



— + —

# LEARNING PROGRESS REVIEW

## Week 1

— + —

**Kelompok 7 - Citizen Data Scientist**

Diaz Jubiary - Hermulia Hadie - Desi Sulistyowati - Farahul Jannah



# Daftar Isi

## Key Presentation Point

- Apa itu Data Science?
- Subset Data Science
- Tipe Machine Learning
  - Supervised Learning
  - Unsupervised Learning
  - Semi-supervised Learning
  - Reinforcement Learning
- Data Science Life Cycle
- Data Scientist vs Profesi Data Lainnya
- Metodologi Data Science

# Data Science

## Apa itu Data Science?

- Bidang ilmu yang menggunakan **metode, proses, algoritma, dan kaidah ilmiah** untuk mengekstrak *knowledge* dan *insight* dari data mentah (*raw data*)
- Data mentah terbagi menjadi 2, yaitu **structured data** (tabel atau spreadsheet) dan **unstructured data** (teks, gambar, audio dan video)

Tujuan Data Science adalah:

- memberikan informasi suatu pola pada data, saran, masukan temuan, ataupun model yang berbasis data (data-driven)
- memberikan solusi (data-driven solutions)
- pengambilan keputusan (data-driven decisions)

terhadap permasalahan yang ingin diselesaikan dalam suatu bidang (finansial, bisnis, perbankan, *e-commerce*, dll)

# Data Science

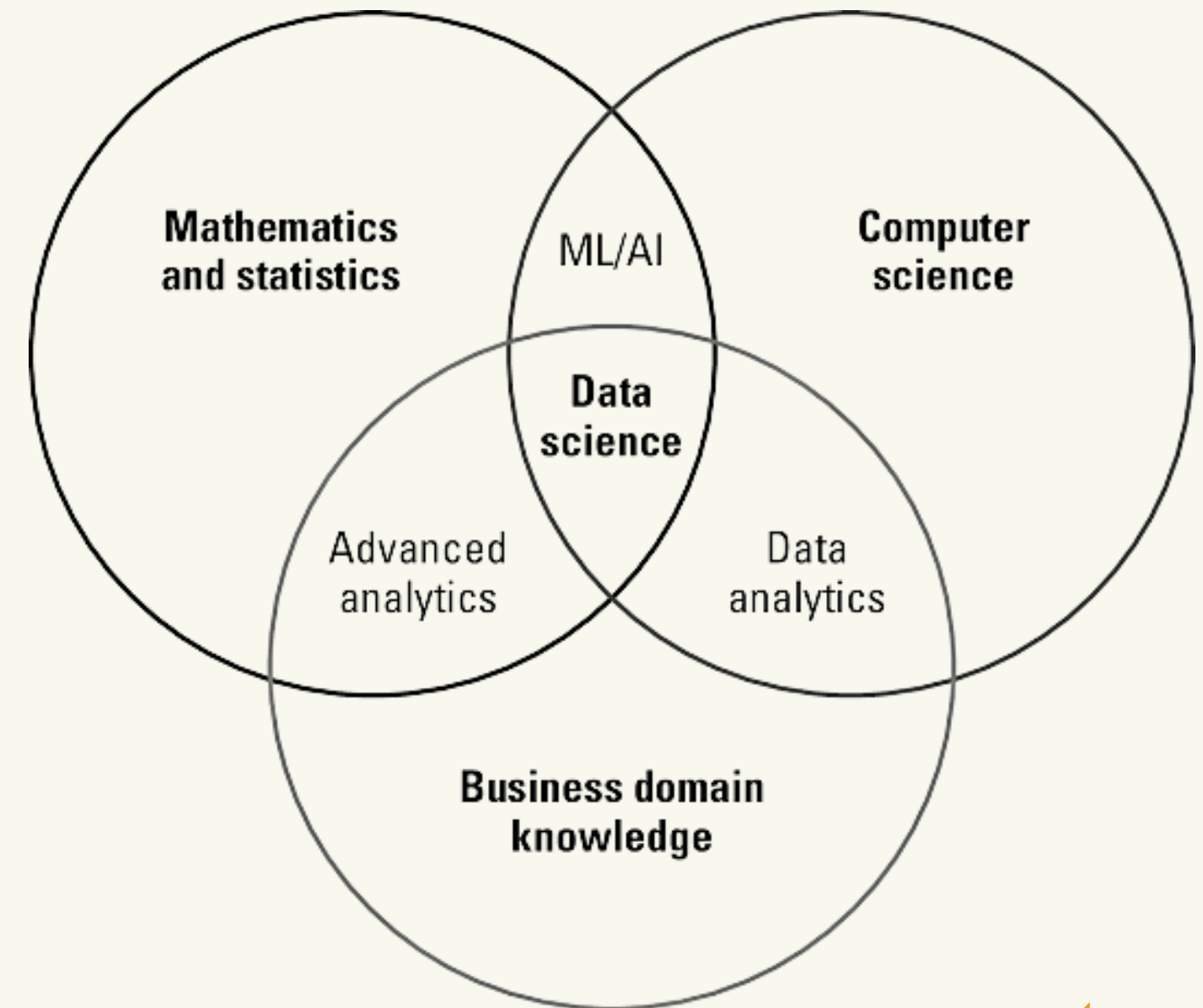
## Studi Interdisipliner

Data Science merupakan bidang ilmu yang menggunakan teknik dan teori dari ilmu:

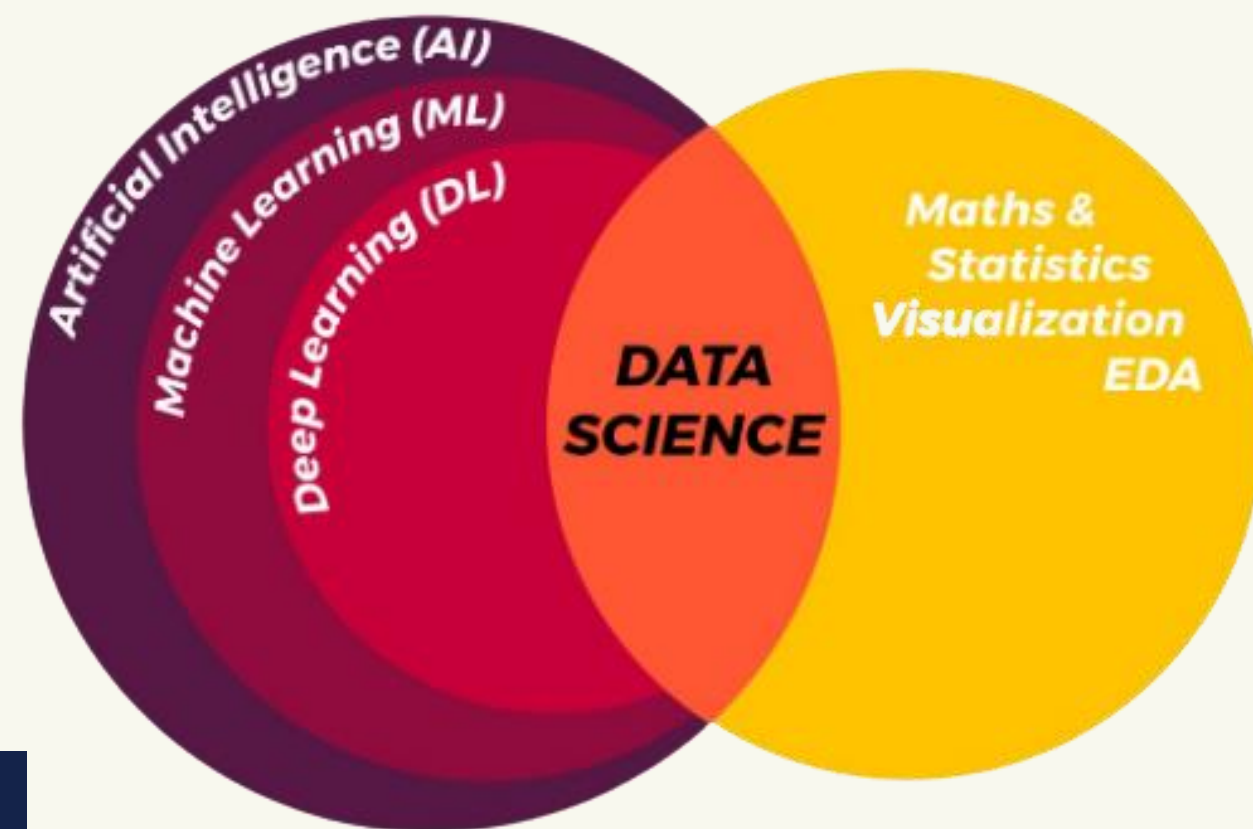
- matematika dan statistik
- ilmu komputer dan ilmu informasi
- *business domain knowledge*

untuk menganalisis fenomena aktual menggunakan data.

Sehingga seorang Data Scientist harus mempunyai kemampuan teknis dan non-teknis dari ketiga bidang keilmuan tersebut



# Subset dan Bidang Terkait Data Science



Data Science adalah **persimpangan** atau **perpotongan** antara **matematika** dan **AI (*artificial intelligence*)** bersama dengan semua subset yang berada di dalamnya (*machine learning & deep learning*).

Hal ini dikarenakan terdapat hal-hal lain diluar Data Science yang berkaitan dengan AI, ML, dan DL.

- **AI (*Artificial Intelligence*)** : cabang ilmu komputer yang mengembangkan sistem komputer dapat melakukan tugas yang biasanya membutuhkan kecerdasan manusia.
- **ML (*Machine Learning*)** : bagian dari AI yang bertujuan mengajari suatu mesin mempelajari pola data tanpa memprogram secara eksplisit.
- **DL (*Deep Learning*)** : bagian dari ML yang menggunakan sekumpulan algoritma berdasarkan strukutr otak manusia atau jaringan saraf



# TIPE MACHINE LEARNING

## SUPERVISED LEARNING

Proses pengelompokan data yang telah memiliki label dan akan dikelompokkan berdasarkan labelnya.

## UNSUPERVISED LEARNING

Proses pengelompokan data yang tidak memiliki label

## SEMI-SUPERVISED LEARNING

Proses pengelompokkan data dalam jumlah besar yang sebagian datanya mempunyai label dan tidak

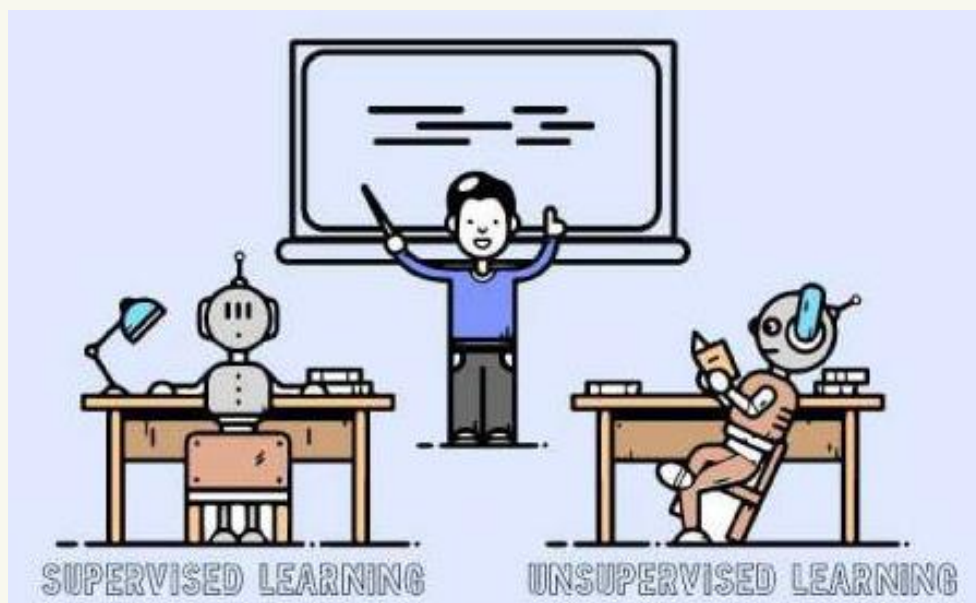
## REINFORCEMENT LEARNING

Komputer sudah dapat bekerja secara otomatis untuk menentukan perilaku yang ideal sehingga dapat memaksimalkan kinerja algoritmanya.

# 1. SUPERVISED LEARNING

## Algoritma

- Support vector machine
- Neural network
- Linear and logistics regression
- Random forest
- Classification trees



- *Supervised Learning* adalah *Machine Learning* model yang mempelajari data dengan label atau target dimana evaluasi model tersebut akan berdasarkan target ini.
- *Supervised Learning* membutuhkan *data training*.
- Model ini dilatih melakukan prediksi berdasarkan pola yang ditentukan dalam menjawab data target

## Contoh:

Film komedi: 21 Jump Street dan Jumanji

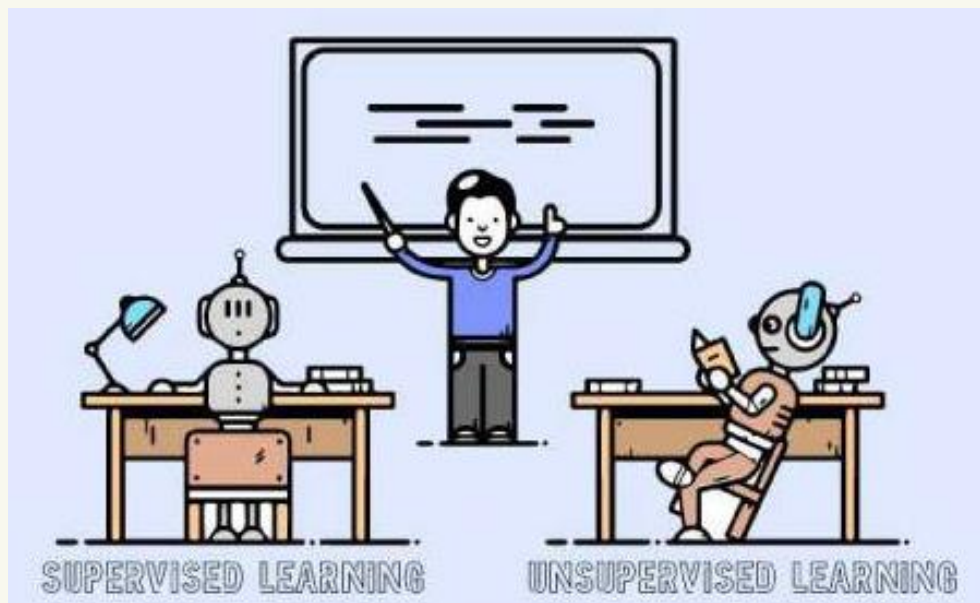
Film horor: The Conjuring dan IT

Pada saat kita menambahkan film A Quiet Place maka ML akan mengidentifikasi apakah film tersebut termasuk dalam kategori film komedi atau horor.

## 2. UNSUPERVISED LEARNING

### Algoritma

- K-Means
- Hierarchical Clustering
- DBSCAN
- Fuzzy C-Means



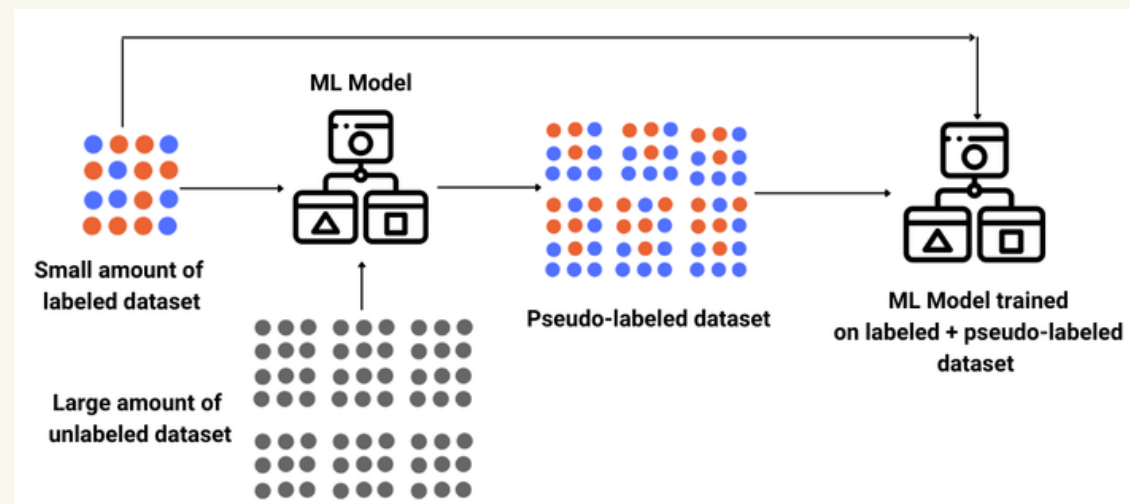
- *Unsupervised learning* merupakan teknik ML yang digunakan pada data yang tidak memiliki informasi/label sebelumnya.
- Teknik melakukan analisis korelasi antar data untuk menemukan pola data yang tidak memiliki label.
- Teknik ini tidak memiliki data apapun yang dijadikan sebagai acuan

### Contoh:

Minggu lalu, Kale membeli sejumlah DVD film, kemudian dia ingin membaginya ke dalam beberapa kategori berdasarkan genre-nya agar mudah ditemukan. Sehingga Kale mulai mengidentifikasi DVD film tersebut berdasarkan kemiripannya jalan ceritanya



### 3. SEMI-SUPERVISED LEARNING



- Gabungan dari dua tipe pembelajaran ML, yaitu supervised learning dan unsupervised learning.
- Dapat bekerja menggunakan data berlabel untuk data dalam skala kecil dan data tidak berlabel untuk data dalam skala besar.
- Mempunyai 2 model, yaitu **metode inductive** (memberikan label pada data baru tanpa melakukan training data) dan **metode transductive** (memberikan label pada data baru dengan melakukan training data)

#### Contoh:

- **Metode inductive:** *image recognition* dan *sentiment analysis*
- **Metode transductive:** model grafik

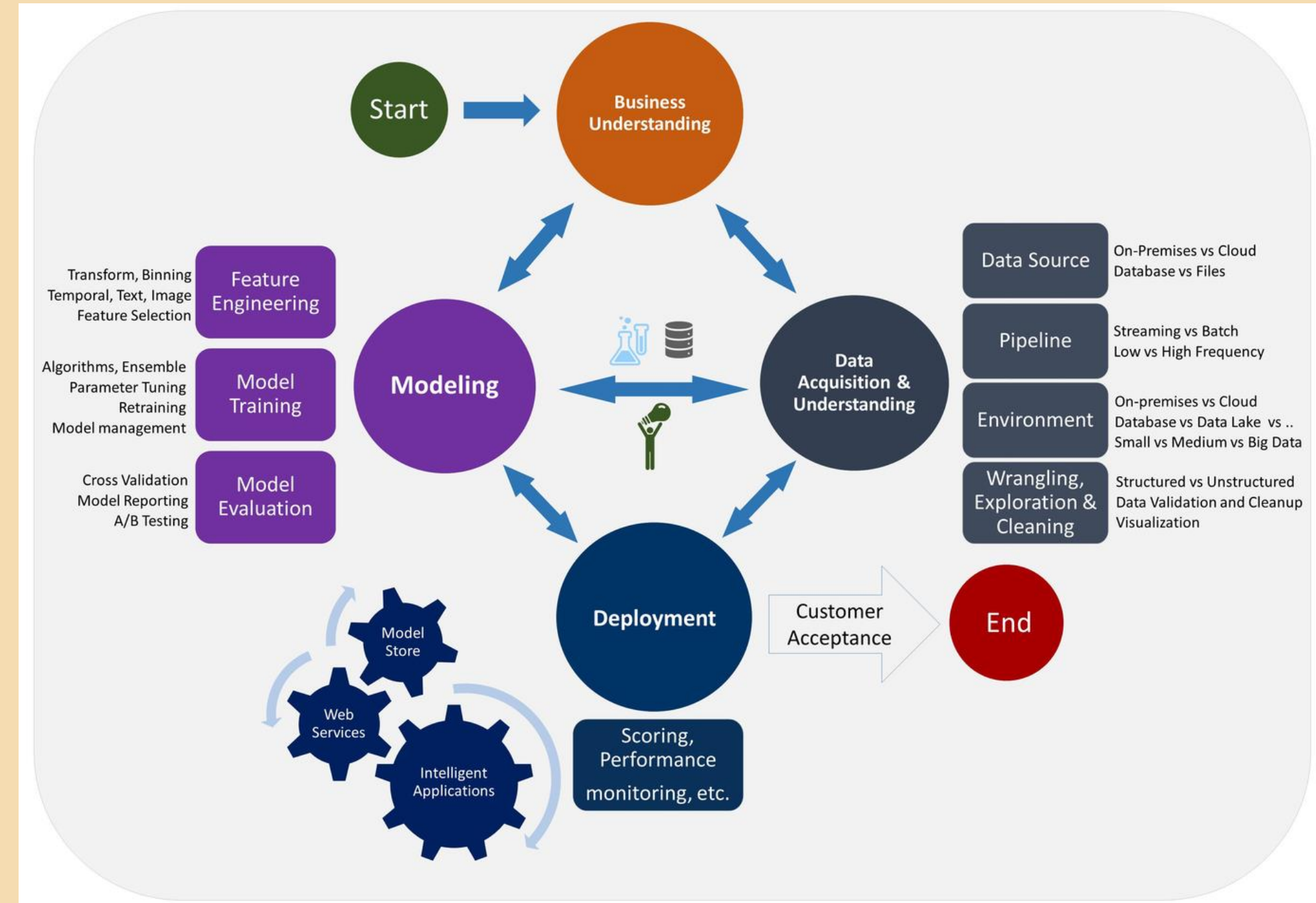
## 4. REINFORCEMENT LEARNING

- Tipe ML untuk pengambilan keputusan
- Mampu menemukan aksi atau perlakuan untuk menghasilkan output terbaik dengan uji coba berulang kali yang didapatkan dari lingkungan yang mempengaruhinya sehingga menambah pengetahuannya agar bisa memecahkan masalah
- Proses ini akan terus berlangsung dan mengurangi interaksi atau keterlibatan manusia serta menghemat waktu dalam memecahkan masalah bisnis
- Biasanya tipe ini digunakan dalam dunia robotik, navigasi, dan develop game



# Data Science Life Cycle

- Memformulasikan permasalahan dan tujuan;
- Menyiapkan data;
- Memproses data agar menjadi bersih;
- Menganalisa atau membuat model;
- Mengkomunikasikan hasil;
- Mengevaluasi dan memonitor terhadap data yang digunakan;
- Menerapkan *knowledge, insight*, dan model yang didapat berdasarkan *business domain knowledge* dalam berbagai macam penggunaan dalam suatu bidang atau industri.





| Data Analyst                   | Data Engineer                             | Data Scientist  |
|--------------------------------|---|---|
| Data Warehousing               | Data Warehousing & ETL                    | Statistical & Analytical skills                       |
| Adobe & Google Analytics       | Advanced programming knowledge            | Data Mining   |
| Programming knowledge          | Hadoop-based Analytics                    | Machine Learning & Deep learning principles           |
| Scripting & Statistical skills | In-depth knowledge of SQL/ database       | In-depth programming knowledge (SAS/R/ Python coding) |
| Reporting & data visualization | Data architecture & pipelining            | Hadoop-based analytics                                |
| SQL/ database knowledge        | Machine learning concept knowledge        | Data optimization                                     |
| Spread-Sheet knowledge         | Scripting, reporting & data visualization | Decision making and soft skills                       |

# Data Scientist VS Profesi Data Lainnya

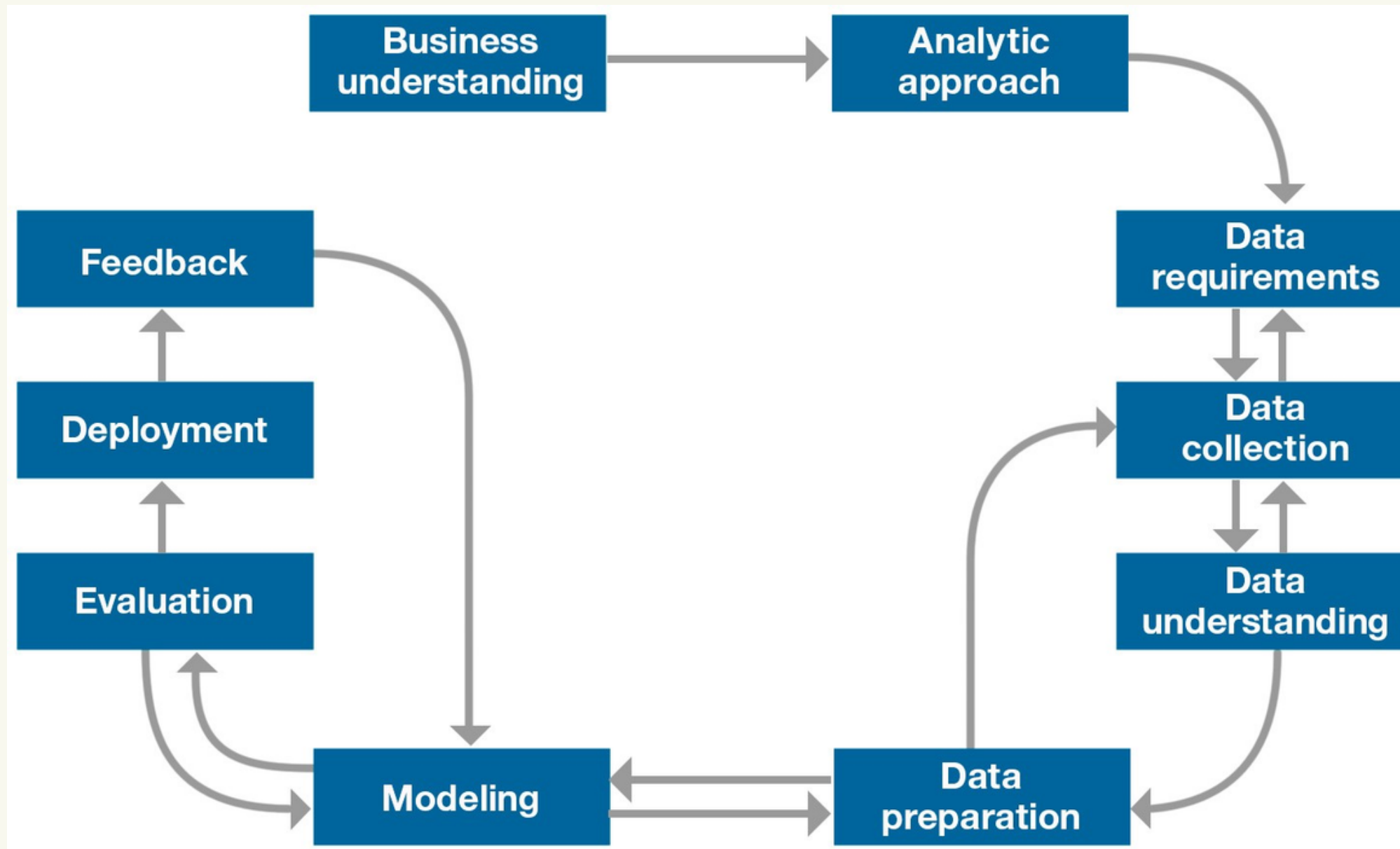
|                        | Domain Expertise                 | Technical Knowledge              | Quantitative Skills              |
|------------------------|----------------------------------|----------------------------------|----------------------------------|
| Data Scientist         | <input type="radio"/>            | <input type="radio"/>            | <input checked="" type="radio"/> |
| Data Engineer          | <input type="radio"/>            | <input checked="" type="radio"/> | <input type="radio"/>            |
| Data Science Architect | <input type="radio"/>            | <input checked="" type="radio"/> | <input type="radio"/>            |
| Data Science Developer | <input type="radio"/>            | <input checked="" type="radio"/> | <input type="radio"/>            |
| Product Owner          | <input type="radio"/>            | <input type="radio"/>            | <input type="radio"/>            |
| Data/Business Analyst  | <input type="radio"/>            | <input type="radio"/>            | <input type="radio"/>            |
| Process Master         | <input type="radio"/>            | <input type="radio"/>            | <input type="radio"/>            |
| Subject Matter Expert  | <input checked="" type="radio"/> | <input type="radio"/>            | <input type="radio"/>            |

Significant Expertise: ☒
 Some Expertise: ☐
 Minimal Expertise: ☐

# Metodologi Data Science

Langkah yang digunakan dalam proyek Data Science agar dapat menghasilkan hasil yang optimal dalam menjawab suatu permasalahan yang ingin diselesaikan





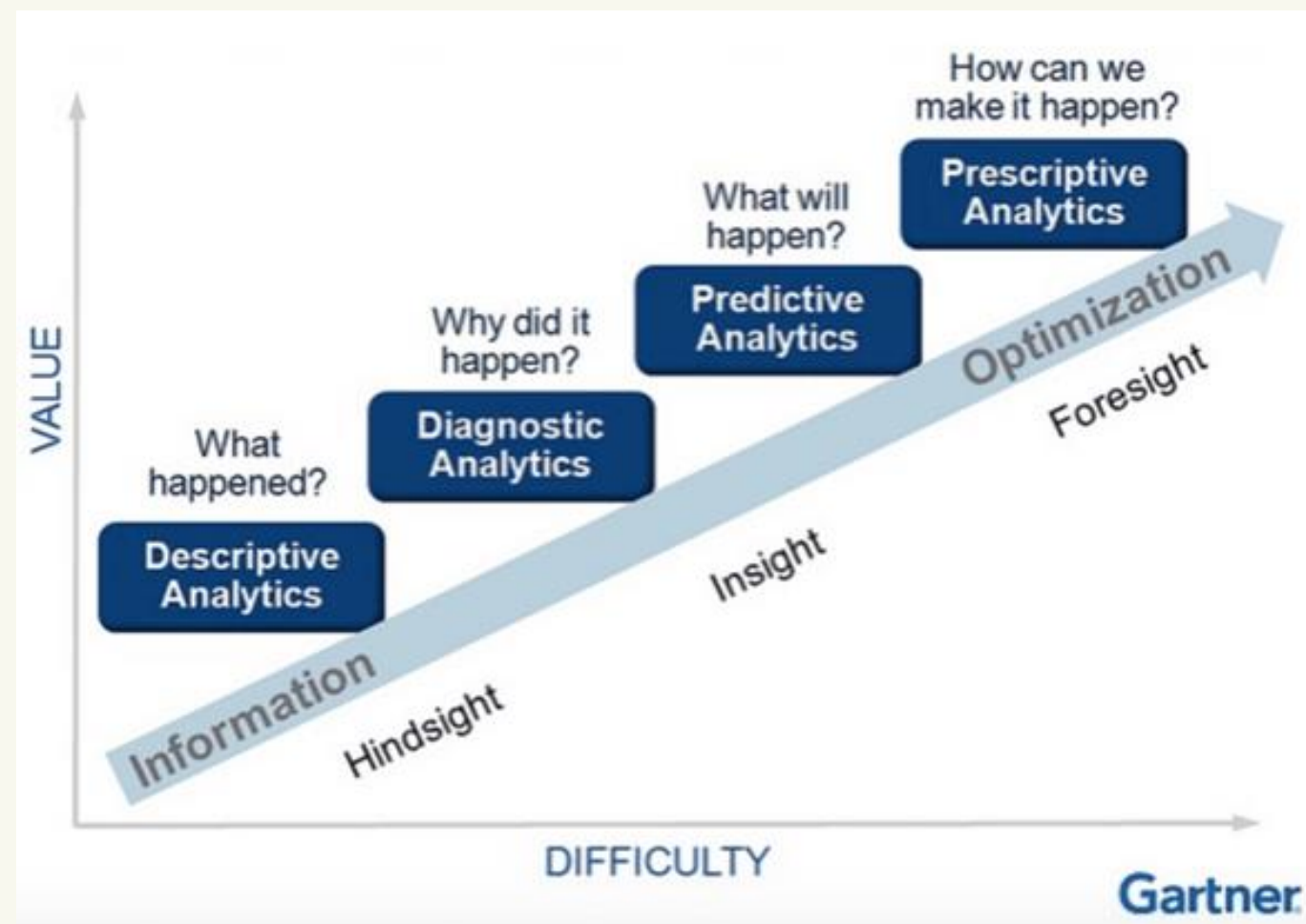
- Business understanding
- Analytic approach
- Data requirements
- Data collection
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment
- Feedback

# 1. Business Understanding



- Ini adalah tahapan pertama pada metodologi Data Science.
- Tahapan ini memiliki peranan penting karena pada tahap ini seorang *data scientist* mendefinisikan permasalahan, mengurutkan prioritas permasalahan, menentukan tujuan, dan menentukan kriteria keberhasilan.
- Tahapan ini dilakukan dengan cara berdiskusi dengan stake holder dan tim lain yang berkaitan dengan permasalahan yang akan diselesaikan.
- Pada akhir tahap ini, *data scientist* memiliki sejumlah daftar business requirements

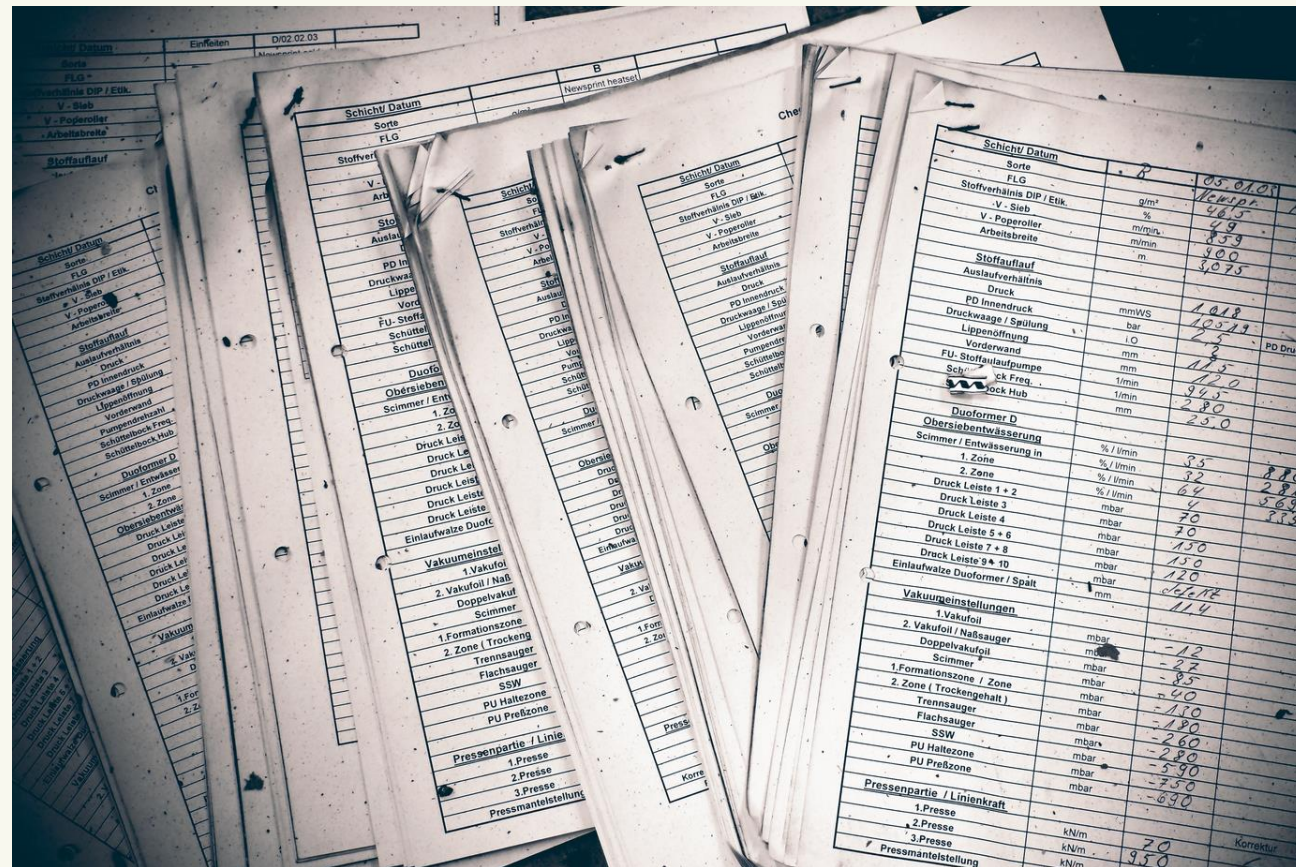
## 2. Analytic Approach



- Tahap kedua ini data scientist menentukan pendekatan analitik apa yang sesuai dengan permasalahan yang dihadapi.
- Pada tahap ini dilakukan pendefinisian masalah dalam konteks statistik atau *machine learning*.
- Pemilihan *analytic approach* ini berdasarkan pertanyaan/permasalahan yang dihadapi dan jenis pola/model apa yang efektif untuk menyelesaikan permasalahan tersebut.
- Ada 4 jenis *analytic approach*, yaitu:
  - **Descriptive analytics**: bertujuan menjelaskan permasalahan yang terjadi.
  - **Diagnostic analytics**: bertujuan mencari tahu kenapa masalah tersebut terjadi.
  - **Predictive analytics**: bertujuan mencari tahu apa yang mungkin terjadi kedepannya.
  - **Prescriptive analytics**: bertujuan untuk mengidentifikasi langkah untuk mencapai apa yang kita inginkan.



# 3. Data Requirements



- Pada tahap ini data scientist membuat list data yang sekiranya dibutuhkan untuk menjawab permasalahan atau pertanyaan yang telah diidentifikasi sebelumnya.
- Misalnya, mengidentifikasi konten, format, dan sumber-sumber data



## 4. Data Collection



- Tahap ini *data scientist* mulai mengumpulkan data yang relevan dengan domain masalah.
- Data yang dikumpulkan dapat dalam bentuk *structured*, *unstructured*, atau *semi-structured*.
- Apabila dalam tahap ini menemukan data yang dibutuhkan kurang maka data scientist bisa mengulangi lagi dari tahap *data requirements*.
- Contoh dari pengumpulan data berdasarkan sumbernya adalah sebagai berikut:

| DB Production  | DB Events Tracker   | Documents  |
|----------------|---------------------|------------|
| Data Transaksi | Data User Click     | File Excel |
| Data User      | Data User Page View | Notes      |
| Data Product   | Data User Scroll    |            |

Internal

| Data Public     | Data 3 <sup>rd</sup> Party | Scraping   |
|-----------------|----------------------------|------------|
| Open Data       | Data Survey                | File Excel |
| Data Repository | Data Vendor                | Notes      |
| Dashboard       |                            |            |

External

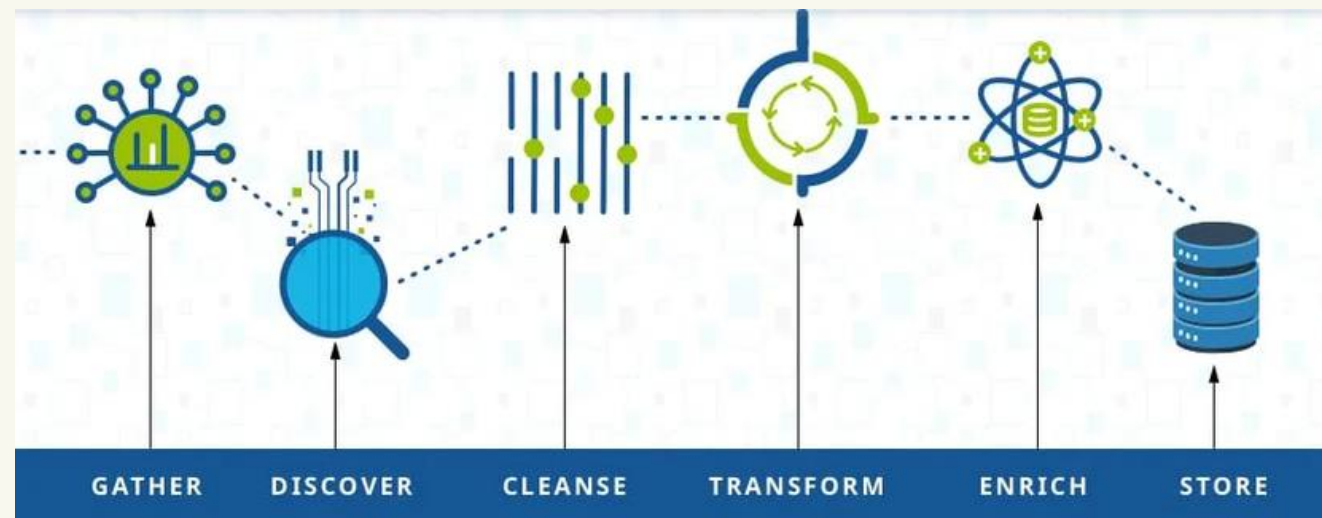


## 5. Data Understanding



- Proses mempelajari data dan melihat kualitas data yang kita miliki, apakah sudah cukup baik untuk proses modeling atau belum.
- Proses validasi apakah ada missing values, data yang imbalanced, outlier, salah format, dan sebagainya yang harus diperbaiki terlebih dahulu.
- Proses data understanding yang populer adalah dengan menggunakan statistik deskriptif dan teknik visualisasi. Teknik ini membantu data scientist memahami isi data, menilai kualitas data, dan menemukan insight awal dari data tersebut.

## 6. Data Preparation



- Proses yang dilakukan untuk membangun dataset yang akan digunakan dalam tahap pemodelan, termasuk membersihkan data, menggabungkan data, dan mengubah data menjadi variabel yang lebih berguna.
- Proses pembersihan data dari missing values, invalid values, dan data duplikat serta memastikan bahwa seluruh data telah memiliki format yang benar. Hal ini dilakukan supaya data dapat diproses secara efektif pada tahap pemodelan
- *Feature engineering* adalah proses transformasi data menjadi fitur-fitur yang lebih representatif dan dalam membantu menyelesaikan masalah dengan lebih baik. Fitur-fitur di dalam data sangat penting untuk model prediktif dan akan berdampak pada hasil yang ingin dicapai.

# 7. Modeling



- Tahap ini *data scientist* membuat model untuk menjawab permasalahan.
- Pemodelan berfokus pada pengembangan model yang deskriptif atau prediktif,
- Contoh:
  - Model deskriptif: Jika seseorang melakukan A, maka mereka mungkin menyukai B.
  - Model prediktif: Jenis model yang menghasilkan tipe jawaban ya/tidak atau maju/berhenti.
- Model ini bergantung pada analytical approach yang telah ditentukan sebelumnya, apakah menggunakan pendekatan statistik atau *machine learning*.

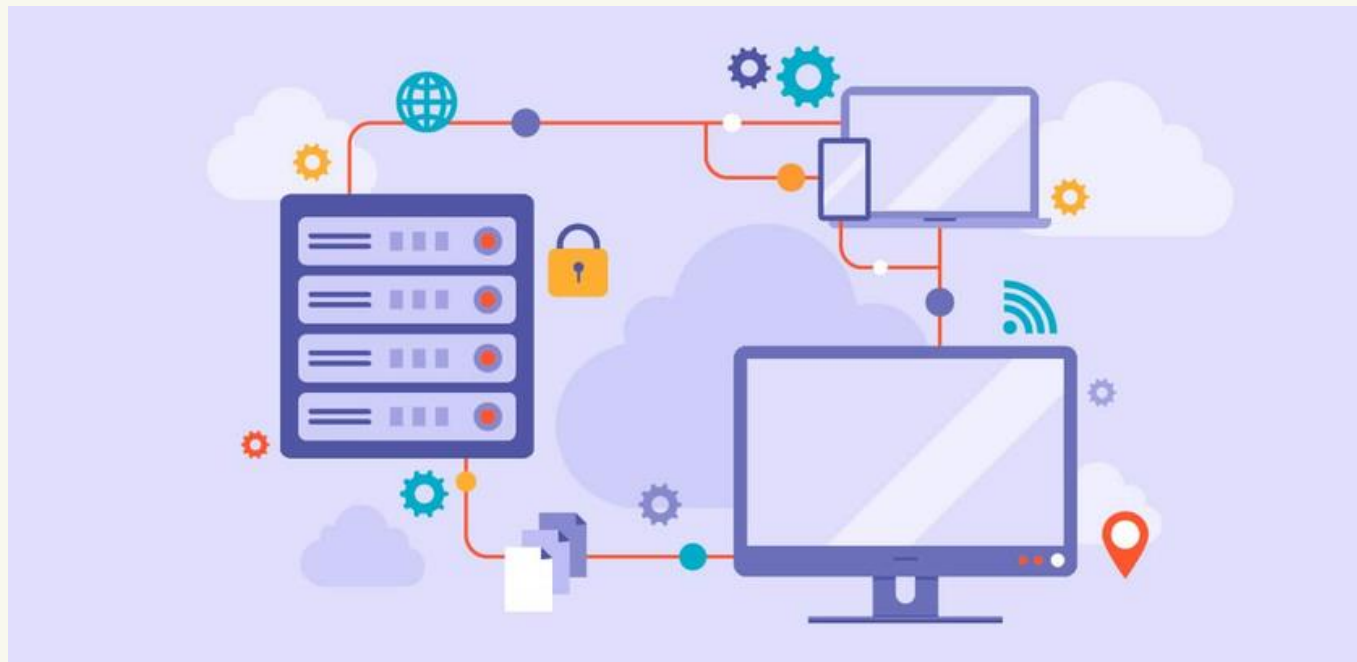


## 8. Model Evaluation



- Tahap untuk mengevaluasi kualitas model dan mengujinya apakah dapat mengatasi permasalahan bisnis dengan tepat.
- Evaluasi model dapat memiliki 2 fase yaitu:
  - **Diagnostic measures** digunakan untuk memastikan model bekerja dengan baik sesuai yang diharapkan
  - **Statistical significance testing** digunakan untuk memastikan bahwa data yang digunakan telah ditangani dan diinterpretasikan dengan benar di dalam model.

## 9. Deployment



Penerapan *machine learning* model untuk memecahkan masalah bisnis, setelah model output yang dikembangkan dirasa memuaskan dan mendapat persetujuan dari stakeholder



# 10. Feedback



- Setelah proses model *deployment*, perusahaan akan mendapatkan *feedback* atau umpan balik tentang kinerja model.
- Analisis *feedback* diperlukan *data scientist* untuk memperbaiki model serta meningkatkan akurasi dan kegunaannya.

CITIZEN DATA SCIENTIST

THANK YOU

 DigitalSkola