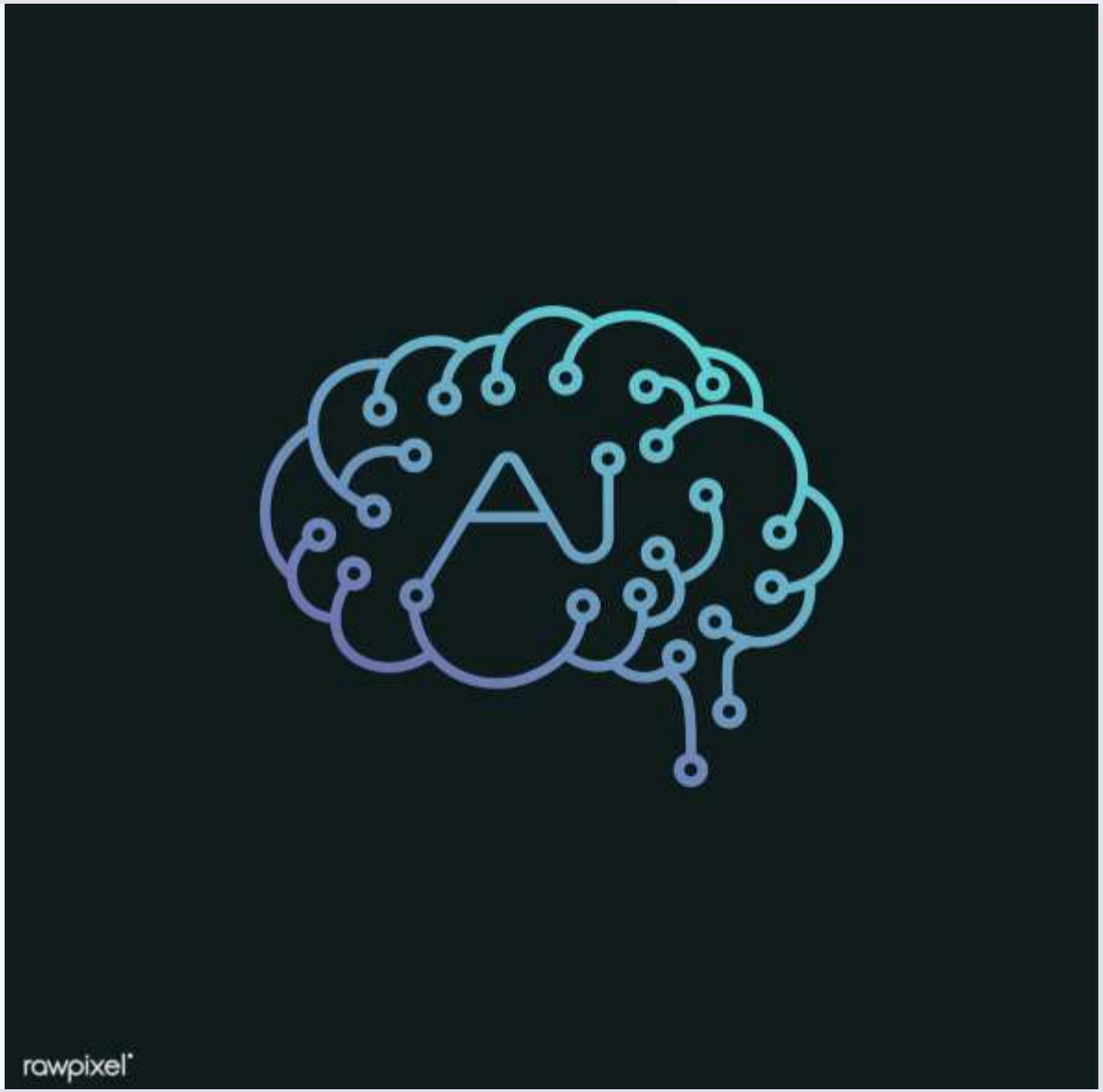
sumber: : https://medium.com/@pritishaw1220/data-science-machine-learning-and-deep-learning-are-they-cut-from-the-same-cloth-ae89abbc64af

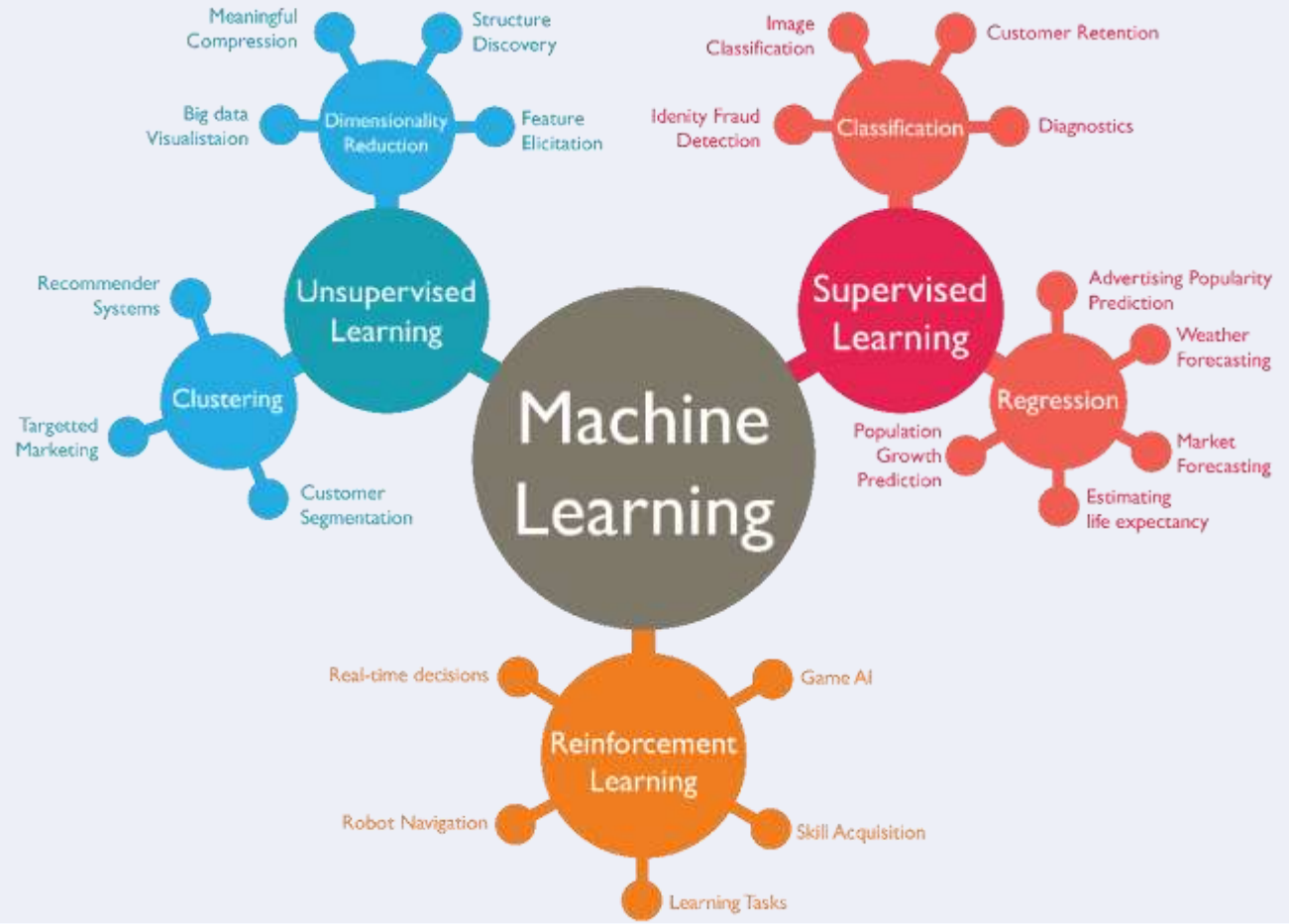
DATA SCIENCE'S SUBSET

1. Artificial Intelligence (AI)
2. Machine Learning (ML)
3. Deep Learning (DL)

ARTIFICIAL INTELLIGENCE (AI)

- Kemampuan mesin untuk meniru kecerdasan atau perilaku manusia.
- Sebuah sistem yang dikembangkan guna dapat menyelesaikan suatu tugas yang biasanya hanya diselesaikan oleh kecerdasan manusia.

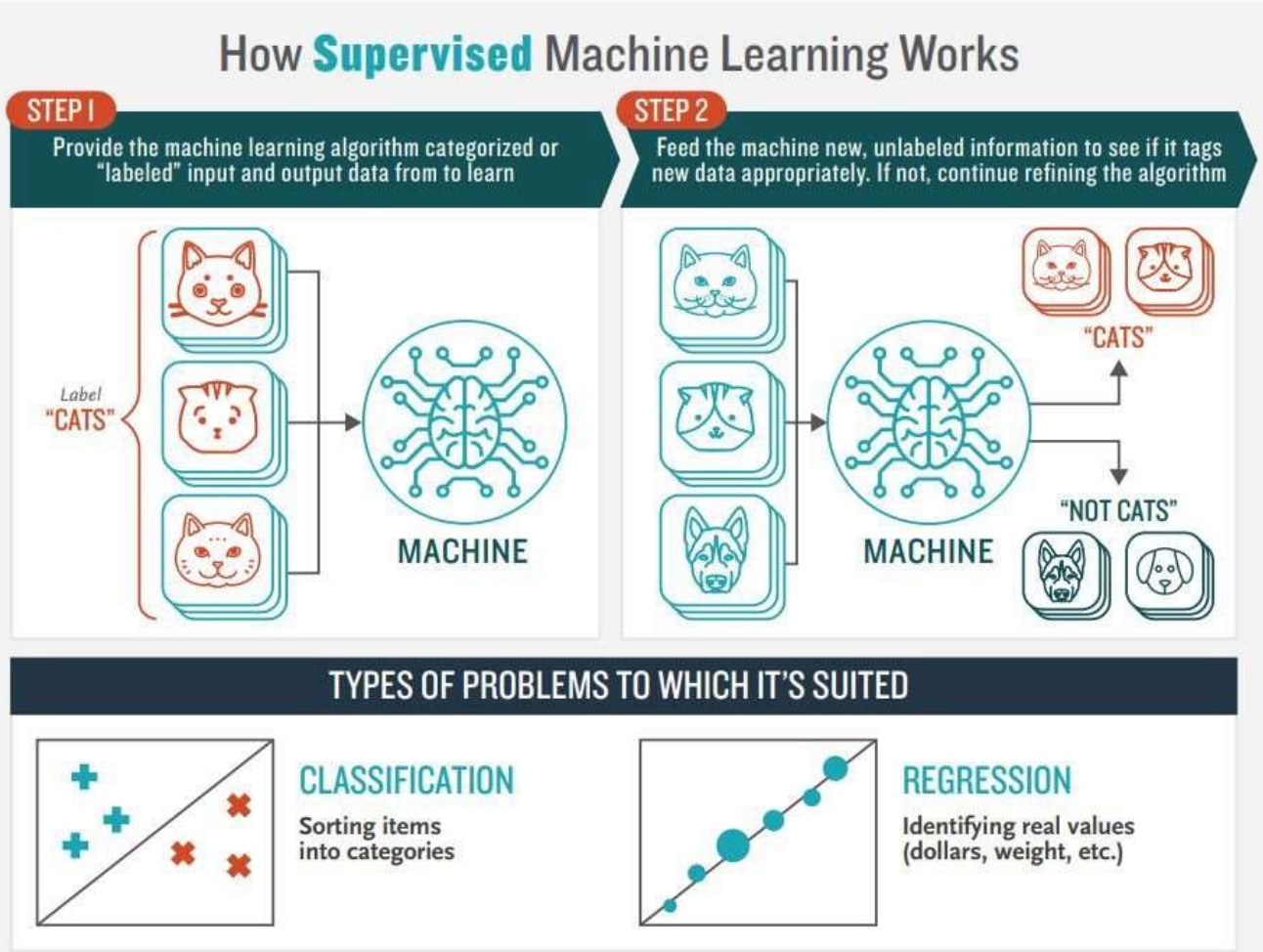




sumber: : https://linkedin.com/pulse/business-intelligence-its-relationship-big-data-geekstyle

MACHINE LEARNING (ML)

- Proses untuk membuat mesin dapat mempelajari pola dari data tanpa diprogram secara eksplisit.
- Manusia belajar dari pengalaman, tetapi Mesin mengikuti aturan (rules). ML membuat komputer belajar dari pengalaman.
- Metode-metode Machine Learning:
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning



sumber: https://www.uc.ac.id/ict/perbedaan-supervised-learning-and-unsupervised-learning/

Supervised Learning

- **Labelled** Raw Data
- Algoritma belajar **membandingkan** output sebenarnya dengan output yang benar (Output yang sudah diketahui) untuk menemukan error.
- Biasanya digunakan untuk **memprediksi** kejadian di masa mendatang berdasarkan data historis.
- Hasil berupa:
  - Accuracy
  - Precision
  - Recall
- Contoh: Mengklasifikasikan spam dalam folder terpisah dari kotak masuk.
- Dua tipe problems di Supervised Learning:
  - Classification
  - Regression

PROBLEMS TYPES IN SUPERVISED LEARNING

CLASSIFICATION

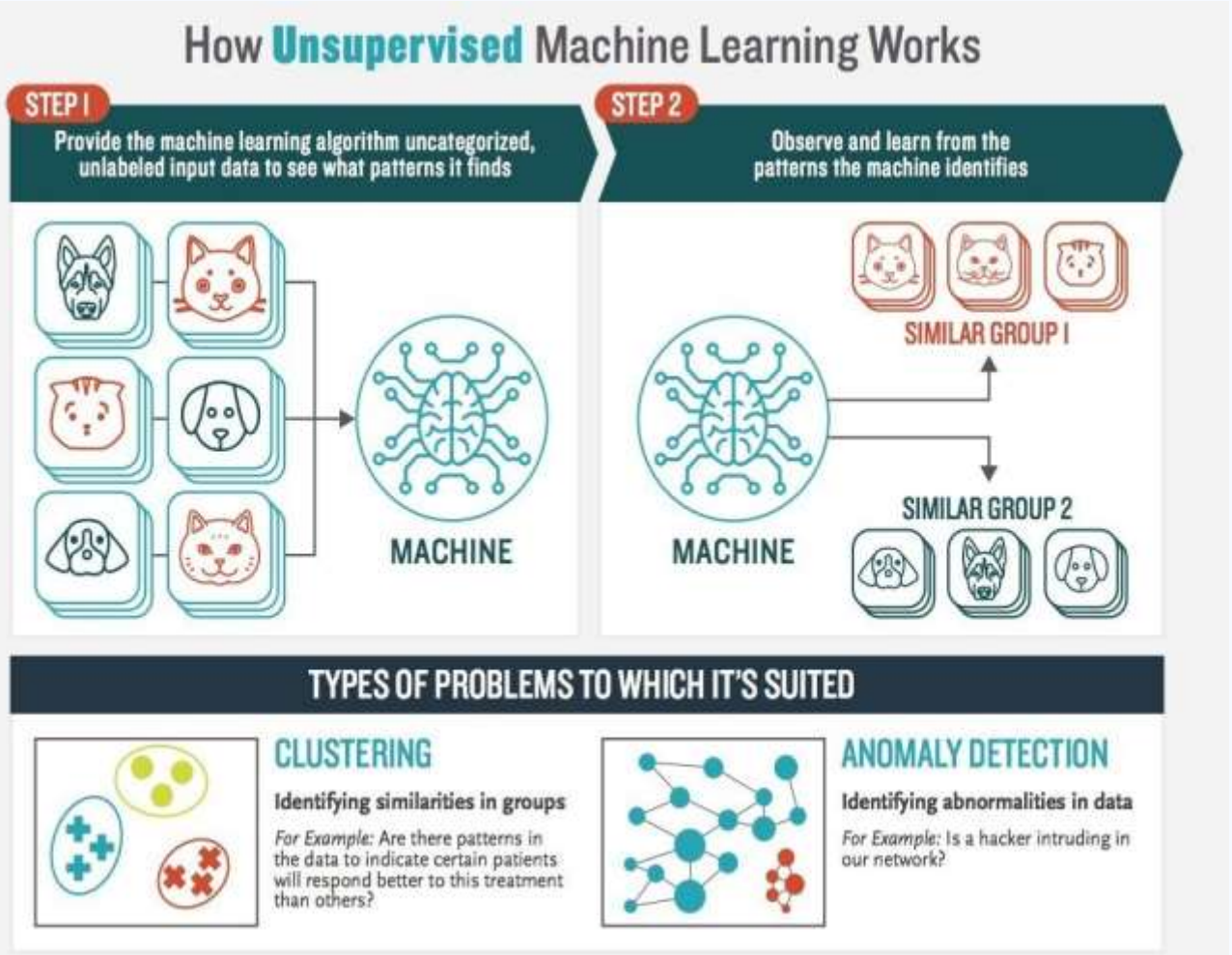
- Menggunakan algoritma untuk secara akurat menetapkan data uji ke dalam kategori tertentu.
- Algoritma:
  - Linear Classifiers
  - Support Vector Machines (SVM)
  - Decision Trees
  - k-nearest neighbor (k-nn)
  - Random Forest
  - Arima
- Contoh:
  - Deteksi Penipuan Identitas
  - Klasifikasi Gambar dan Objek
  - Retensi Pelanggan
  - Diagnostik

REGRESSION

- Memahami hubungan antara variabel terikat dan variabel bebas.
- Algoritma:
  - Linear Regression
  - Logistical Regression
  - Polynomial Regression
  - Lasso
  - Ridge
- Contoh:
  - Prediksi Popularitas Iklan
  - Prakiraan Cuaca
  - Prakiraan Pasar
  - Memperkirakan Harapan Hidup
  - Prediksi Pertumbuhan Penduduk

PERBEDAAN DARI CLASSIFICATION DAN REGRESSION

Dasar untuk perbandingan	Regresi	Klasifikasi
Yang Diprediksi	Angka	Kelas
Melibatkan prediksi	Continuous	Discrete
Model Sebagai	Best Fit Line	Decision Boundary
Algoritma	Pohon regresi (Hutan acak), Regresi linear, dll.	Pohon keputusan, regresi logistik, dll.
Sifat data yang diprediksi	Dipesan	Tidak dipesan
Metode perhitungan	Pengukuran root mean square error	Mengukur akurasi



sumber: https://www.uc.ac.id/ict/perbedaan-supervised-learning-and-unsupervised-learning/

Unsupervised Learning

- **Unlabelled** Raw Data
- Algoritma akan menganalisis dan mengelompokkan kumpulan data yang tidak berlabel.
- Algoritma ini menemukan **pola** tersembunyi atau pengelompokan data tanpa perlu campur tangan manusia.
- Dua tipe problems di Unsupervised Learning:
  - Clustering
  - Dimensionality Reduction

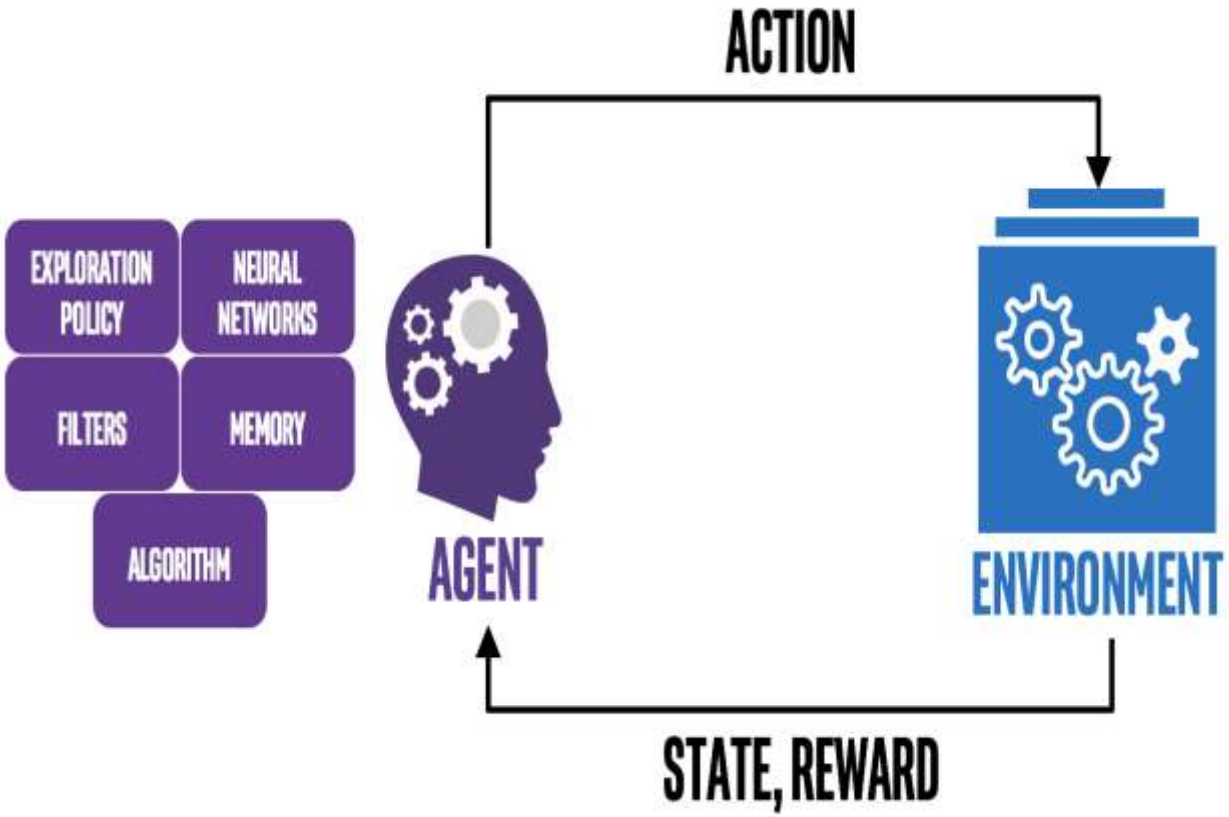
PROBLEMS TYPES IN UNSUPERVISED LEARNING

CLUSTERING

- Mengelompokkan data yang tidak berlabel berdasarkan persamaan atau perbedaannya.
- Contoh:
  - Sistem Rekomendasi
  - Pemasaran yang ditargetkan
  - Segmentasi Pelanggan

DIMENSIONALITY REDUCTION

- Digunakan ketika jumlah fitur, atau dimensi, dalam kumpulan data yang diberikan terlalu tinggi untuk mengurangi jumlah input data ke ukuran yang dapat dikelola.
- Contoh:
  - Visualisasi Big Data
  - Meaningful Compression
  - Penemuan Struktur
  - Fitur Elisitasi



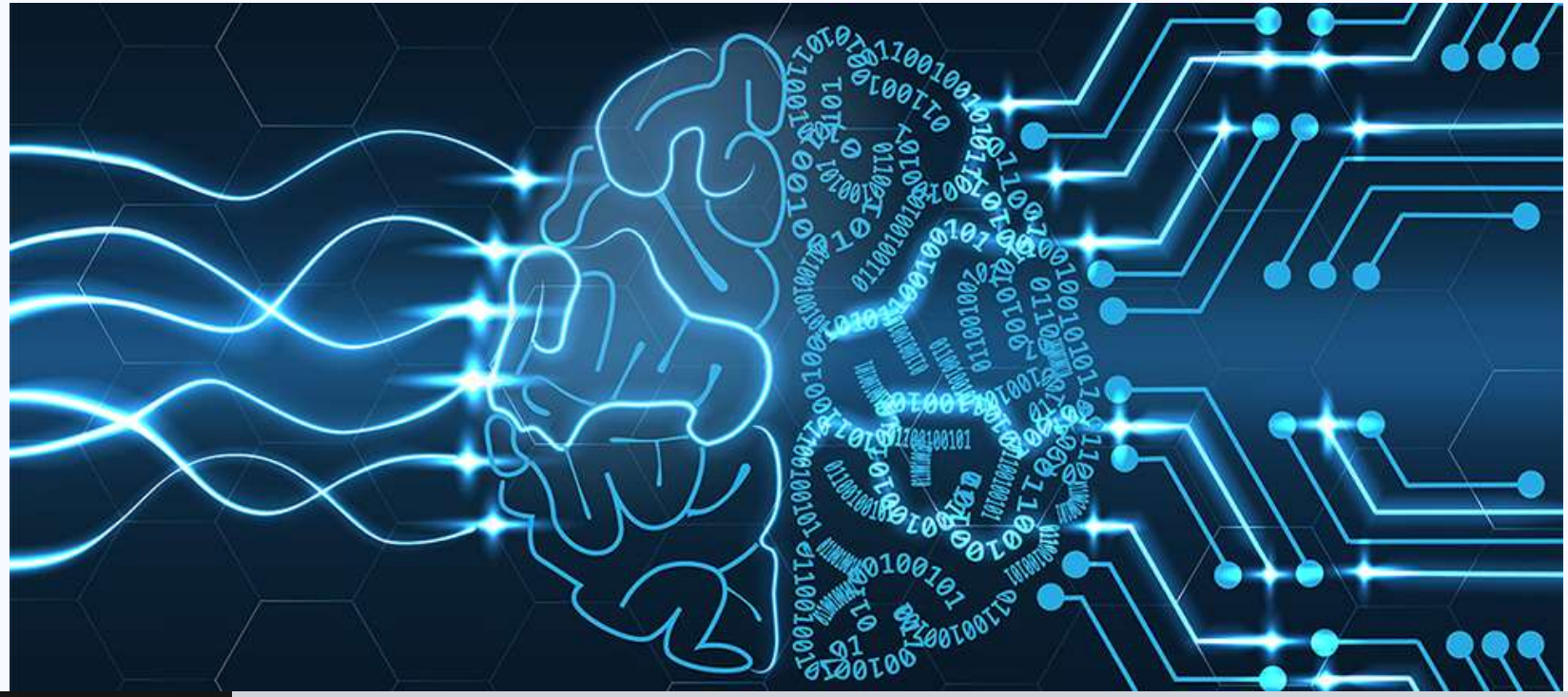
sumberhttps://intellabs.github.io/coach/_images/design.png

Reinforcement Learning

Apa yang mesti dilakukan (mengimplementasikan aksi kedalam situasi) pada sebuah masalah/problem untuk mendapatkan hasil/reward yang maksimal.

DEEP LEARNING (DL)

Proses untuk membuat mesin dapat mengerti pola dari data yang menggunakan algoritma yang meniru cara kerja pikiran (neural network).



DATA SCIENCE IN **BUSINESS**

FINANCIAL INDUSTRY

- Investment Decision

Memaksimalkan pengembalian berdasarkan banyaknya markey signals dan alternatif sumber data.

- Fraud Prevention

Mendeteksi dan mencegah aktivitas penipuan (fraud) dengan memanfaatkan Machine Learning untuk memprediksi anomali secara real time.

- Credit Scoring

Menentukan kelayakan kredit seseorang atau bisnis kecil yang dioperasikan oleh owner.

MEDIA & ENTERTAINMENT

- Sentiment Analysis

Memahami bagaimana konten beresonansi di saluran sosial.

- Content Personalization

Menampilkan konten yang dipersonalisasi saat mengakses platform untuk meningkatkan pengalaman user.

- Churn Prediction

Memprediksi pengguna mana yang cenderung melakukan churn, menganalisis faktor-faktornya, dan mencegah mereka dari churn.

EXPLORATORY DATA ANALYSIS (EDA)

- **Critical process** dalam melakukan investigasi awal pada data
- Tujuan:
  - Menemukan pola dari data.
  - Menguji validitas hipotesa dan asumsi.

DATA SCIENCE METHODOLOGY

WEEK 1 – SESSION 3

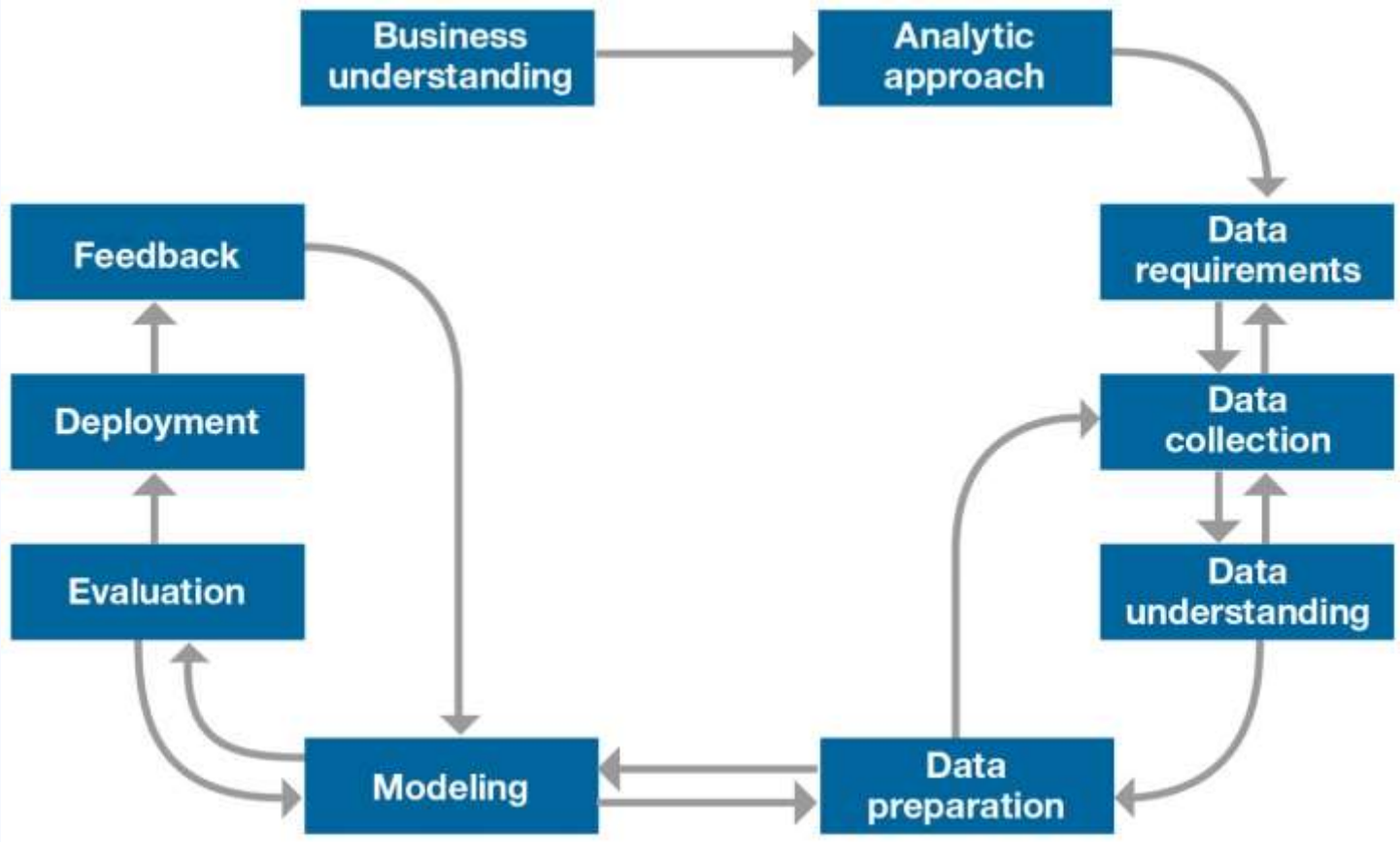
DS OMICRON – BATCH 11

DigitalSkola

OUTLINE

1. Business Understanding
2. Analytic Approach
3. Data Requirements
4. Data Collection
5. Data Understanding
6. Data Preparation
7. Modelling
8. Model Evaluation
9. Deployment
10. Feedback

DATA SCIENCE METHODOLOGY



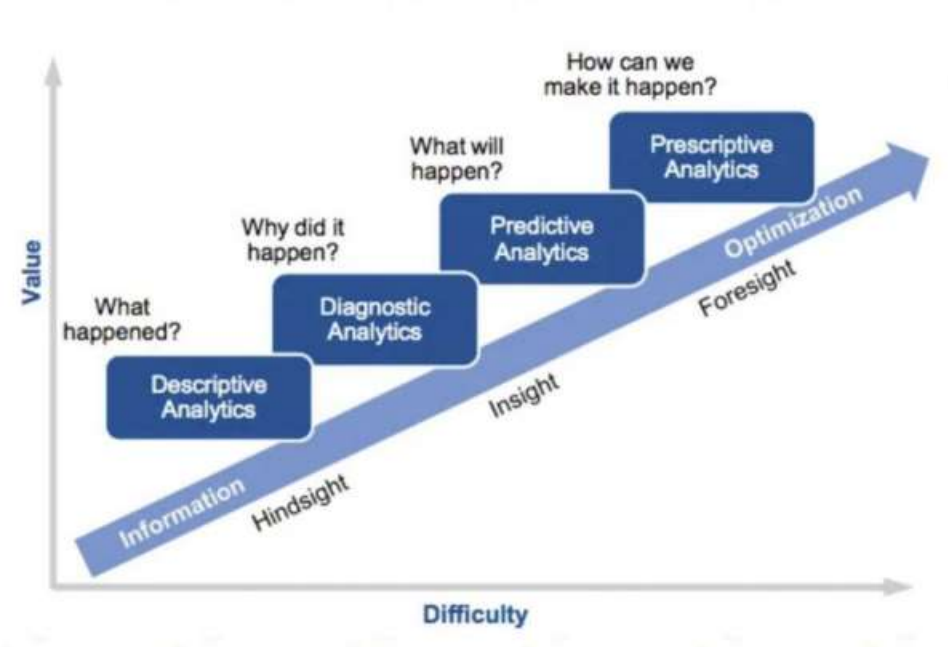
sumber: https://www.ibmbigdatahub.com/blog/why-we-need-methodology-data-science

BUSINESS UNDERSTANDING

- Pada tahapan business understanding, kita harus menanyakan secara detail kepada pihak terkait untuk memahami masalah apa yang sedang terjadi hingga pada akhirnya kita dapat memperolah business requirement.

- Tahapan business understanding:
  - Mencatat dan mendefinisikan masalah bisnis yang sedang terjadi, lalu membuat daftar skala prioritas.
  - Menetapkan tujuan bisnis.
  - Menentukan kriteria tingkat kesuksesan.

ANALYTIC APPROACHES



sumber: : https://www.datascienceassn.org/content/descriptive-predictive-prescriptive-analytics

Analytical approach digunakan untuk membuat analisis terhadap masalah bisnis yang terjadi. Ada 4 jenis analisis yang bisa digunakan, antara lain :

1. DESCRIPTIVE
   - Keadaan sekarang / Masalah yang dihadapi sekarang
2. DIAGNOSTIC
   - Apa yang terjadi?
   - Mengapa ini terjadi?
3. PREDICTIVE
   - Bagaimana jika tren ini berlanjut?
   - Apa yang akan terjadi?
4. PRESCRIPTIVE
   - Bagaimana cara menyelesaikannya?

ANALYTIC APPROACHES

DESCRIPTIVE ANALYTIC

1. Membuat matriks bisnis
2. Mengidentifikasi data yang diperlukan
3. Menarik dan menyiapkan data
4. Menganalisa data
5. Menyajikan data

DIAGNOSTIC ANALYTIC

1. Mengidentifikasi anomali yang terjadi
2. Menemukan akar permasalahan
3. Menentukan penyebabnya

ANALYTIC APPROACHES

PREDICTIVE ANALYTIC

1. Mengidentifikasi tujuan bisnis
2. Menentukan required data sebagai training data
3. Menentukan tipe analisis yang digunakan
4. Validasi hasil
5. Tes predicted data

PRESCRIPTIVE ANALYTIC

1. Mengidentifikasi tujuan bisnis
2. Menentukan required data sebagai training data
3. Menentukan tipe analisis yang digunakan
4. Validasi hasil
5. Tes predicted data

DATA REQUIREMENTS

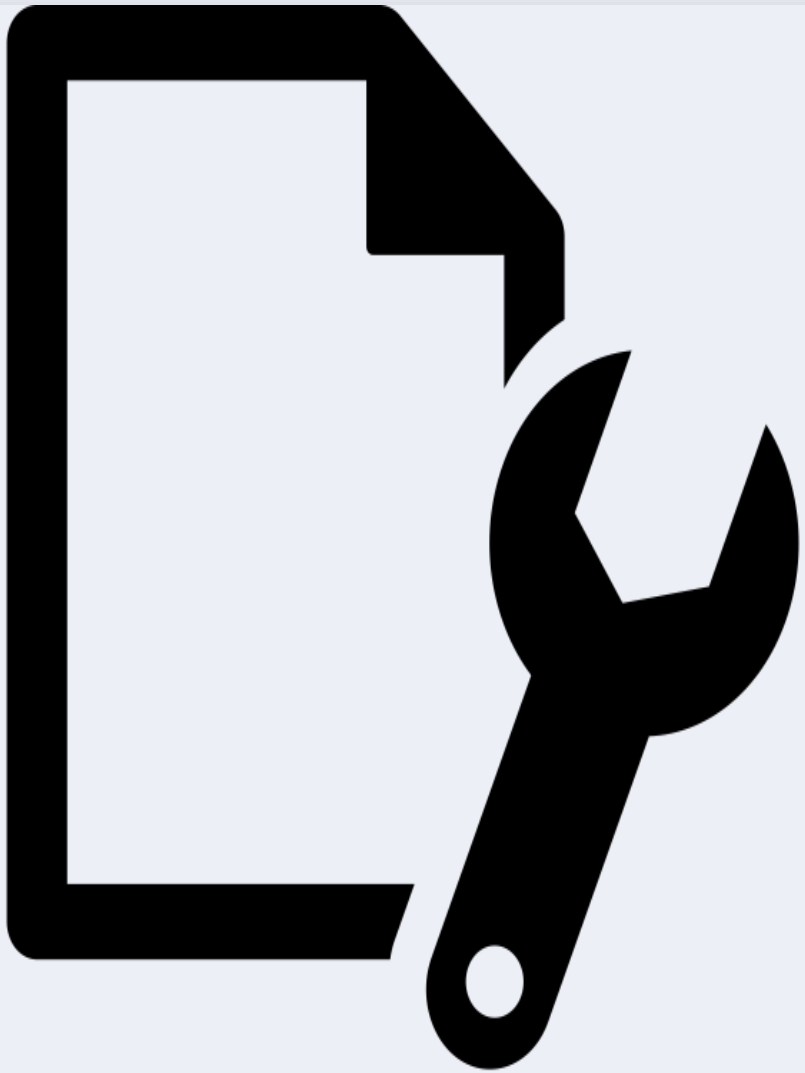
Diperlukan untuk mengidentifikasi data/ konten yang diperlukan sebagai basis analisis proses

DATA COLLECTION

- Diperlukan untuk menentukan sumber data yang tersedia apakah relevan dengan permasalahannya, sehingga seluruh sumber data akan dikumpulkan.
- Sumber data collection ada yang bersifat internal maupun eksternal. Data internal seperti data transaksi, data produk, dokumen excel dsb. Data eksternal seperti hasil survei public, data repository dsb.

DATA UNDERSTANDING

- Diperlukan untuk membangun data set dan memahami anomali-anomali yang terjadi
- Perlu eksplorasi data apakah data yang dibutuhkan sudah cukup untuk menangani problem yang ditentukan.



DATA PREPARATION

- Bersifat teknikal, handling data seperti menghapus data duplikat, mengatur data yang hilang dengan drop maupun impute data, mengecek format data apakah sudah sesuai
- Setelah proses ini, diharapkan tidak ada data yang eror

MODELLING

Modelling adalah proses pengembangan sekumpulan data yang telah dipersiapkan melalui proses data preparation untuk menentukan algoritma ML yang akan digunakan.

ANALYTIC APPROACHES

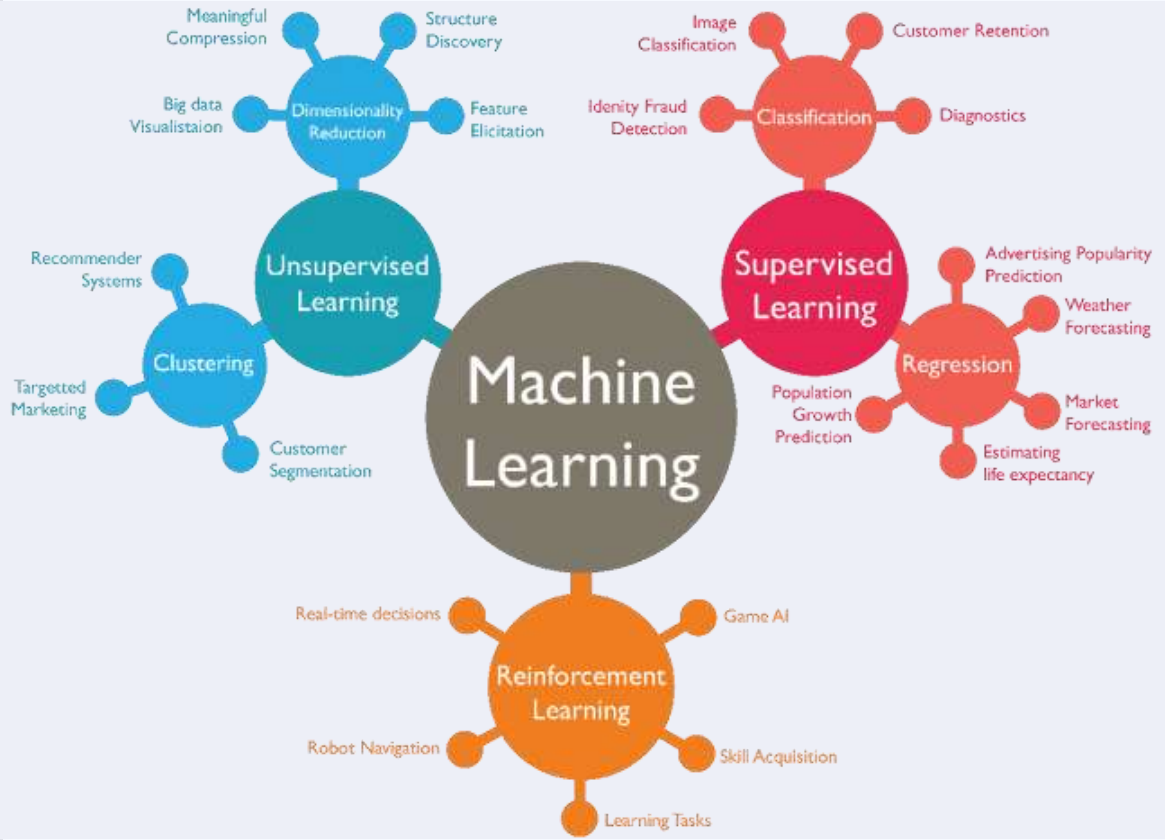
DESCRIPTIVE

Proses matematis yang dapat menjelaskan suatu kondisi di dunia nyata secara aktual dan hubungan antara faktor-faktor yang mempengaruhinya.

PREDICTIVE

Proses yang menggunakan data mining dan probabilitas untuk meramalkan apa yang akan terjadi di masa mendatang.

MODELLING



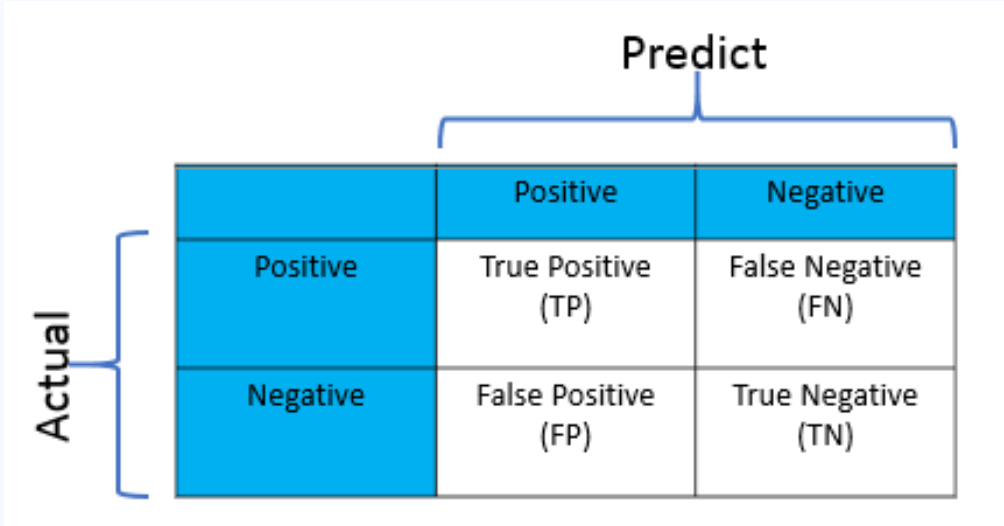
sumber: : https://linkedin.com/pulse/business-intelligence-its-relationship-big-data-geekstyle

Tipe algoritma yang digunakan disesuaikan dengan kasus yang berhubungan.

Misal:

Dalam kasus klasifikasi, bisa menggunakan ML Decision Tree yang menggunakan Gini Index, atau SVM yang memperhitungkan jarak hyperplan.

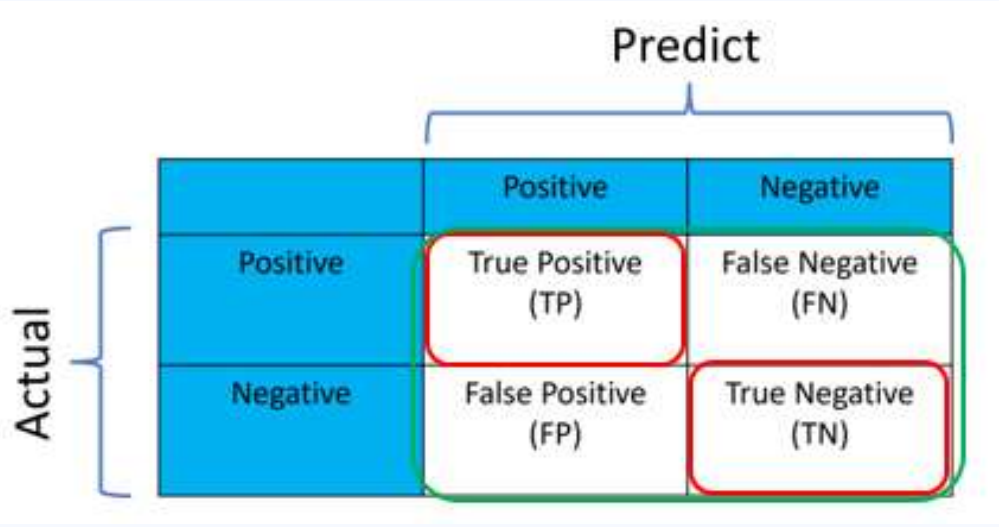
MODEL EVALUATION



- Proses pengujian hasil perhitungan algoritma ML dengan current data menggunakan confusion matrix
- Standar yang diuji adalah nilai accuracy, precision dan recall

MODEL EVALUATION

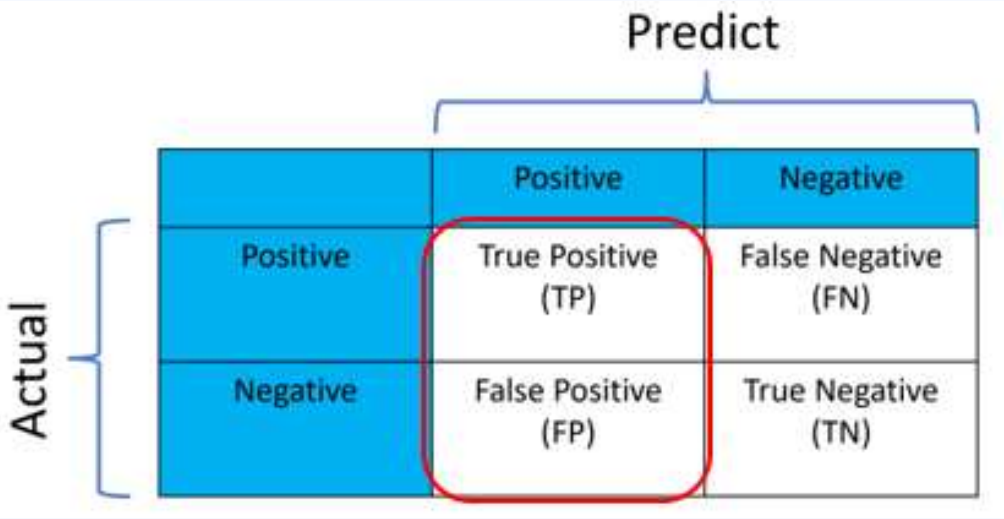
Accuracy



- Accuracy adalah total data yang terprediksi dengan tepat dari total data yang dimiliki
- Accuracy = $\frac{TP+TN}{TP+FP+TN+FN}$

MODEL EVALUATION

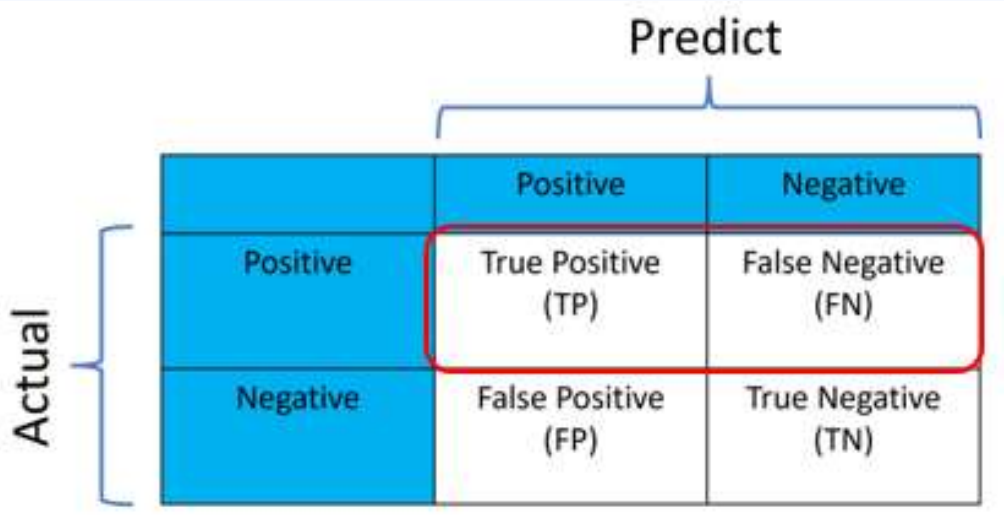
Precision



- Precision adalah nilai ketepatan prediksi positif, yaitu persentase sample yang benar positif dari total sample yang diprediksi positif
- Precision = $\frac{TP}{TP+FP}$

MODEL EVALUATION

Recall



- Recall adalah nilai sensitivitas prediksi atau kemampuan algoritma untuk mengenali sample positif, yaitu persentase sample yang terprediksi positif dari total sample yang aktual positif

- Recall = $\frac{TP}{TP+FN}$

DEPLOYMENT

DESCRIPTIVE

- Menyajikan temuan-temuan yang didapat dari proses analisis
- Menggambarkan situasi yang sedang terjadi
- Menemukan faktor yang menyebabkan situasi yang sedang terjadi

PREDICTIVE

- Mengembangkan Machine Learning Model untuk memprediksi data yang akan datang
- Membuat daftar aktivitas untuk mencapai target di masa mendatang

FEEDBACK

DESCRIPTIVE

Menjelaskan dengan baik penyebab terjadinya temuan

PREDICTIVE

Menjaga performa model ML yang digunakan

Memastikan angka prediksi tepat atau bahkan melampaui syarat yang telah ditentukan

Jika terjadi penurunan, segera dilakukan re-modelling dari feature engineering sampai ML model building

TERIMA KASIH

DS OMICRON – BATCH 11