

# DATA SCIENCE



Jefina Tri Kumalasari, M.kom

Data science adalah bidang ilmu yang mempelajari cara mengumpulkan, mengolah, menganalisis, dan menginterpretasikan data untuk memperoleh informasi yang berguna. Dalam era digital saat ini, data menjadi sangat penting karena dapat membantu organisasi membuat keputusan yang lebih baik, meningkatkan efisiensi, dan menciptakan nilai tambah. Oleh karena itu, memahami data science menjadi sangat penting bagi siapa saja yang ingin berkarier di bidang ini atau ingin meningkatkan kemampuan analisis datanya.

Proses pengolahan data dalam data science melibatkan beberapa tahap, yaitu: pengumpulan data, pembersihan data, transformasi data, analisis data, dan visualisasi data. Pada tahap pengumpulan data, data dikumpulkan dari berbagai sumber seperti database, file, atau sensor. Kemudian, data dibersihkan untuk menghilangkan kesalahan atau ketidakakuratan. Hasil analisis data disajikan dalam bentuk visualisasi data untuk memudahkan interpretasi.

Tujuan utama dari data science adalah untuk mengubah data menjadi informasi yang berguna dan dapat ditindaklanjuti. Dengan memahami data science, kita dapat meningkatkan kemampuan analisis data, membuat keputusan yang lebih baik, dan menciptakan nilai tambah bagi organisasi. Selain itu, data science juga dapat membantu kita memahami fenomena yang kompleks, mengidentifikasi peluang baru, dan meningkatkan efisiensi operasional. Dengan demikian, memahami data science menjadi sangat penting bagi siapa saja yang ingin berkarier di bidang ini atau ingin meningkatkan kemampuan analisis datanya.

## Daftar isi

### Contents

BAB 1 .....	1
Pengenalan Data Science .....	1
1.1 Mengenal Data Science.....	1
1.2 Sejarah Data Science .....	2
1.3 Asal Data .....	4
1.4 Hubungan AI, Machine Learning Deep Learning dan Data Science .....	7
1.5. Data Science dalam industri.....	9
1.6 Tujuan Data science .....	11
Latihan Soal.....	11
BAB II .....	15
METODOLOGI DATA SCIENCE .....	15
2.1 Data Value Chain.....	15
2.2 Data Science Framework and Project Flow.....	16
2.2.1. Business Undertanding .....	16
2.2.2 Analitic Approach.....	17
2.2.3 Data Requirements .....	17
2.2.4 Data Collection.....	18
2.2.5 Data Undestanding .....	19
2.2.6 Data Preparation.....	19
2.2.7 Modeling .....	21
2.2.8 Evaluation .....	21
2.2.9 Deployment.....	22
Tugas: .....	22
Pertemuan 3 .....	26
Teknologi AI Berbasis Data .....	26
3.1 Perbedaan Artificial Intellegent dan Data Science.....	26
3.2 Lingkup dan Aplikasi.....	27
Analisis Foto Medis .....	28
3.3 Tools data science .....	28
Soal Latihan.....	30
BAB 4 .....	34

Sains data Dan Machine Learning .....	34
Perkembangan data:.....	35
4.1 Hubungan data science dan Machine learning.....	35
4.2 Kategori Machine Learning .....	36
Soal Latihan :.....	44
BAB 5 .....	49
Pemahaman Bisnis Proses Data .....	49
5.1 Data Science Lifecycle .....	49
5.2 Pemahaman Bisnis .....	49
Soal Latihan.....	51
BAB 6 .....	54
Tools Data Sains .....	54
6.1 INFRASTRUKTUR BIG DATA .....	55
6.2 Cloud Platforms.....	56
6.3 Jenis Penyimpanan Data .....	58
Soal Latihan.....	59
Pertemuan 7 .....	61
Teknik Pemrosesan Data.....	61
7.1 Data Preprocessing .....	61
7.2 Data Preparation.....	62
7.3 Tantangan Data Preparation.....	62
7.4 CRISP-DM .....	63
Soal Latihan.....	63
BAB 8 .....	66
8.1 Analisis dan Penafsiran data .....	66
8.2 Model Analisis Data .....	67
Soal Latihan.....	67
BAB 9 .....	71
Visualisasi Data .....	71
9.1 Perlunya Visualisasi Data .....	71
9.2 Manfaat Visualisasi Data:.....	72
9.3 Grafik.....	73
Soal Latihan.....	75
Bab 10 .....	78
Teknik Sampling .....	78

10.1 Big Data .....	78
10.2 Sampling Data .....	79
Metode Sampling.....	79
1. Probability Sampling (Sampling Probabilitas) .....	80
2. Non-Probability Sampling (Sampling Non-Probabilitas) .....	80
<b>Tahapan sampling .....</b>	<b>81</b>
Resampling.....	82
Latihan Soal.....	83
Daftar Pustaka.....	86

## BAB 1

### Pengenalan Data Science

#### 1.1 Mengenal Data Science

**Sebelum membahas tentang data science kita perlu mengenal data.**

Data menjadi sedemikian sentral dalam kehidupan modern. Pengembangan sains dan teknologi yang sedemikian revolusioner didasarkan pada data. Kegiatan bisnis dan ekonomi juga semakin mengandalkan ketersediaan data. Bahkan, dinamika sosial-politik serta budaya dan seni tidak terlepas dari data. Diplomasi dan negosiasi internasional makin bertumpu pada data.

Data yang dalam pemahaman umum adalah kumpulan fakta-fakta dan menjadi sumber informasi dan basis (ilmu) pengetahuan. Penguasaan dan pemilikan atas data selanjutnya menjadi ukuran kemampuan, sumber kekuatan, dan modalitas yang sangat penting untuk melakukan apapun atau untuk menjadi apapun. Sebagai kumpulan fakta, data dengan demikian tersebar, ada dimana-mana, sehingga sejarah peradaban umat manusia bisa disebut sebagai tumpukan atau akumulasi fakta. Fakta-fakta ini hanya menjadi (lebih) bermanfaat ketika berubah menjadi data dan selanjutnya menjadi informasi dan pengetahuan untuk menentukan pilihan-pilihan strategi dan keputusan.

Apa sih data science dan bagaimana data science akan menyelesaikan pemasalahan data yang besar dan tidak terbatas, perhitungan dan tools apa yang digunakan?

Data Science merupakan inti dari pertumbuhan bisnis modern saat ini mulai dari masalah kesehatan, pemerintahan hingga periklanan. Pengumpulan analisa ilmu data akan memberi potensi untuk meningkatkan kualitas, efektifitas, dan efisiensi hasil kerja guna keperluan profesional dan pribadi. Seseorang yang berkecimpung pada data science biasa disebut **data scientist**.

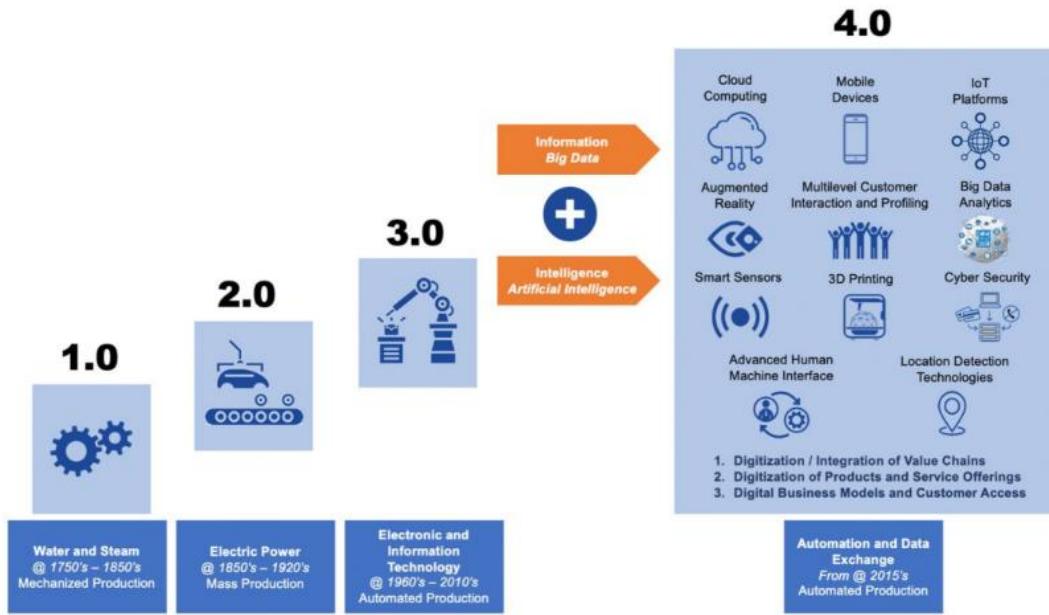
**Data science** adalah ilmu yang menggabungkan matematika, statisika dengan ilmu computer dengan tujuan analisa data (data analysis) dari suatu himpunan data baik skala kecil ( sampel) maupun besar (populasi ) dengan mengaplikasikan algoritma tertentu untuk tujuan menggal I data (data mining) dan mendapatkan pola data serta dapat melakukan prediksi data (prediction) dengan cukup akurat yang dapat membantu dalam pengambilan keputusan dan dapat digunakan untuk membuat system yang cerdas (AI) yang dapat terus belajar dengan sendirinya (machine learning).

## 1.2 Sejarah Data Science

<i>Tahun</i>	<i>Keterangan</i>
<b>1962</b>	John W. Tukey menulis dalam “The Future of Data Analysis” bahwa ia mengira dirinya adalah seorang ahli statistik, tetapi menyadari bahwa minat utamanya adalah pada analisis data
<b>1968</b>	Kongres IFIP (International Federation for Information. Processing) dengan tajuk “Datalogy, the Science of Data and Data Processes and Its Place in Education” Naur, menawarkan definisi data science, yaitu adalah Ilmu yang menangani data, begitu data ditetapkan, sedangkan “ hubungan data dengan apa yang diwakilinya didelegasikan ke bidang dan mu lain.
<b>1974</b>	<i>Peter Naur menerbitkan “Concise Survey of Computer Methods” di Swedia dan US. Buku ini merupakan survei metode pemrosesan data kontemporer yang digunakan dalam berbagai aplikasi.</i>
<b>1977</b>	<i>Tukey menerbitkan “Exploratory Data Analysis” yang menjelaskan pentingnya penggunaan data dalam memilih hipotesis untuk diri dan menjelaskan bahwa analisis data konfirmatori dan analisis data eksplorasi harus bekerja sama.</i>
<b>1989</b>	Howard Dresser mendeskripsikan Business Intellegence sebagai “concept and method to improve business decision making by using fact-based support systems.
<b>1994</b>	Memasuki dunia marketing, Business Week membuat berita mengenai database marketing yang mengungkapkan bahwa perusahaan telah mulai mengumpulkan sejumlah besar informasi konsumen.

1997	Mulai dikenal istilah machine learning. Mitchell, T.M
2002	<i>Committee on Data for Science and Technology</i> dari <i>Internal Council for Science</i> menerbitkan jurnal <i>data science</i> . Jurnal ini berfokus pada masalah-masalah seperti deskripsi sistem data, publikasinya di internet, aplikasi, dan masalah hukum.
2006	Sebuah <i>database</i> non-relasional <i>open source</i> dan berbasis Nutch bernama Hadoop 0.1.0 dirilis
2009	Alon Halevy, Peter Norvig, dan Fernando Pereira menuliskan <i>The Unresonable Effectiveness Of Data</i> , dimana algoritma dapat digunakan untuk membangun model data dengan baik
2010	Masalah Big Data muncul dengan fenomena <i>The Data Deluge</i> bagaimana data dapat diubah sehingga memberi manfaat pada bidang bisnis
2011	Lowongan <i>data scientist</i> meningkat sebesar 15%. Seminar dan konferensi <i>data science</i> dan <i>big data</i> juga ikut meningkat. Lalu, Sekitar 90% data di dunia sudah melalui tahap analisis.
2015	<i>Data science</i> makin berkembang dengan bantuan Artificial Intelligence (AI). Tercatat bahwa total proyek <i>software</i> Google yang menggunakan AI meningkat menjadi lebih dari 2.700 proyek. Selain itu, dengan menggunakan teknik <i>deep learning</i> , pengenalan Google <i>speech recognition</i> dan Google Voice mengalami peningkatan sebesar 49%.

#### Revolusi Industri Keempat

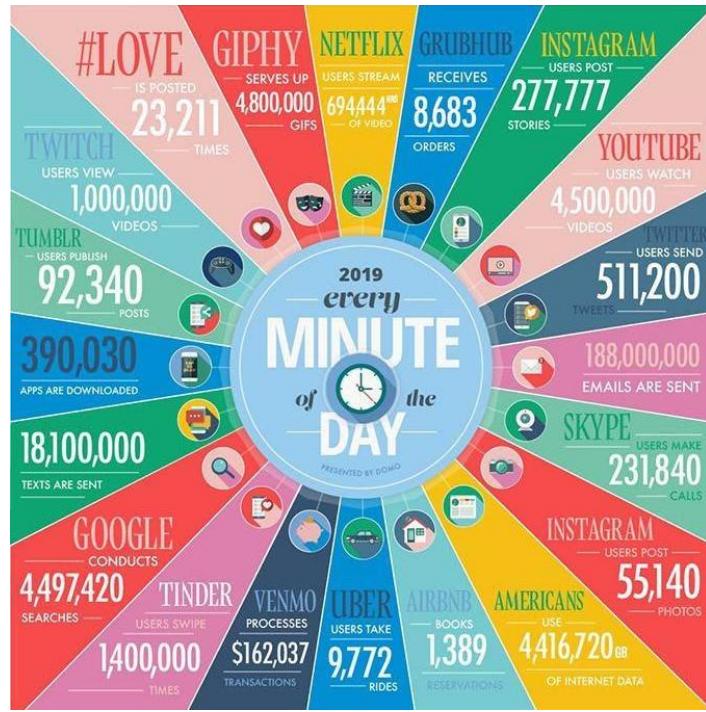


Dari Revolusi Industri pertama (mekanisasi melalui air dan tenaga uap), Industri 4.0 akan mengambil elektronik dan teknologi informasi yang dimulai pada Revolusi Industri 3.0 dan membawanya ke tingkat yang lebih tinggi berkat data besar dan kecerdasan buatan (*Machine Learning*).

### 1.3 Asal Data

Banyaknya jumlah data yang dapat diambil dari berbagai kegiatan. Dalam setiap menit dari penjuru dunia akan mengumpulkan berbagai macam data baik berasal dari media social atau informasi domestic suatu negara.

Data berasal dari berbagai aktifitas.



Sumber : Forbes.com

### Istilah dalam Big Data

Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it. Atau Big Data adalah data yang melebihi dari kapasitas pemrosesan sistem database konvensional. Data yang terlalu besar dan terlalu cepat atau tidak sesuai dengan struktur arsitektur database yang ada. Untuk mendapatkan nilai dari data, maka anda harus memilih jalan alternatif untuk memproses data data tersebut.

Pendefinisian big data sendiri pada umumnya dibagi menjadi 4 bagian yaitu:

### Volume

Mengacu pada sejumlah big data yang dihasilkan setiap detiknya Artinya sekumpulan data dalam jumlah dan volume yang sangat besar dan kadang tidak terstruktur Contohnya feed Twitter, feed

Istagram data teks chat dan status Whatsapp alur klik user dari halaman web Arus data data tersebut bisa berukuran hingga ribuan Terrabyte ( per detiknya)

### **Velocity**

Data dapat diakses dengan kecepatan yang sangat cepat sehingga dapat langsung digunakan pada detik itu juga lebih real time) Salah satu buktinya antara lain, adanya sistem operasi online berbasis Microsoft Silverlight, aplikasi perkantoran office berbasis web seperti Office 365 cloud storage seperti Dropbox dan GDrive

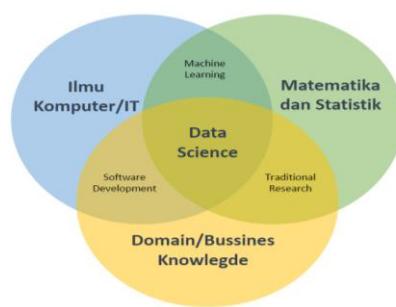
### **Veracity**

Big data memiliki kerentanan dari sisi keakuratan dan kevaliditasan sehingga memerlukan kedalaman untuk menganalisis big data agar bisa menghasilkan keputusan yang tepat Karakter veracity mengarah kepada seberapa akurat dan dapat dipercaya suatu data

### **Value**

Value berarti big data memiliki nilai yang sangat tinggi apabila diolah dengan cara yang tepat guna atau dapat juga dikatakan seberapa bernilainya atau bermaknanya suatu data Contohnya biodata karyawan suatu perusahaan penjualan bahan baku makanan tidak akan bernilai untuk kepentingan analisis prediksi penjualan bahan baku ke customer Data tersebut mungkin tidak penting dan tidak bernilai untuk satu hal namun bisa sangat penting dan sangat bernilai untuk hal lain Data yang tidak memiliki nilai di bagian mana pun tidak akan terfilter di sistem aplikasi analisis Big data

Irisan Bidang Ilmu dalam data Science

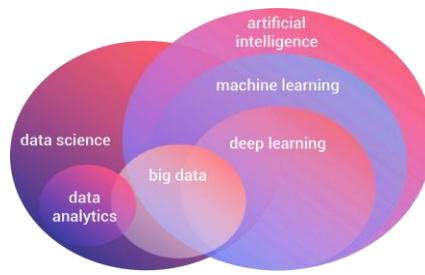


Irisan Bidang Ilmu dalam data Science

Kemampuan yang dibutuhkan Data Science

1. Matematika dan Statistika.
2. Pemrograman (R, Python, dan lainnya)
3. Database dan Query (SQL dan lainnya) dan pengolahan data.
4. Analisa data dan visualisasi data.
5. Pemahaman masalah terkait bisnis atau suatu bidang lainnya

#### **1.4 Hubungan AI, Machine Learning Deep Learning dan Data Science**



##### **1. Artificial intelligence**

Kecerdasan buatan digunakan untuk menggambarkan komputer yang dapat mensimulasikan kemampuan dan perilaku berpikir manusia.

##### **2. Mechine Learning**

Mechine Learning menggunakan kecerdasan buatan untuk memungkinkan mesin mempelajari dan memprediksi hasil dengan lebih akurat tanpa diprogram secara eksplisit untuk

melakukannya. Dimana sebuah algortima dalam computer yang mampu belajar secara otomatis berdasarkan pengalaman

### **3. Deep Learning**

Merupakan salah satu bagian dari machine learning, yang berbasis pada artificial neural network, dengan cara kerja mengikuti pola saraf otak manusia.

Deep learning menggunakan data untuk meningkatkan kemampuan mesin untuk mengklasifikasikan, mengenali, dan mendeteksi dimana akan membuat model lebih akurat.

### **4. Data Sience**

Merupakan sebuah bidang lintas disiplin ilmu yang didalamnya metodologi scientific, process, algoritma dan system yang mengekstraksi pengetahuan untuk mencari pengetahuan baru dari sebuah data baik terstruktur maupun tidak terstruktur. Data Science berhubungan dengan data mining, deep learning dan big data

### **5. Big Data**

"big data" mengacu pada data yang sangat besar, cepat atau kompleks sehingga sulit atau tidak mungkin untuk diproses menggunakan metode tradisional. **Alasan Data Science Menarik**

**Lalu apa perbedaan data science dengan data mining?**

Data mining adalah proses pengambilan informasi dari pola data yang berasal dari himpunan data sebelumnya tidak diketahui, biasa disebut juga data discovery. Data mining berfokus pada mengekstrak pola menggunakan metode statistic untuk menganalisa dan juga memprediksi.

Machine learning merupakan bagian dari AI yang digunakan agar system computer secara otomatis dapat belajar dengan sendirinya.

## 1.5. Data Science dalam industri

*Data science* adalah ilmu yang saat ini sangat populer dan dapat diterapkan di berbagai industry, sebagai:

### 1. Kompetensi Wajib di Zaman Digital

Banyak perusahaan yang mulai menyadari betapa pentingnya data untuk membuat keputusan.

Dari data yang ada, kita



a bisa

menemukan insight yang menarik dan bisa dijadikan bahan pertimbangan untuk membuat keputusan. Secara umum, data bisa dikelompokkan menjadi 3 kelompok besar, yaitu data terstruktur, data semi terstruktur, serta data yang tidak terstruktur.

### 2. Dibutuhkan Hampir Semua Sektor

## DOMAIN EXPERIENCES

Data Analytics Use Cases with Related Domain Experiences

data crawling, data warehousing, data visualization & collaboration analysis	SMART CITY			TOURISM & HOSPITALITY	visitor profiling, counting & appearance (face mask detection) sentiment & aspect analysis
tracing suspect, morbidity, early warning, healthcare facility occupancy	HEALTHCARE			RETAIL & LOGISTICS	visitor demography, behaviour, counting & appearance, face mask detection
speed estimation, counting & classification, traffic modelling, license plate recognition, law enforcement	ROAD TRAFFIC			BANKING & FINTECH	fraud detection, log management system / security incident & event management, customer behavior
residential & commercial area analysis, human daily activity (HDA) analysis, people mobility analysis	PUBLIC HOUSING			HUMAN RESOURCES	intolerant detection, lifestyle analysis, employee engagement, assessment profiling
product intelligence, cyber patrol, local business analysis, product segmentation	E-COMMERCE			OPERATIONS MANAGEMENT	performance evaluation, efficiency analysis, predictive maintenance
sentiment analysis, influencer & people network analysis, campaign effectiveness analysis, e-clipping	SOCIAL & ONLINE MEDIA			MARKET SEGMENTATION	customer behavior, market trend analysis, market segmentation analysis

### 3. Membantu Pengambilan Keputusan (*Decision-Making*)

*Data Science* membantu organisasi dalam mengambil keputusan berdasarkan data yang ada. Dengan analisis data yang akurat dan komprehensif, organisasi dapat membuat keputusan yang lebih baik dan lebih cepat.

### 4. Meningkatkan Efisiensi

*Data Science* dapat membantu organisasi dalam mengidentifikasi pola-pola dalam data dan mengoptimalkan proses bisnis untuk meningkatkan efisiensi dan produktivitas.

### 5. Meningkatkan keuntungan

Dengan memahami perilaku pelanggan dan tren pasar, *Data Science* dapat membantu organisasi dalam mengidentifikasi peluang bisnis yang baru dan pada akhirnya bisa meningkatkan keuntungan.

### 6. Meningkatkan pengalaman pelanggan.

Dengan memahami perilaku pelanggan dan preferensi mereka, *Data Science* dapat membantu organisasi dalam meningkatkan pengalaman pelanggan dan memberikan layanan yang lebih baik.

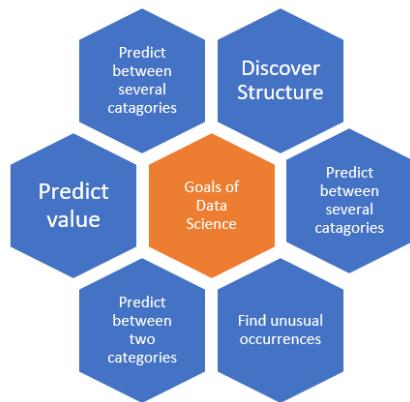
### 7. Mengurangi risiko, fraud, dan eror.

*Data Science* dapat membantu organisasi dalam mengidentifikasi risiko bisnis dan membuat keputusan yang lebih cerdas dalam mengatasinya.

risiko tersebut. Selain itu, ia juga bisa digunakan untuk mendeteksi fraud dan meminimalisir eror.

8. **Meningkatkan kualitas produk dan layanan.** Dengan menganalisis data kualitas produk dan layanan, Data Science dapat membantu organisasi dalam meningkatkan kualitas produk dan layanan tersebut.
9. **Mendukung penelitian dan pengembangan.** Data Science dapat membantu penelitian dan pengembangan dalam mengidentifikasi tren dan pola dalam data yang kompleks.
10. **Meningkatkan efektivitas pemasaran.** Data Science dapat membantu organisasi dalam memahami perilaku pelanggan dan kecenderungan pasar untuk membuat strategi pemasaran yang lebih efektif dan efisien.
11. **Meningkatkan keamanan.** Data Science dapat membantu organisasi dalam mengidentifikasi ancaman keamanan dan mengambil langkah-langkah untuk mengurangi risiko dan melindungi data sensitif.
12. **Menjaga daya saing.** Data Science dapat membantu organisasi dalam mempertahankan daya saing mereka dengan memahami tren dan perubahan dalam pasar dan industri.

## 1.6 Tujuan Data science



Dengan menggunakan algoritma membuat prediksi menggunakan data tidak terstruktur, semi terstruktur dan data terstruktur untuk membangun model / trend data sehingga individu maupun organisasi membantu membuat keputusan yang baik di masa akan datang.

### Latihan Soal

1. Apa yang menjadi awal mula perkembangan konsep data science?
  - a. Abad ke-21
  - b. Abad ke-20

- c. Abad ke-19
  - d. Abad pertengahan
  - e. Abad kuno
2. Siapakah yang mendeskripsikan Business Intelligence sebagai dasar perusahaan membuat keputusan
- a. Howard Dresser
  - b. Isaac Newton
  - c. Halevy
  - d. Alan Turing
  - e. Aristotles
3. Manakah yang bukan merupakan salah satu langkah dalam proses data science?
- a. Ekstraksi data
  - b. Visualisasi data
  - c. Penelitian eksperimental
  - d. Interpretasi data
  - e. Pemodelan data
4. Data science bertujuan untuk melakukan apa terhadap data?
- a. Mengumpulkan dan menyimpannya
  - b. Menganalisis dan mengekstraksi informasi
  - c. Menyembunyikan data sensitif
  - d. Menjual data kepada perusahaan lain
  - e. Membuat data menjadi tidak terbaca
5. Apa yang dimaksud dengan "Big Data" dalam konteks data science?
- a. Data yang memiliki ukuran yang kecil
  - b. Data yang diproses secara manual
  - c. Data yang terdiri dari berbagai jenis

- d. Data yang diproduksi dengan cepat
  - e. Data yang memiliki volume besar, kecepatan, dan variasi
6. Perusahaan mana yang pertama kali menggunakan istilah "Data Scientist" secara resmi?
- a. Google
  - b. Facebook
  - c. IBM
  - d. Amazon
  - e. Microsoft
7. Algoritma apa yang umumnya digunakan untuk pembelajaran mesin dalam data science?
- a. Algoritma matematika
  - b. Algoritma pembelajaran manusia
  - c. Algoritma berbasis aturan
  - d. Algoritma pembelajaran mesin
  - e. Algoritma perhitungan statistik
8. Apa tujuan utama dari visualisasi data dalam data science?
- a. Memperkecil data
  - b. Memperjelas pola dan tren dalam data
  - c. Membuat data tidak bisa diakses
  - d. Mempercepat proses analisis data
  - e. Menyembunyikan informasi dalam data
9. Seseorang yang berkecimpung dalam dalam data science disebut?
- a. Programmer
  - b. Data scientist
  - c. Data mining
  - d. Manager data

e. Data visualization

10. Apa yang dimaksud dengan "Descriptive Analytics" dalam data science?

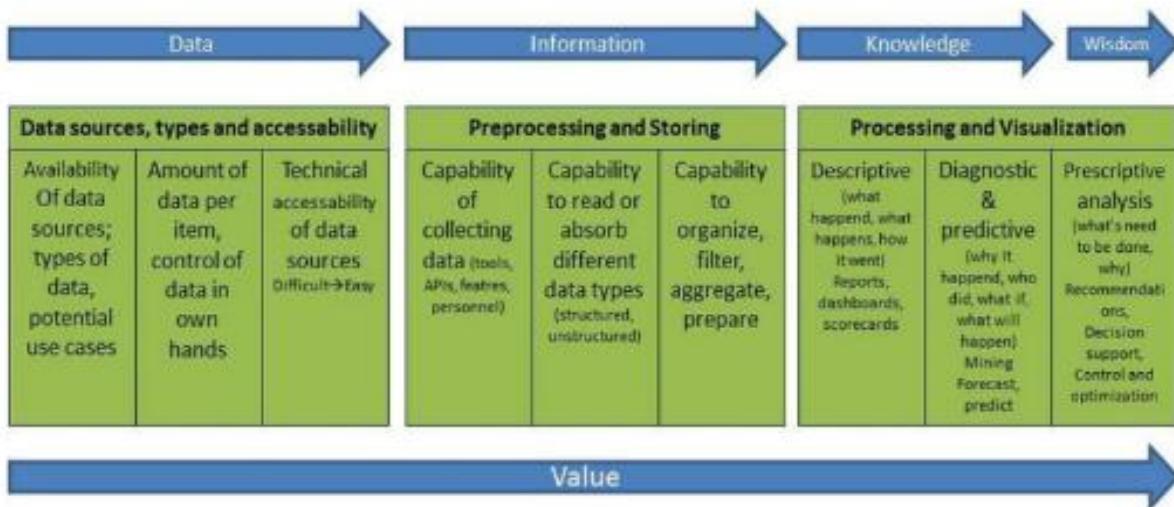
- a. Menganalisis masa lalu
- b. Memprediksi masa depan
- c. Menggambarkan karakteristik data
- d. Mengoptimalkan hasil
- e. Mengembangkan strategi bisnis

## BAB II

### METODOLOGI DATA SCIENCE

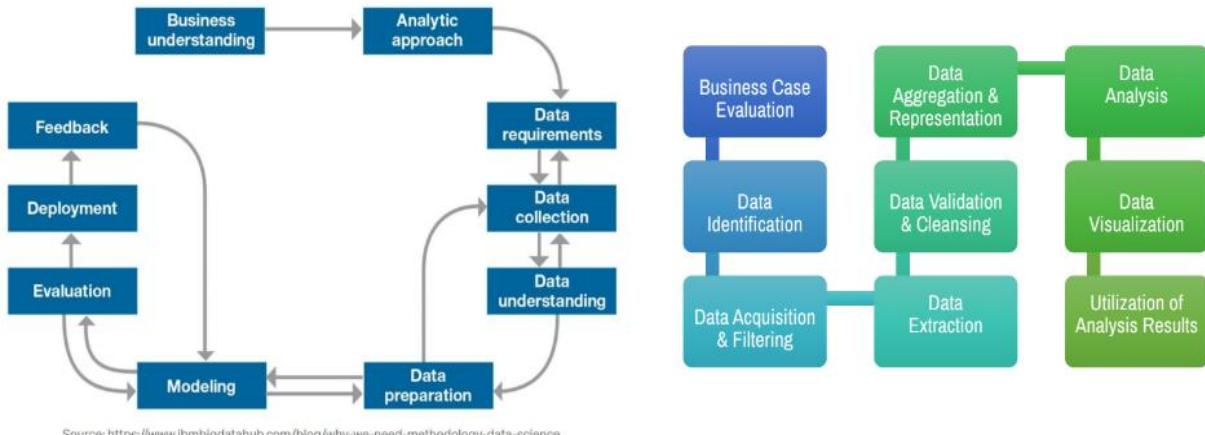
Metodologi data science adalah langkah-langkah digunakan dalam proyek data science agar dapat menghasilkan hasil yang optimal yang dapat menjawab pertanyaan dari suatu masalah yang ingin diselesaikan. Metodologi

#### 2.1 Data Value Chain



Rantai nilai atau value chain adalah strategi untuk memandang suatu bisnis dan dipandang sebagai rangkaian kegiatan yang mengubah input menjadi output yang bernilai bagi pelanggan. Konsep ini menyediakan suatu kerangka kerja untuk memvisualisasikan bagaimana perusahaan dapat menambah nilai.

## 2.2 Data Science Framework and Project Flow



### 2.2.1. Business Understanding

Fase ini mendefinisikan masalah tujuan dan persyaratan solusi dari perspektif bisnis. Berdasarkan pemahaman bisnis sebelumnya, kita harus memutuskan pendekatan analitis mana yang harus diikuti, yaitu:

- Deskriptif** → status saat ini dan informasi yang diberikan.
- Diagnostik** → analisis statistik, apa yang terjadi dan mengapa itu terjadi.
- Prediktif** → meramalkan tren atau kemungkinan kejadian di masa depan.
- Preskriptif** → bagaimana masalah harus diselesaikan

Contoh:

perusahaan asuransi ingin menggunakan data science untuk menyelesaikan masalah katakanlah pertanyaannya Bagaimana cara terbaik untuk mengalokasikan dana kesehatan yang terbatas agar dapat memaksimalkan penggunaannya dalam memberikan layanan yang berkualitas Sebelum memulai mengumpulkan data, target dan tujuan dari pertanyaan tersebut perlu didefinisikan terlebih dahulu Kita memerlukan penjelasan dari si pemberi pertanyaan untuk dalam mengetahui lebih detail target dan tujuannya Misalnya dalam kasus ini targetnya adalah menyediakan layanan kesehatan tanpa menaikkan biaya sedangkan

tujuannya adalah meninjau kembali proses yang sudah berjalan untuk mengidentifikasi ketidakefektifan inefficiencies.

Setelah target dan tujuan ditentukan misalnya tim data scientist memprioritaskan perawatan kembali pasien sebagai area yang efektif untuk ditinjau ulang Dengan bekal target dan tujuan yang sudah ditentukan ditemukan bahwa 25-35 pasien yang telah selesai menjalani perawatan akan kembali menjalani perawatan dalam waktu satu tahun sementara 50 pasien akan kembali menjalani perawatan dalam waktu lima tahun Dan pasien gagal jantung merupakan pasien terbanyak yang kembali menjalani perawatan.

### **2.2.2 Analitic Approach**

Tahap metodologi data science selanjutnya yang dilakukan adalah menentukan pendekatan analitik untuk menyelesaikan masalah. Dalam tahap ini dilakukan pendefinisian masalah dalam konteks statistik atau *machine learning* untuk memperoleh hasil yang diinginkan.

### **2.2.3 Data Requirements**

Metode analisis yang telah dipilih sebelumnya menunjukkan isi, format, dan sumber data yang diperlukan untuk dikumpulkan. Selama proses kebutuhan data, kita harus menemukan jawaban atas pertanyaan-pertanyaan seperti apa, dimana, kapan, mengapa, bagaimana, siapa. Contoh: dalam proses diatas diketahui:



masalah diatas adalah bagaimana membuat masakan yang enak dengan data bahan-bahan makanan. Untuk menyelesaikan masalah kita membutuhkan prosedur yaitu resep. maka yang

perlu kita identifikasi adalah data apa saja yang diperlukan, bagaimana mengumpulkan data tersebut, bagaimana mengolah data tersebut, dan bagaimana menyiapkan data tersebut agar sesuai dengan hasil yang diinginkan.

## 2.2.4 Data Collection

proses pengumpulan, pengukuran, dan analisis berbagai tipe informasi menggunakan teknik berstandar. Tujuan utama *data collection* adalah untuk mengumpulkan informasi dan data terpercaya sebanyak-banyaknya, yang kemudian dianalisis untuk membuat sebuah keputusan bisnis yang krusial. Ketika sudah berhasil dikumpulkan, data ini kemudian melalui sejumlah proses meliputi pembersihan dan pemrosesan data agar dapat digunakan oleh perusahaan. Organisasi mengumpulkan data menggunakan berbagai metode untuk membuat keputusan yang lebih baik. Tanpa data, akan sulit bagi organisasi untuk mengambil keputusan yang tepat, sehingga data dikumpulkan dari audiens yang berbeda pada waktu yang berbeda-beda



### Jenis Data

Generally, data collection can yield two kinds of data: qualitative and quantitative.

- **Qualitative data** refers to data that describes the characteristics, qualities, and other non-quantifiable traits of a certain subject. This includes personal opinions, descriptions of a certain place, event, or behavior, or the quality of a certain item. Qualitative data is often hard to measure with numbers, and so they are analyzed based on their qualities or patterns.

- **Quantitative data**, on the other hand, refers to quantifiable or countable data such as statistics, the number of respondents or test subjects, and those under certain standards of measurement such as temperature.

### 2.2.5 Data Understanding

Pada tahap ke lima dari metodologi data science ini, kita akan mengecek apakah ada *missing values*, data yang *imbalanced*, *outlier*, salah format, dan sebagainya yang harus diperbaiki terlebih dahulu.

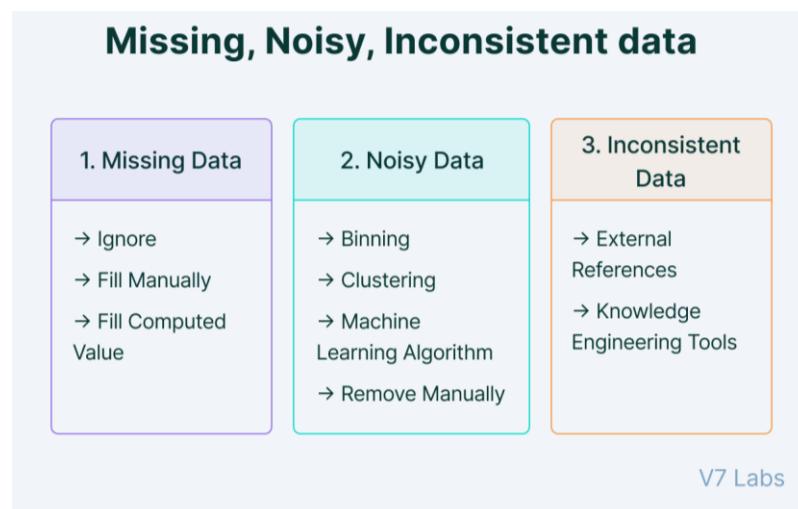
Proses data *understanding* yang populer adalah dengan menggunakan statistik deskriptif dan teknik visualisasi. Teknik ini membantu *data scientist* memahami isi data, menilai kualitas data, dan menemukan *insight* awal dari data tersebut.

### 2.2.6 Data Preparation

Pada tahapan ini dilakukan pembersihan data, menggabungkan data, dan mengubah data menjadi variabel yang lebih berguna. Agar data dapat diproses secara efektif pada tahap pemodelan, data harus dipersiapkan dengan baik.

data Cleaning, dengan membersihkannya dari *missing values*, *invalid values*, dan data duplikat construct data, memastikan bahwa seluruh data telah memiliki format yang benar.

#### 1. Data Cleaning



#### 2. Data Integration

Data integration adalah langkah preprocessing untuk menggabungkan data dari berbagai sumber menjadi satu basis data besar, seperti data warehouse. Ini penting terutama dalam kasus nyata, seperti mendeteksi nodul dari citra CT Scan yang memerlukan penggabungan data dari berbagai sumber medis

Dalam proses data integration, beberapa masalah yang mungkin muncul antara lain: perbedaan skema dan pencocokan objek, adanya atribut yang redundant, serta konflik nilai data yang perlu dideteksi dan diselesaikan.

### **3. Normalization**

Normalisasi adalah teknik transformasi data yang penting, di mana atribut numerik diskalakan ke dalam rentang tertentu untuk membangun korelasi antar data.

### **4. Attribute Selection**

Pembuatan atribut baru dari data yang sudah ada membantu proses data mining. Contohnya, atribut tanggal lahir dapat diubah menjadi is\_senior\_citizen untuk mendukung prediksi seperti risiko penyakit atau peluang bertahan hidup.

### **5. Data Reduction**

Ukuran dataset dalam data warehouse bisa terlalu besar untuk dianalisis secara langsung, sehingga diperlukan strategi reduksi data untuk mempermudah proses analisis dan data mining

- 1. Data cube aggregation,** Reduksi data dapat dilakukan dengan menyajikan data dalam bentuk ringkasan

#### **2. Dimensionality reduction**

Teknik reduksi dimensi digunakan untuk mengekstraksi fitur dengan mengurangi jumlah atribut yang redundant dalam data, sehingga mempermudah algoritma machine learning. Salah satu teknik yang digunakan adalah Principal Component Analysis (PCA)

#### **3. Data compression**

Menggunakan teknologi encoding akan mengurangi ukuran data.

#### **4. Discretization**

Discretisasi data digunakan untuk membagi atribut kontinu menjadi interval karena fitur kontinu cenderung memiliki korelasi rendah dengan variabel target. Contohnya, atribut usia dapat dibagi menjadi kelompok seperti di bawah 18, 18–44, 44–60, dan di atas 60.

#### **5. Numerosity reduction**

Pemilihan atribut yang tepat sangat penting dalam pemodelan data, seperti regresi. Atribut yang tidak relevan sebaiknya dihapus untuk menghindari data berdimensi tinggi yang dapat menyebabkan masalah underfitting atau overfitting..

#### ***Feature engineering***

*Feature engineering* adalah proses transformasi data menjadi fitur-fitur yang lebih representatif dan membantu menyelesaikan masalah dengan lebih baik. Tahapan ini memakan banyak waktu. Tahap ini bisa menghabiskan sekitar 70% atau bahkan 80% dari keseluruhan proses dalam projek *data science*.

#### **2.2.7 Modeling**

*Modeling* atau pemodelan adalah tahap dalam metodologi *data science* dimana *data scientist* membuat model untuk menjawab permasalahan. Pemodelan data berfokus pada mengembangkan model, baik itu model deskriptif atau prediktif. Model ini bergantung pada *analytical approach* yang telah ditentukan sebelumnya, apakah menggunakan pendekatan statistik atau *machine learning*. Proses pemodelan untuk model prediktif menggunakan data training.

Tahap untuk membuat model menggunakan Machine Learning/Deep Learning atau metode data mining lainnya. Memformulasikan data ke dalam model yang tepat (*fit data*)

#### **2.2.8 Evaluation**

Pengujian terhadap kualitas model apakah model yang dirancang sebelumnya tersebut dapat mengatasi permasalahan bisnis dengan tepat. Evaluasi model memiliki dua fase yaitu:

1. Fase *diagnostic measures*

*Diagnostic measures* digunakan untuk memastikan model bekerja dengan baik sesuai yang diharapkan.

## 2. *Statistical significance*

*Testing* dapat digunakan untuk memastikan bahwa data yang digunakan telah ditangani dan diinterpretasikan dengan benar di dalam model.

### 2.2.9 Deployment

Setelah mengasumsikan bahwa model yang dikembangkan menghasilkan hasil yang memuaskan dan disetujui oleh pemangku kepentingan, model tersebut dapat diterapkan atau digunakan dalam lingkungan bisnis.

- Deploy plan
- Monitoring and maintain plan
- Produce final report
- Review project

Menghadirkan model yang telah dibuat sebagai solusi atas permasalahan bisnis Dapat berupa penerapan machine learning atau cukup dengan generate report Penerapan strategi dan kebijakan sesuai hasil data mining yang telah dilakukan.

#### Tugas:

Budi seorang manager cabang di sebuah hotel. Atasannya meminta Budi untuk mengobservasi review tamu yang mengunjungi hotelnya. Tujuannya untuk meningkatkan performa hotel dari review yang diberikan oleh tamu, menganalisa competitor dari review, pelayanan apa yang sering mendapat komplain. Budi mengoleksi review dari online travel agent yang telah bekerjasama dengan hotel dan digunakan oleh tamu.

Buatlah modelling dari kasus berikut

Jawaban :

Bussinis understanding

Memetakan Online Travel Agent (OTA) yang bekerja sama dengan hotel dan dipakai oleh tamu

Memetakan competitor hotel bedasarkan kelas hotel yang berada disekitar lokasi hotel  
Memanfaatkan data media digital untuk mempercepat dan mempermudah proses analisis kompetitor, sehingga dapat merespon dan menentukan strategi pemasaran yang cepat dan tepat

### **Data Understanding**

Data apa saja yang disajikan oleh Online Travel Agent

- Review hotel
- Nama tamu
- Harga

### **Cara mengumpulkan data tersebut:**

- Crawling
- Call API
- Download data dari OTA

### **Mendeskripsikan data yang telah diperoleh**

- Data mana yang penting untuk dikumpulkan → Melakukan eksploring terhadap data
- Data mana yang relevan dengan proyek ini → Melakukan verifikasi apakah data tersebut layak untuk digunakan

### ***Data***

Data preparation, Setelah data terkumpul ternyata data tersebut perlu dirapikan, *Melakukan observasi:*

- *Clean data*
- *Missing value*
- *Inconsistent*
- *Outlier*

### ***Modelling***

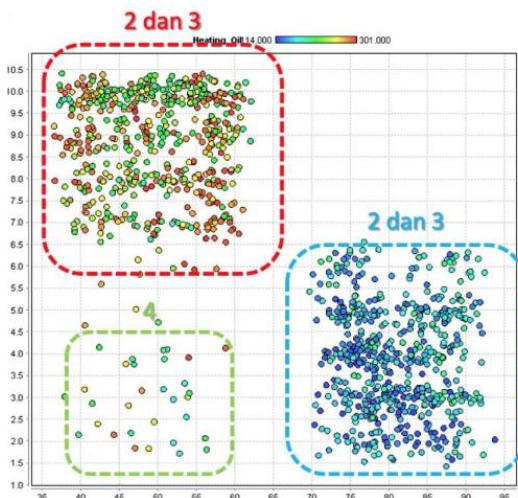
*Data akan diklasifikasikan atau diestimasi atau diklastering?*

```
In [18]:  
from sklearn.model_selection import train_test_split  
from sklearn.feature_extraction.text import TfidfVectorizer  
from sklearn.naive_bayes import MultinomialNB  
from sklearn.svm import SVC, LinearSVC  
from sklearn.metrics import accuracy_score  
from sklearn.metrics import confusion_matrix  
from sklearn.linear_model import LogisticRegression
```

```
In [20]:  
import joblib  
  
cols = df.iloc[:,3:12].columns  
  
for col in cols:  
    X = df['cleaned']  
    y = y1[col]  
    print(col)  
  
    # vectorizer  
    vectorizer = TfidfVectorizer()  
    X_vect = vectorizer.fit_transform(X)  
    joblib.dump(vectorizer, "../models/tfidf-%s"%col)  
  
    # classification  
    clf = LinearSVC()  
    clf.fit(X_vect, y)  
    joblib.dump(clf, "../models/clf-%s"%col)
```

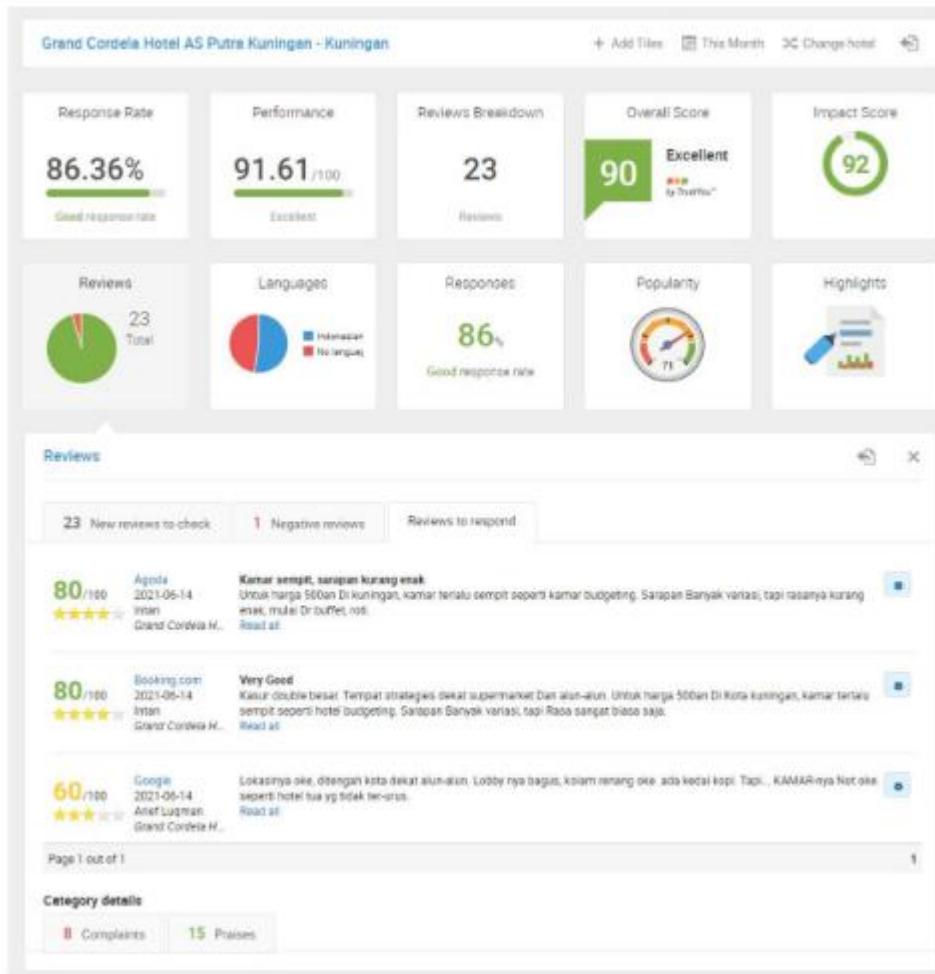
## Evaluasi

**Data yang dievaluasi untuk mengetahui pola yang ada**



## Deployment

Membuat dashboard agar dapat diakses oleh atasan dengan mudah



## Feedback

Umpulan balik dari stackholder tentang kinerja model yang telah dibangun oleh data scientist, sehingga dapat meningkatkan akurasi dan kegunaan modelnya

## Pertemuan 3

### Teknologi AI Berbasis Data

#### 3.1 Perbedaan Artificial Intellegent dan Data Science

Baik AI dan Data Science adalah pilihan yang populer saat ini. Perhatikan dan pahami beberapa perbedaan berikut :

	Data Science	Artificial Intelligence
<b>Arti</b>	Data Science adalah teknik kurasi data massa untuk visualisasi dan analitik.	Artificial Intelligence adalah seni menerapkan data di mesin.
<b>Ketrampilan</b>	Ini adalah teknik statistik desain dan pengembangan	Ini adalah teknik algoritma desain dan pengembangan
<b>Teknik</b>	Ilmu data dianggap sebagai teknik analisis data.	Kecerdasan buatan dikenal sebagai teknik pembelajaran mesin
<b>Penggunaan</b>	Teknologi Data Science menggunakan pembelajaran statistik untuk analisis	Kecerdasan buatan menggunakan pengetahuan machine learning
<b>Pengamatan</b>	Ini mengamati pola dalam data untuk keputusan	Ini menggunakan kecerdasan dalam data untuk keputusan
<b>Pemecahan</b>	Teknologi ini cenderung memanfaatkan berbagai bagian loop untuk memecahkan suatu masalah tertentu	Teknologi AI merepresentasikan loop persepsi dan perencanaan dengan tindakan
<b>Pengolahan Data</b>	Data sains adalah pengolahan data tingkat menengah untuk manipulasi data	Artificial intelligence tingkat tinggi pengolahan data ilmiah untuk manipulasi

<b>Grafis</b>	Teknologi Data Science terutama melibatkan representasi data dalam format grafis yang berbeda.	Teknologi kecerdasan buatan melibatkan representasi node jaringan algoritma
<b>Kontrol</b>	Data Science adalah tentang kontrol dan manipulasi data dengan berbagai teknik.	Kecerdasan buatan adalah tentang kontrol dengan AI dan teknik machine learning

Sehingga dapat di simpulkan :

**Data Science** : mencakup teknik analisis data, penggunaan algoritma machine learning dan pemodelan prediktif. Dimana berfokus membangun model prediktif, memahami dan mengekstrak nilai dari data melalui analisis, dan mendukung pengambilan keputusan berbasis data, biasanya menggunakan metode statistic.

**Artificial Intelligence (AI)** : AI mengadopsi ruang lingkup belajar, menalar dan memecahkan masalah. Sehingga berdasarkan ketiga hal itu, membuat AI berfokus pada meniru kecerdasan manusia untuk memecahkan masalah atau tugas yang diberikan kepada sistemnya.

### 3.2 Lingkup dan Aplikasi

*Data science* mencakup beragam aplikasi yang memberikan dampak signifikan, terutama dalam konteks bisnis dan ilmu pengetahuan. Dalam bisnis, *data science* digunakan untuk prediksi penjualan guna membantu perusahaan merencanakan persediaan dan strategi pemasaran. Analisis pasar menggunakan teknik *data science* untuk memahami perilaku konsumen dan tren industri. Segmentasi pelanggan juga menjadi area penting, memungkinkan personalisasi layanan dan penawaran.

Di ilmu pengetahuan, *data science* mendukung penelitian dengan menyederhanakan analisis data eksperimental, mempercepat proses penemuan, dan membantu memahami pola yang

kompleks dalam data ilmiah. Dalam konteks kedokteran, misalnya, *data science* dapat digunakan untuk menganalisis pola genetik atau meramalkan hasil klinis.

### **Dalam Bidang E-commerce**

Memberikan rekomendasi produk yang sesuai dengan kebutuhan berdasarkan data history pencarian atau pembelian barang sebelumnya

Contoh data yang dapat digunakan untuk Analisa data

<https://www.kaggle.com/ekrembayar/apriori-association-rules-grocery-store>.

### **Dalam Bidang Industri**

Melakukan prediksi harga komoditas dan permintaan komoditas contoh data yang dapat digunakan

<https://tradingeconomics.com/forecast/commodity>

### **Analisis Foto Medis**

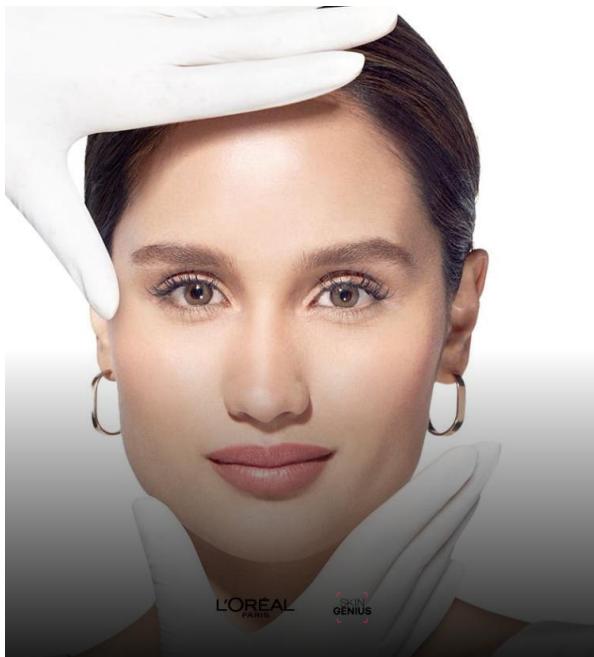
Sektor kesehatan mendapatkan manfaat pencitraan medis dengan BioMed Research. Penelitian tersebut, teknik pencitraan yang populer termasuk magnetic resonance imaging (MRI), sinar-X, computed tomography, mamografi, dan sebagainya. Berbagai metode digunakan untuk mengatasi perbedaan dalam modalitas, resolusi, dan dimensi gambar-gambar ini. Banyak lagi yang sedang dikembangkan untuk meningkatkan kualitas gambar, mengekstrak data dari gambar dengan lebih efisien, dan memberikan interpretasi yang paling akurat. Algoritme berbasis pembelajaran mendalam meningkatkan akurasi diagnostik dengan belajar dari contoh sebelumnya dan kemudian menyarankan solusi perawatan yang lebih baik.

### **3.3 Tools data science**

1. Pyton
2. R
3. SQL
4. Jupiter Notebook
5. Tableau
6. Hadoop
7. Azure dll

## Aplikasi Data Science

1. L'oreal Paris memanfaatkan data science dan AI untuk pemeriksaan kulit dengan memanfaatkan kamera handphone



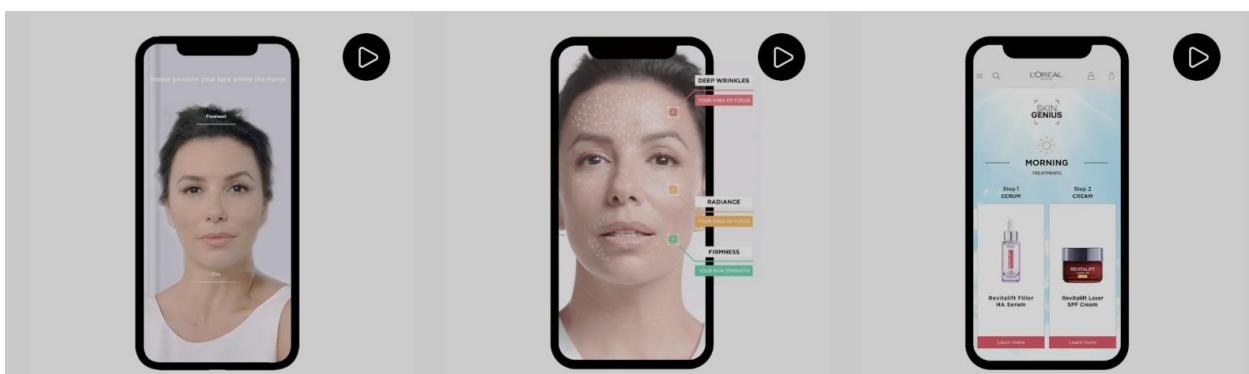
### Analisa kulit yang akurat hanya dengan satu selfie

Skin Genius adalah alat diagnosa kulit berbasis teknologi AI yang menganalisa kulitmu dan memberikan rekomendasi rutinitas perawatan sesuai dengan jenis dan masalah kulitmu dengan tingkat akurasi hingga 95%.

MULAI



Scan kode QR dengan kamera handphone



The image shows three sequential screenshots of the L'Oréal Skin Genius app. Step 1 shows a camera view with a white frame and a play button. Step 2 shows the same camera view with various skin concerns highlighted with colored boxes: DRY WRINKLES (red), FIRMNESS (orange), RADIANCE (yellow), and HYDRATION (green). Step 3 shows the app's main interface with a play button, featuring 'MORNING treatments' and two product options: 'Step 1 SERUM' (Revitalift Filler AA Serum) and 'Step 2 CREAM' (Revitalift Laser SPF Cream).

**Langkah 1**  
Siap untuk mengambil selfie? Lepas kacamata,

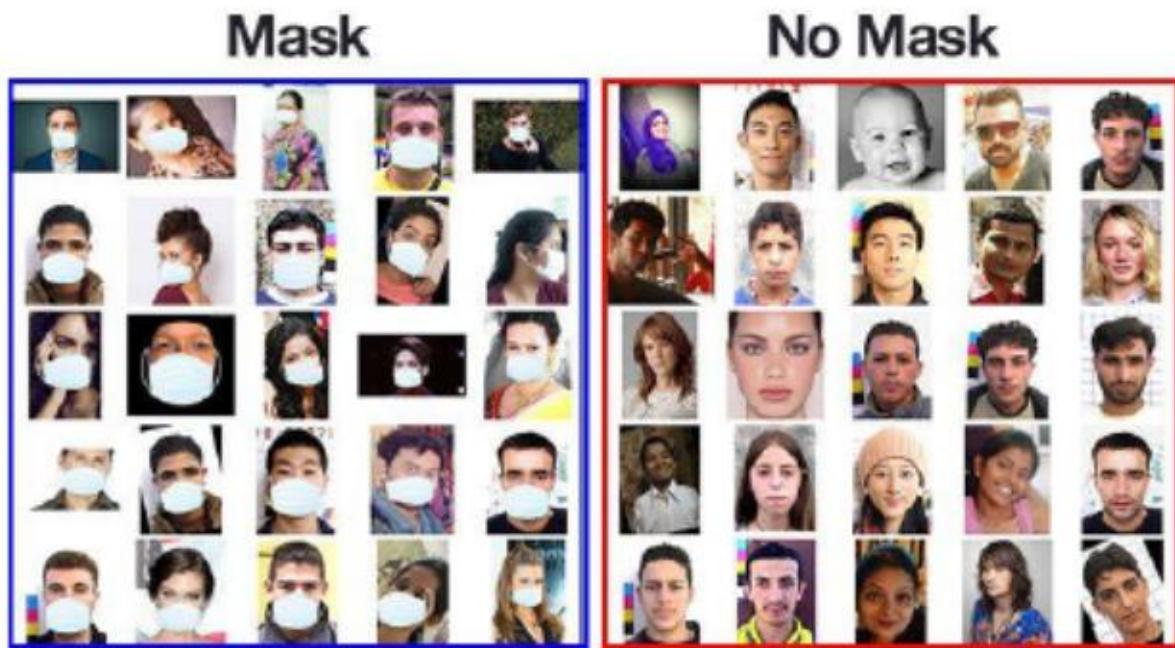
**Langkah 2**  
Setelah kualitas selfie telah terverifikasi,

**Langkah 3**  
Setelah Skin Genius selesai melakukan analisis,

2. Face Recognition



Dalam aplikasi pendekripsi penggunaan masker mesin dapat mengenali orang yang mengenakan masker dan tidak



Dalam kasus deteksi masker, datascience berperan dalam proses identifikasi data seperti melakukan Analisa ciri dari wajah seperti: mata, alis, hidung, mulut, janggung/kumis

### Soal Latihan

1. Apa perbedaan utama antara Data Science dan Artificial Intelligence (AI)?
  - a. Data Science berfokus pada pengumpulan dan analisis data, sedangkan AI lebih fokus pada pengembangan sistem yang dapat belajar secara mandiri.

- b. Data Science berfokus pada pengolahan gambar, sedangkan AI berfokus pada analisis teks.
  - c. Data Science hanya menggunakan statistik, sedangkan AI menggunakan matematika kompleks.
  - d. Data Science berfokus pada hardware komputer, sedangkan AI berfokus pada software
  - e. Data Science dan AI akan memberikan dampak signifikan, terutama dalam konteks bisnis dan ilmu pengetahuan.
2. Manakah pernyataan berikut yang benar mengenai aplikasi Data Science?
- a. Memprediksi hasil pertandingan sepak bola berdasarkan data historis.
  - b. Menciptakan robot yang bisa berpikir seperti manusia.
  - c. Mengembangkan sistem untuk mendeteksi wajah dalam foto.
  - d. Menganalisis pola pembelian pelanggan untuk meningkatkan strategi pemasaran
  - e. Menciptakan kamera yang dapat mendeteksi manusia / benda bergerak secara otomatis
3. Apa tujuan utama dari AI dalam konteks pengembangan teknologi?
- a. Menganalisis data untuk menemukan pola dan insight.
  - b. Membuat keputusan berdasarkan data historis.
  - c. Mempelajari perilaku manusia untuk menirunya.
  - d. Meningkatkan kecepatan komputasi dalam pengolahan data
  - e. Membuat keputusan sendiri dengan cepat dan akurat
4. Manakah pernyataan berikut yang paling tepat menggambarkan peran AI dalam perbedaannya dengan Data Science?
- a. AI berfokus pada pengolahan besar-besaran data, sedangkan Data Science hanya menggunakan data kecil.
  - b. AI lebih fokus pada pengembangan algoritma yang dapat belajar sendiri, sedangkan Data Science lebih pada analisis statistik tradisional.

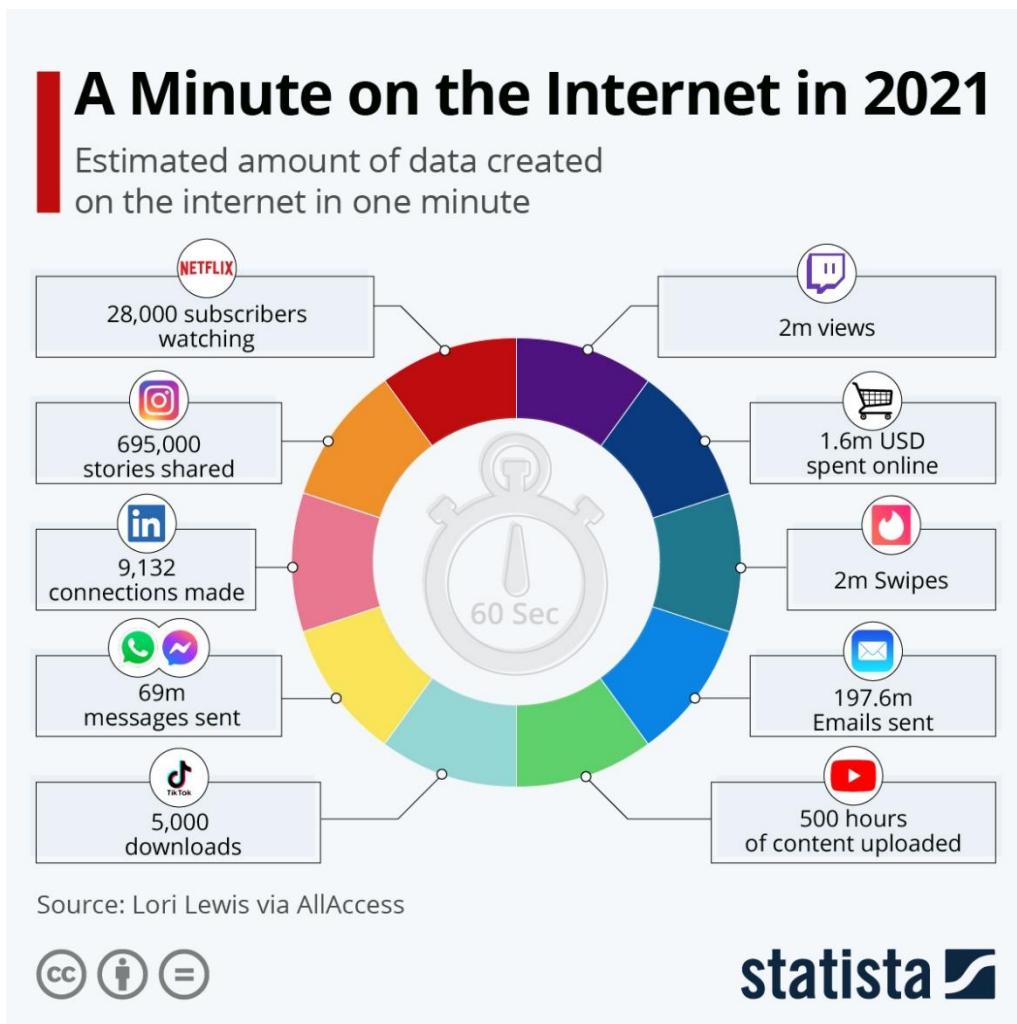
- c. AI berfokus pada analisis data, sedangkan Data Science lebih pada pengembangan perangkat lunak.
  - d. AI berfokus pada perhitungan matematis, sedangkan Data Science lebih pada visualisasi data.
  - e. AI berfokus pada perangkat *forecasting* sedangkan Data Science lebih kepada pengembangan perhitungan analisa matematika
5. Apa tujuan utama dari teknologi AI berbasis data dalam data science?
- A. Mengumpulkan sebanyak mungkin data
  - B. Menganalisis data untuk menemukan pola dan insight
  - C. Menyimpan data secara acak
  - D. Menjual data kepada perusahaan lain
  - E. Menghapus data yang tidak relevan
6. Algoritma Machine Learning mana yang digunakan untuk mengelompokkan data berdasarkan kemiripan karakteristiknya?
- A. Decision Tree
  - B. Linear Regression
  - C. K-Means Clustering
  - D. Support Vector Machine
  - E. Naive Bayes
7. Apa yang dimaksud dengan "Feature Engineering" dalam konteks data science dan AI?
- A. Proses menghapus fitur-fitur yang tidak penting dari data
  - B. Proses mengidentifikasi dan memilih fitur yang paling relevan untuk analisis
  - C. Proses mengekstrak data dari sumbernya
  - D. Proses menggandakan data agar lebih banyak
  - E. Proses menyimpan data dalam bentuk yang berbeda

8. Metode apa yang sering digunakan untuk menghindari overfitting dalam model Machine Learning?
- A. Menambahkan lebih banyak data training
  - B. Mengurangi kompleksitas model
  - C. Mengabaikan data training yang kompleks
  - D. Mengurangi jumlah epoch saat training
  - E. Menggunakan data yang kecil saja
9. Apa yang dimaksud dengan "Bias" dalam konteks data science dan AI?
- A. Data yang tidak relevan
  - B. Kesalahan yang disebabkan oleh model yang terlalu sederhana
  - C. Kesalahan yang disebabkan oleh model yang terlalu kompleks
  - D. Kecenderungan atau preferensi yang tidak diinginkan dalam analisis data
  - E. Variabel yang tidak berpengaruh dalam prediksi
10. Apa yang dimaksud dengan "Supervised Learning" dalam Machine Learning?
- A. Metode untuk belajar dari pengalaman sendiri
  - B. Proses untuk memperbaiki model setelah training
  - C. Algoritma untuk belajar dari data yang tidak memiliki label
  - D. Teknik untuk belajar dari data yang memiliki label
  - E. Teknik untuk belajar dari trial dan error

## BAB 4

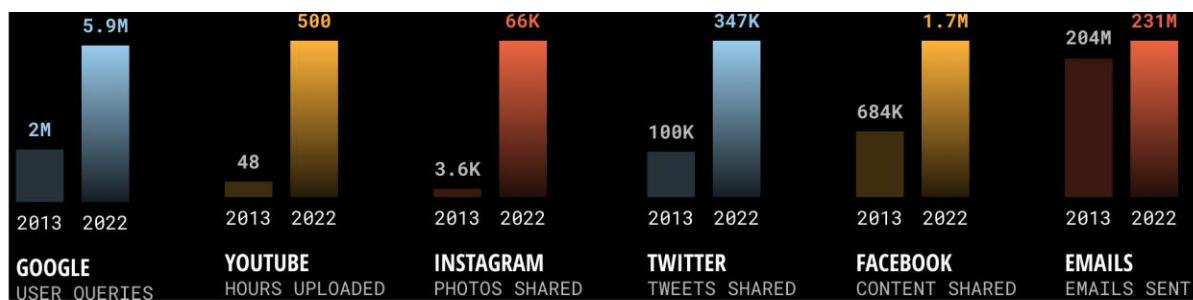
### Sains data Dan Machine Learning

Jumlah data yang kami hasilkan setiap hari sungguh mencengangkan. Terdapat 2,5 triliun byte data yang dihasilkan setiap hari dengan kecepatan yang kita miliki saat ini, namun kecepatan tersebut semakin meningkat seiring dengan pertumbuhan Internet of Things (IoT). Selama dua tahun terakhir saja, 90 persen data di dunia telah dihasilkan. Faktanya, jumlah pengguna internet telah meningkat lebih dari satu miliar dalam lima tahun terakhir, lebih dari separuh lalu lintas web dunia kini berasal dari telepon seluler



Gambar Penggunaan Internet per Menit

## Data Never Sleeps



Perkembangan data:

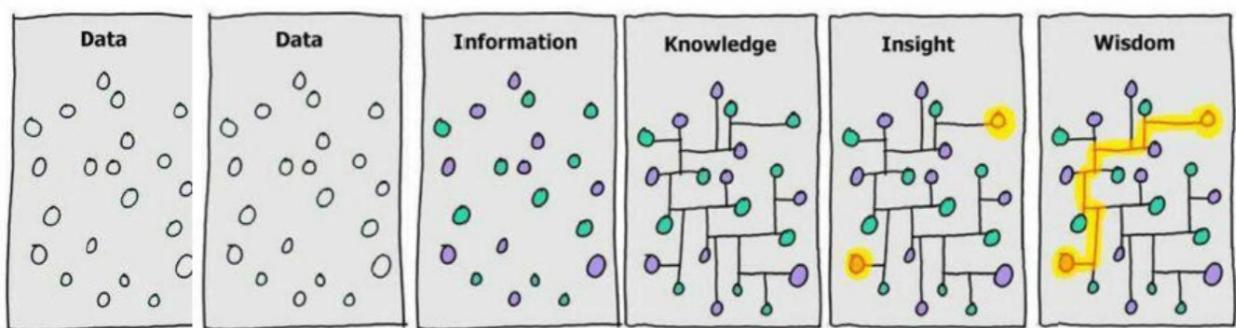
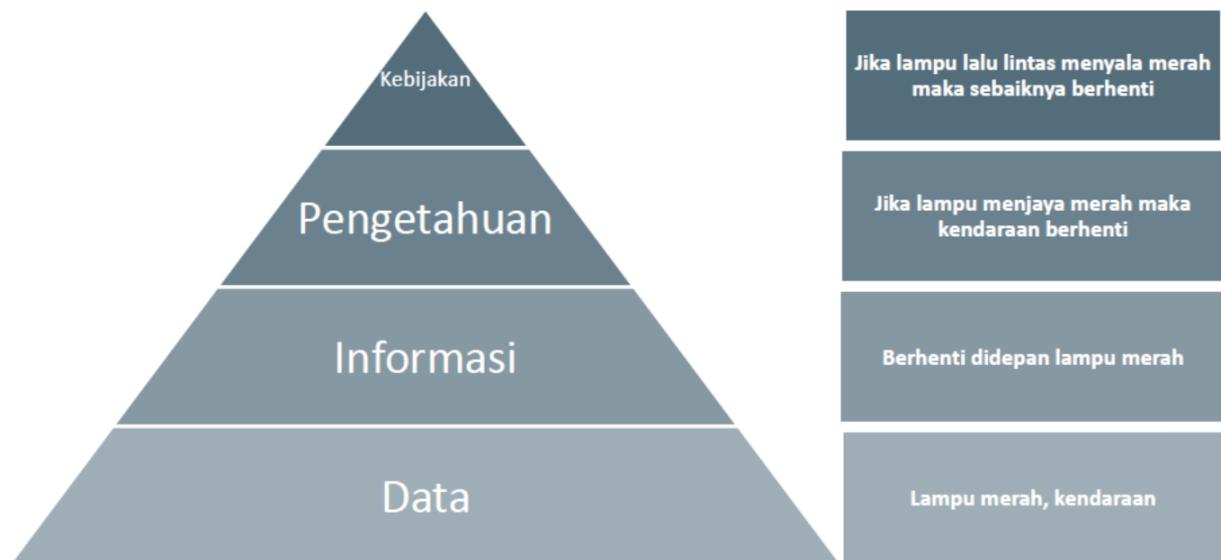
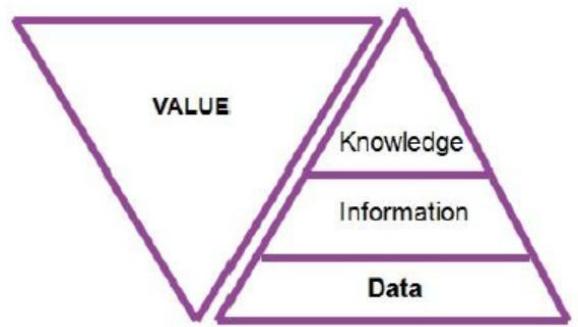
kilobyte (kB)	$10^3$
megabyte (MB)	$10^6$
gigabyte (GB)	$10^9$
terabyte (TB)	$10^{12}$
petabyte (PB)	$10^{15}$
exabyte (EB)	$10^{18}$
zettabyte (ZB)	$10^{21}$
yottabyte (YB)	$10^{24}$

### 4.1 Hubungan data science dan Machine learning

Analisis data dan proses bisnis dilakukan oleh data science dimana Pemodelan menggunakan learning machine. Data perlu kita olah menjadi pengetahuan agar bermanfaat bagi manusia.

Dengan pengetahuan kita dapat:

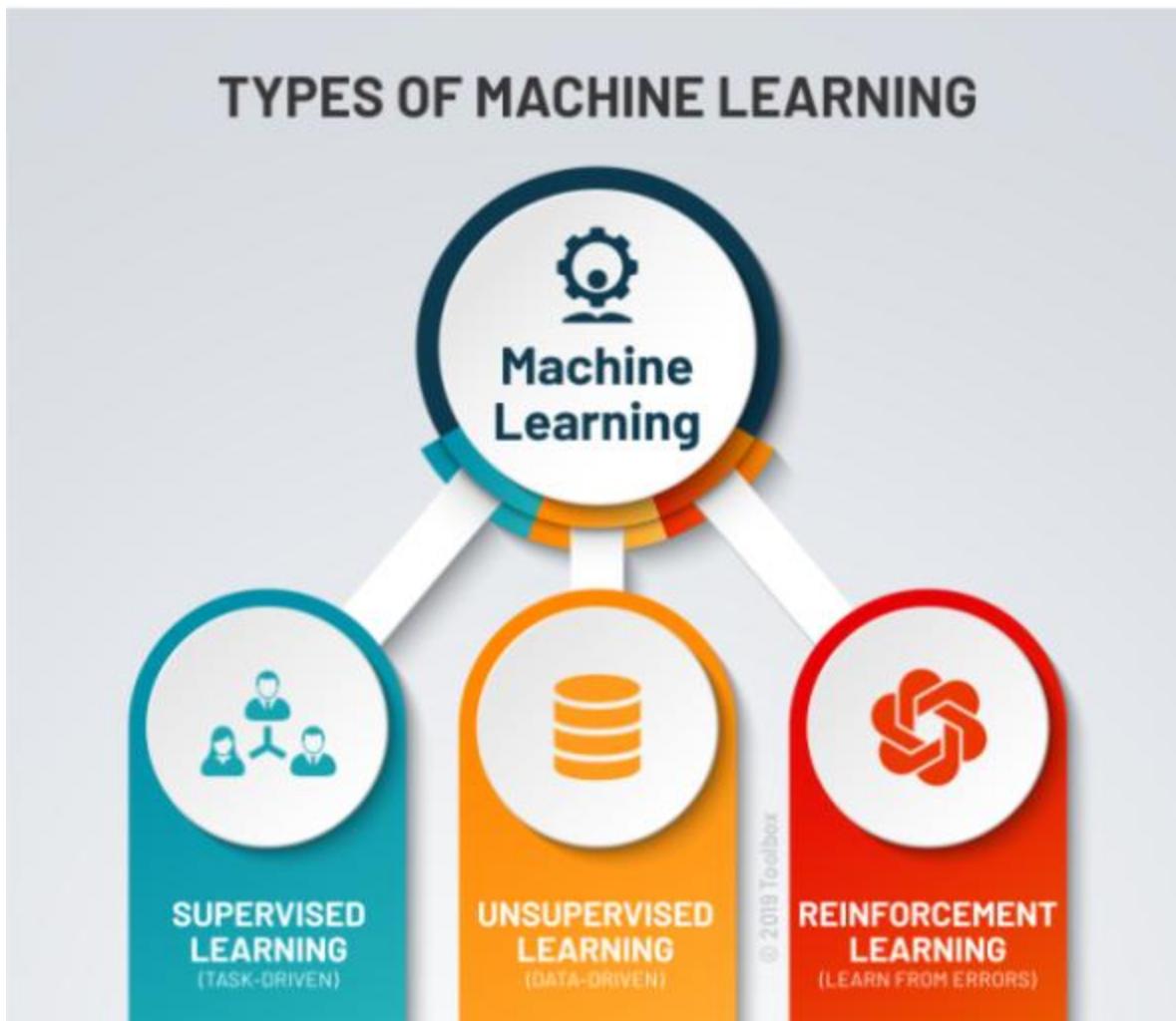
- Melakukan estimasi dan prediksi apa yang akan terjadi di depan
- Melakukan analisis tentang asosiasi, korelasi dan pengelompokan antar data dan atribut
- Membantu pengambilan keputusan dan pembuatan kebijakan



## 4.2 Kategori Machine Learning

- Supervised Learning

- Unsupervised Learning
- Reinforcement Learning



Tiga metode atau tipe machine learning (Sumber: Potentiaco.com)

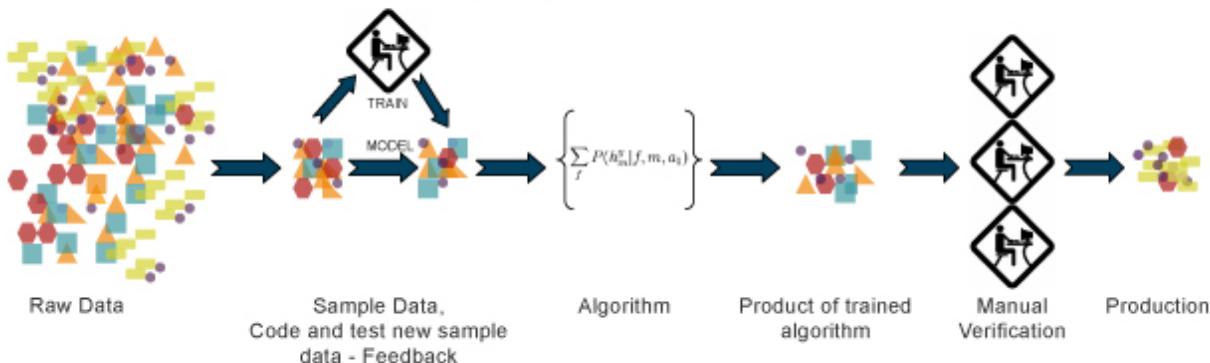
### 1. Supervised Learning

**Supervised Learning** adalah subkategori pembelajaran mesin (Machine Learning) dan kecerdasan buatan (AI).

Model ini ditentukan oleh penggunaan kumpulan data berlabel untuk melatih algoritma yang mengklasifikasikan data atau memprediksi hasil secara akurat. Dimana pembelajaran yang terawasi dimana jika output yang diharapkan telah diketahui sebelumnya. Biasanya pembelajaran ini dilakukan dengan menggunakan data yang telah ada.

## SUPERVISED LEARNING

Reliance on algorithm trained by human input, reduced expenditure on manual review for relevance and coding



*E-Discovery Concepts: Machine Learning*

Hudson | LEGAL

	Estimation	Prediction	Classification	Clustering	Assosiation
Attribute	Numeric	Numeric + Timeseries	Nominal & Numeric	Numeric	Nominal
Class/Label	Numeric	Numeric	nominal	-	-
Algorithm	Linear Regression, Neural Network, SVM, etc.	Linear Regression, Neural Network, SVM, etc.	Naive Bayes, k-Nearest Neigbor, C45, ID3 Logistic Regression, etc	K-Means, K-Medoids, Self-Organizing Map (SOM), etc	FP-Growth, A Priori, Coefficient of Correlation.
Evaluation	Root Mean square Error (RMSE), mse, etc	Root Mean square Error (RMSE), mse, etc	Accuracy, ROC, Area Under Curve (AUC)	Internal: dunn index, davies-bouldin index External: f-measure, confusion matrix	Lift Ratio, Precision and Recall (f-measure)

**Supervised Learning dapat dibagi kedalam 2 jenis berdasarkan cara pemecahan masalahnya**

### a. Klasifikasi (Classification)

Menggunakan algoritma, untuk menetapkan data uji secara akurat menetapkan ke dalam kategori tertentu. Jenis ini mengenali entitas tertentu dalam kumpulan data dan mencoba menarik beberapa kesimpulan tentang bagaimana entitas tersebut harus diberi label atau didefinisikan. Algoritma umum yang digunakan, SVM, Decission Tree, CNN, linear classifier dll.

fitur

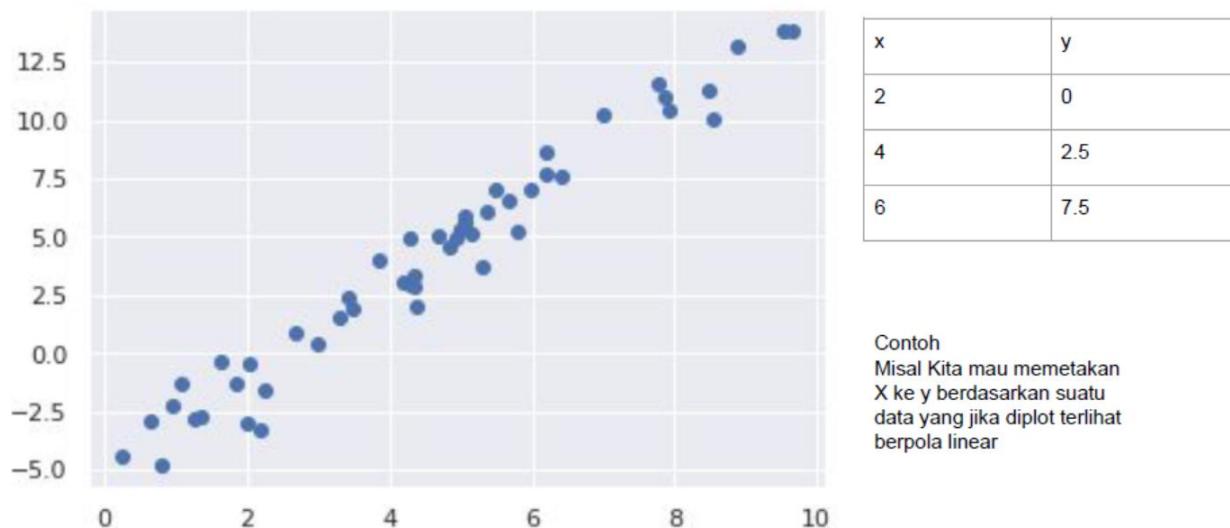
Data	MeanRed	MeanGreen	MeanBlue	Label
Gambar1.jpg	200	180	20	Kucing (0)
Gambar2.jpg	20	250	40	Kelinci
Gambar3.jpg	30	100	120	?

Komputer akan dilatih untuk menebak kelas Gambar3.jpg

### b. Regresi (Regression)

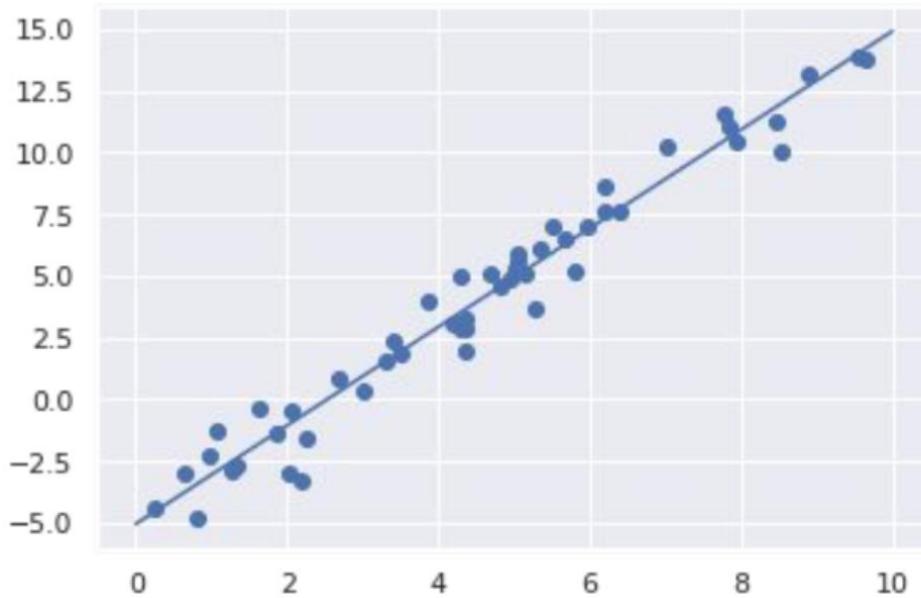
digunakan untuk mengetahui hubungan antara variabel dependen dan variabel bebas. Jenis Ini biasanya digunakan untuk membuat proyeksi, seperti untuk pendapatan penjualan untuk bisnis tertentu. Regresi yang paling umum digunakan saat ini adalah : Regresi Linier, Polinomial dan logistic regression

Contoh data sederhana:



Sumber : AINO Indonesia

[Contoh Sederhana Linear Regression?](#)



Contoh  
Misal Kita mau memetakan  
X ke y berdasarkan suatu  
data yang berpola linear

Sumber: AINO Indonesia

## 2. UnSupervised Learning

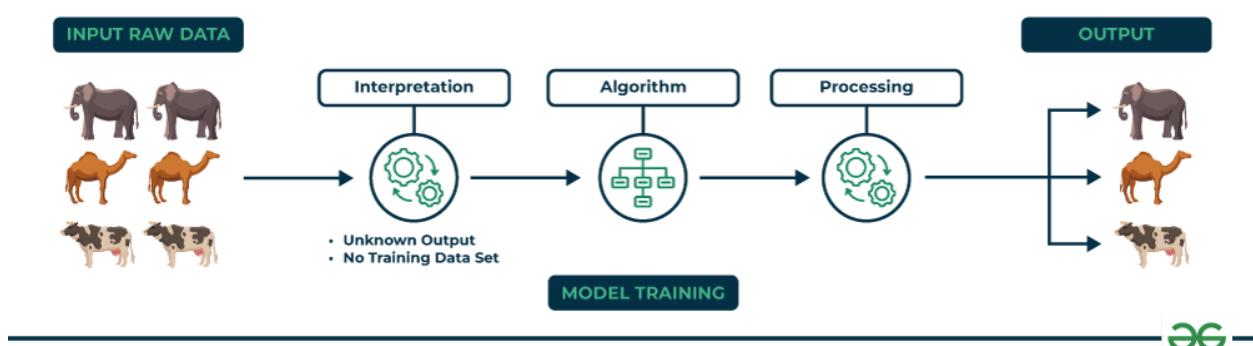
### Model UnSupervised Learning/Descriptive

Unsupervised Learning menggunakan algoritma pembelajaran mesin (machine learning) untuk menganalisis dan mengelompokkan kumpulan data yang tidak berlabel.

- **Algoritma ini menemukan pola tersembunyi atau pengelompokan data tanpa perlu campur tangan manusia, berbanding terbalik dengan Supervised Learning yang erat dengan campur tangan manusia.**
- **Kemampuan Unsupervised Learning menemukan persamaan dan perbedaan informasi menjadikannya solusi ideal untuk analisis data eksplorasi, strategi penjualan silang, segmentasi pelanggan, dan pengenalan gambar**
- **Dataset sama seperti klasifikasi hanya sekarang tanpa kolom kelas, Komputer dilatih untuk mengelompokkan datatersebut kedalam n kelompok**

Contoh algoritma yang digunakan di model ini: K-Means Clustering Algorithm.

# Unsupervised Learning

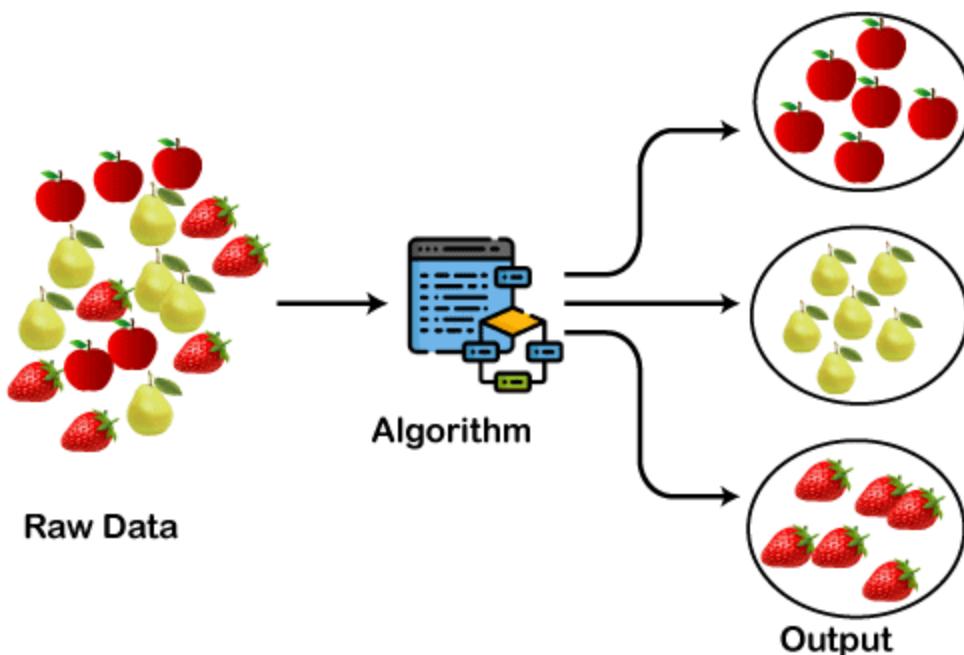


Sumber: <https://www.geeksforgeeks.org/ml-types-learning-part-2/>

### 3. Jenis-jenis Unsupervised Learning

#### Clustering

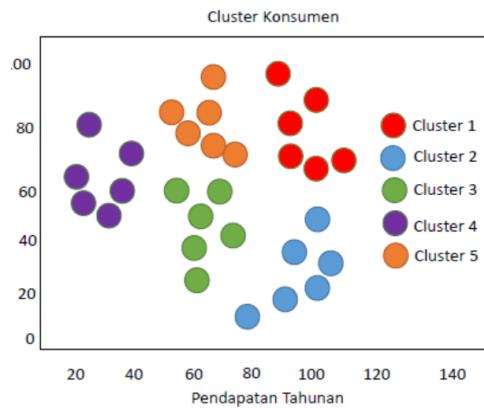
Clustering adalah Teknik data mining yang mengelompokkan data yang tidak berlabel berdasarkan persamaan atau perbedaannya. Algoritma pengelompokan digunakan untuk memproses objek data mentah yang tidak diklasifikasikan ke dalam kelompok yang diwakili oleh struktur atau pola dalam informasi.



Sumber : <https://static.javatpoint.com/tutorial/machine-learning/images/clustering-in-machine-learning.png>

## Hierarchical Clustering

Hierarchical Clustering adalah algoritma pengelompokan tanpa pengawasan yang dapat jika tegorikan dalam dua cara, yakni bisa menjadi aglomerat atau memecah belah.



## Probabilistik Clustering

### Probabilistik Clustering

Model probabilistik adalah teknik Unsupervised Learning yang memecahkan estimasi kepadatan atau masalah pengelompokan lunak. Dalam pengelompokan probabilistik, titik data dikelompokkan berdasarkan kemungkinan bahwa mereka termasuk dalam distribusi tertentu. Model Campuran Gaussian (GMM) adalah salah satu metode pengelompokan probabilistik yang paling umum digunakan. Model probabilistik adalah teknik Unsupervised Learning yang memecahkan estimasi kepadatan atau masalah pengelompokan lunak. Dalam pengelompokan probabilistik, titik data dikelompokkan berdasarkan kemungkinan bahwa mereka termasuk dalam distribusi tertentu.

## Association Rule

Association Rules adalah metode berbasis aturan untuk menemukan hubungan antara variabel dalam kumpulan data yang diberikan. Metode ini sering digunakan untuk analisis keranjang belanja, memungkinkan perusahaan untuk lebih memahami hubungan antara produk yang berbeda.

## Apriori Algoritm

Apriori Algorithms dipopulerkan melalui analisis keranjang belanja yang mengarah ke berbagai mesin rekomendasi untuk platform musik dan toko online. Mereka biasa digunakan dalam kumpulan data transaksional untuk mengidentifikasi kumpulan barang yang sering, atau

kumpulan barang, untuk mengidentifikasi kemungkinan mengkonsumsi suatu produk mengingat konsumsi produk lain

Transaksi 1				
Transaksi 2				
Transaksi 3				
Transaksi 4				
Transaksi 5				
Transaksi 6				
Transaksi 7				
Transaksi 8				

### 3. Reinforcement Learning

Reinforcement learning (RL) bekerja melalui sebuah proses feedback, dan akan terus melakukan aktivitasnya sampai ia mencapai tujuannya. Jika ia mencapai tujuan maka ia akan mendapatkan reward. Proses ini akan terus berlangsung dengan tujuan besarnya adalah memaksimalkan reward yang didapat



Melalui gambar di atas bisa dilihat bahwa agen melakukan sebuah aksi, yang kemudian aksi itu akan diterapkan di lingkungannya (baik dunia luar atau berupa simulasi).

Kemudian akan dilihat apakah tujuannya tercapai atau tidak.

Jika tercapai maka ia akan mendapat reward, jika tidak maka ia mendapat hukuman (punishment) atau bisa juga disetting tidak terjadi apa-apa.

Setiap keputusan (aksi) yang ia ambil, maka ia berada di kondisi (state) yang baru. Begitu seterusnya sampai reward yang didapat maksimal.

Contoh penggunaan:

- **Mobil Self Driving**

kasus yang ideal, komputer tidak boleh diberikan instruksi mengemudi apa pun. Pemrogram akan menghindari memasang kabel apa pun yang terkait dengan tugas dan sebaliknya membiarkan mesin belajar dari kesalahannya. Fungsi *rewards* akan menjadi satu-satunya fitur terprogram dalam pengaturan yang ideal.

- **Natural Language Processing (NLP)**

Peringkasan teks, penjawab pertanyaan, dan terjemahan mesin hanyalah beberapa dari aplikasi

### **Soal Latihan :**

1. Mana dari berikut ini yang merupakan metode validasi model dalam supervised learning?

- a. K-Means Clustering
- b. Principal Component Analysis (PCA)
- c. Cross-Validation
- d. Dimensionality Reduction
- e. Feature Selection

2. Apa hubungan antara Data Science dan Machine Learning?

- a. Data Science adalah cabang dari Machine Learning.
- b. Machine Learning adalah teknik yang digunakan dalam Data Science untuk menganalisis data.
- c. Data Science dan Machine Learning adalah dua bidang yang sepenuhnya terpisah tanpa hubungan.
- d. Data Science hanya mencakup pengolahan data, sedangkan Machine Learning berfokus pada statistik.
- e. Data Science mengutamakan perangkat keras, sementara Machine Learning fokus pada perangkat lunak.

3. Dalam supervised learning, model dilatih menggunakan:
- Data yang tidak memiliki label.
  - Data dengan label yang telah ditentukan sebelumnya.
  - Data yang hanya memiliki fitur tanpa target.
  - Data yang sepenuhnya acak.
  - Data yang hanya digunakan untuk validasi.
4. Apa tujuan utama dari unsupervised learning?
- Mengklasifikasikan data ke dalam kategori yang sudah ditentukan.
  - Memprediksi nilai numerik dari data yang sudah ada.
  - Mengidentifikasi pola atau struktur yang tersembunyi dalam data tanpa menggunakan label.
  - Mengoptimalkan keputusan dalam lingkungan yang dinamis.
  - Mengukur kinerja model terhadap data yang telah dilabeli.
5. Dalam reinforcement learning, agen belajar dengan:
- Menggunakan data yang telah dilabeli dan memprediksi hasilnya.
  - Menjelajahi lingkungan dan mendapatkan umpan balik dalam bentuk reward atau punishment.
  - Menggunakan data yang tidak memiliki label untuk mencari pola.
  - Menggunakan teknik clustering untuk mengelompokkan data.
  - Mengoptimalkan model berdasarkan data historis tanpa interaksi langsung
6. Apa tujuan utama dari algoritma Apriori dalam data mining?
- Mengklasifikasikan data ke dalam kategori yang telah ditentukan.
  - Memprediksi nilai numerik berdasarkan data historis.
  - Menemukan aturan asosiasi dalam data transaksi
  - Mengurangi dimensi data untuk visualisasi.
  - Mengelompokkan data ke dalam kelompok yang homogen.
7. Algoritma mana yang paling sering digunakan untuk regresi dalam data mining?
- K-Means
  - Naive Bayes

- c. Decision Trees
  - d. Linear Regression
  - e. Apriori
8. Apa tujuan utama dari data mining?
- a. Menghitung statistik dasar
  - b. Menghapus data yang tidak relevan
  - c. Menemukan pola dan informasi berguna dalam data besar
  - d. Mengatur data dalam tabel
  - e. Membuat laporan manual
9. Metode pembelajaran apa yang digunakan untuk membuat model berdasarkan data yang telah diberi label?
- a. Unsupervised Learning
  - b. Reinforcement Learning
  - c. Supervised Learning
  - d. Clustering
  - e. Association Rule
10. Apa tujuan dari unsupervised learning?
- a. Mengklasifikasikan data ke dalam kategori yang telah ditentukan
  - b. Mengoptimalkan keputusan berdasarkan umpan balik
  - c. Menemukan struktur atau pola dalam data tanpa label
  - d. Menyaring data yang tidak relevan
  - e. Mengatur data ke dalam tabel
11. Dalam konteks clustering, apa yang dimaksud dengan ‘centroid’?
- a. Titik data yang paling relevan
  - b. Titik pusat dari sebuah cluster
  - c. Titik yang memiliki jarak terjauh dari cluster
  - d. Data yang tidak bisa diklasifikasikan
  - e. Nilai rata-rata dari semua fitur

12. Metode apa yang digunakan untuk memperkirakan nilai output berdasarkan input dalam supervised learning?

- a. K-Means
- b. Random Forest
- c. Apriori
- d. Q-Learning
- e. Principal Component Analysis (PCA)

13. Apa itu reinforcement learning?

- a. Metode untuk mengelompokkan data tanpa label
- b. Teknik untuk menemukan aturan asosiasi dalam data
- c. Proses pembelajaran melalui trial and error dengan umpan balik
- d. Pembelajaran berdasarkan data yang sudah dilabeli
- e. Metode untuk menurunkan dimensi data

14. Apa yang dilakukan oleh algoritma K-Means dalam clustering?

- a. Menciptakan aturan asosiasi dari data
- b. Mengoptimalkan keputusan dengan feedback
- c. Mengelompokkan data ke dalam k kelompok berdasarkan kesamaan fitur
- d. Mengurangi dimensi data
- e. Mengklasifikasikan data berdasarkan label

15. Apa itu association rule dalam data mining?

- a. Metode untuk mengklasifikasikan data
- b. Teknik untuk mengurangi dimensi data
- c. Aturan yang menggambarkan hubungan antara item dalam dataset
- d. Metode untuk memprediksi nilai numerik
- e. Algoritma untuk memperbaiki data yang hilang

16. Dalam supervised learning, apa yang dimaksud dengan 'overfitting'?

- a. Model terlalu sederhana dan tidak menangkap pola yang kompleks
- b. Model terlalu kompleks dan menangkap noise sebagai pola\*
- c. Model tidak bisa dioptimalkan

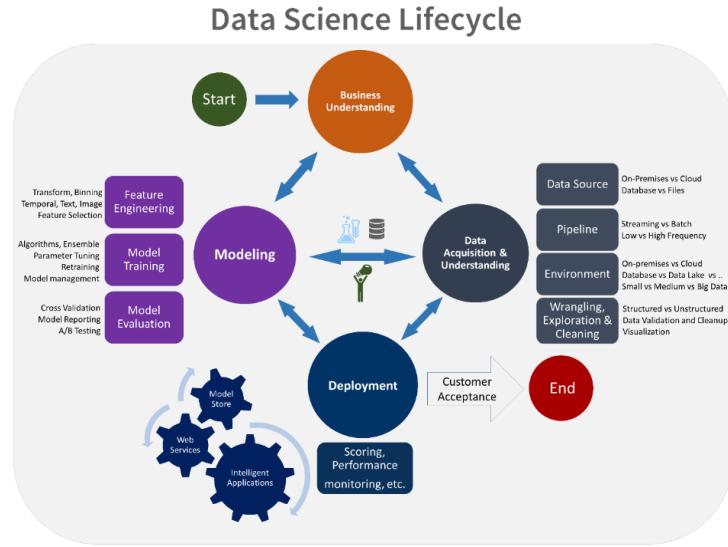
- d. Model memerlukan lebih banyak data
  - e. Model tidak mampu membedakan kelas
17. Metode unsupervised learning mana yang digunakan untuk reduksi dimensi data?
- a. K-Nearest Neighbors (KNN)
  - b. Principal Component Analysis (PCA)\*
  - c. Naive Bayes
  - d. Q-Learning
  - e. Apriori
18. Dalam reinforcement learning, apa yang dimaksud dengan 'reward'?
- a. Hadiah yang diberikan pada akhir proses
  - b. Nilai yang diperoleh dari aksi yang diambil dalam lingkungan\*
  - c. Label yang diberikan untuk data
  - d. Ukuran kesalahan model
  - e. Data yang digunakan untuk pelatihan
19. Metode klasifikasi apa yang menggunakan decision trees sebagai model?
- a. K-Means
  - b. Naive Bayes
  - c. Random Forest\*
  - d. Apriori
  - e. Principal Component Analysis (PCA)
20. Dalam association rule, apa itu 'confidence'?
- a. Probabilitas suatu aturan berlaku di seluruh data\*
  - b. Ukuran jarak antar item
  - c. Frekuensi kemunculan item dalam dataset
  - d. Evaluasi kesalahan model
  - e. Jumlah data yang hilang

## BAB 5

### Pemahaman Bisnis Proses Data

#### 5.1 Data Science Lifecycle

Data Science Lifecycle (DSL) digunakan untuk membantu organisasi dalam mengelola data mereka dengan lebih efektif dan efisien, sehingga dapat meningkatkan pengambilan keputusan dan mencapai tujuan bisnis.



#### 5.2 Pemahaman Bisnis

Pertama yang perlu dilakukan ialah tentukan metrik target, yang menentukan keberhasilan proyek. Identifikasi sumber data relevan yang dapat diakses atau perlu diperoleh oleh bisnis.

Langkah-langkah yang diperlukan dalam tahapan business understanding (IBM,2011) meliputi:

1. Menentukan Tujuan Bisnis
2. Melakukan Assesment
3. Menentukan tujuan data analytics
4. Merencanakan proyek data analytics

Tujuan Objectif AI dalam Bisnis yaitu meniru keputusan dan tindakan kecerdasan manusia

Contoh

- Autonomous Vehicle
- Stock Price Prediction

- **Otomatisasi** berfokus pada penyederhanaan tugas instruktif yang berulang

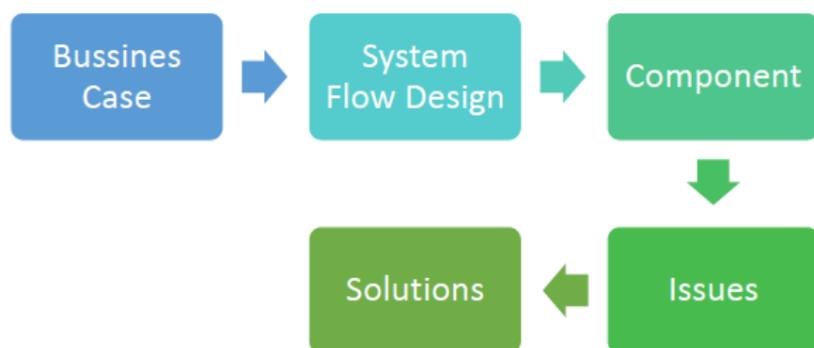
Contoh

Menyalakan lampu menggunakan smartphone setiap jam 18:00 Alat otomatis pemberi makan hewan peliharaan

### 5.3 Identifikasi Problem Machine Learning

Terdapat beberapa sistem pengolahan data, atau sistem pembelajaran yang memerlukan beberapa tahapan pengolahan yang masing-masing memiliki keterikatan satu sama lain untuk menghasilkan output.

1.Tahapan-tahapan Business Understanding di dalam AI



### 2.Rencana Proyek Data Science

Setiap proyek dimulai dengan pemahaman bisnis. Proyek Data Science merupakan proyek Bisnis, sehingga harus selalu berorientasi pada pencapaian hasil yang berfokus pada bisnis. Proyek Data Science harus memiliki visi global yang selaras dengan strategi bisnis. Sponsor bisnis membutuhkan solusi analitik

### Tahapan Proyek Data Science



### **Contoh bussines understanding:**

*Autonomous Vehicle Problem*: Kendaraan autonomous yang belum aman

### **Clear question:**

Bagaimana cara agar kendaraan autonomous dapat secara aman beroperasi?

#### 1. Analytic Approach

Setelah masalah bisnis dinyatakan dengan jelas, data scientist dapat menentukan pendekatan analitik untuk memecahkan masalah. Tahap ini mencakup pengungkapan masalah dalam konteks teknik statistik dan pembelajaran mesin. Tahap ini dapat membantu organisasi dalam mengidentifikasi solusi yang paling tepat

Pendekatan analitik yang dipilih akan menentukan kebutuhan data. Secara khusus, metode analitik yang akan digunakan memerlukan konten, format, dan representasi data tertentu, hal ini juga dipengaruhi oleh problem domain yang dipilih.

#### 2. Data Collection

Pada tahap pengumpulan data awal, data scientist dapat melakukan:

- a. Mengidentifikasi dan mengumpulkan sumber data yang tersedia—Terstruktur
- b. tidak terstruktur, semi-terstruktur—yang relevan dengan domain masalah
- c. Gathering Data
- d. Data dapat dikumpulkan melalui beberapa sumber, seperti
- e. Data internal perusahaan (excel, database internal, dll)
- f. Web API, Web scraping
- g. Kumpulan data melalui data publik
- h. Kumpulan data melalui data terbuka

### **Soal Latihan**

1. Apa yang dimaksud dengan Data Science Life Cycle?
  - a. Proses pengumpulan data saja
  - b. Proses analisis data saja
  - c. Proses pengumpulan, analisis, dan implementasi data
  - d. Proses pengumpulan, analisis, dan evaluasi data

- e. Proses pengumpulan, analisis, implementasi, dan evaluasi data\*
2. Tahap apa yang pertama kali dilakukan dalam Data Science Life Cycle?
- Data Collection
  - Data Cleaning
  - Data Analysis
  - Business Problem Framing\*
- e. Data Integration
3. Apa yang dimaksud dengan Identifikasi Problem Machine Learning?
- Proses pengumpulan data untuk machine learning
  - Proses analisis data untuk machine learning
  - Proses identifikasi masalah bisnis yang dapat diselesaikan dengan machine learning\*
  - Proses implementasi model machine learning
  - Proses evaluasi model machine learning
4. Tahap apa yang dilakukan setelah Data Collection dalam Data Science Life Cycle?
- Data Cleaning\*
  - Data Analysis
  - Data Integration
  - Business Problem Framing
  - Modeling
5. Apa yang dimaksud dengan Data Integration dalam Data Science Life Cycle?
- Proses pengumpulan data dari berbagai sumber
  - Proses analisis data untuk memperoleh wawasan
  - Proses menggabungkan data dari berbagai sumber\*
  - Proses implementasi model machine learning
  - Proses evaluasi model machine learning
- Jawaban: c) Proses menggabungkan data dari berbagai sumber
6. Tahap apa yang dilakukan setelah Data Analysis dalam Data Science Life Cycle?
- Modeling\*
  - Deployment

- c. Evaluation
- d. Data Integration
- e. Business Problem Framing

Jawaban: a) Modeling

7. Apa yang dimaksud dengan Deployment dalam Data Science Life Cycle?
- a. Proses implementasi model machine learning dalam lingkungan produksi\*
  - b. Proses analisis data untuk memperoleh wawasan
  - c. Proses menggabungkan data dari berbagai sumber
  - d. Proses evaluasi model machine learning
  - e. Proses identifikasi masalah bisnis
8. Tahap apa yang dilakukan setelah Deployment dalam Data Science Life Cycle?
- a. Evaluation\*
  - b. Modeling
  - c. Data Analysis
  - d. Data Integration
  - e. Business Problem Framing
- Jawaban: a) Evaluation
9. Apa yang dimaksud dengan Evaluation dalam Data Science Life Cycle?
- a. Proses analisis data untuk memperoleh wawasan
  - b. Proses menggabungkan data dari berbagai sumber
  - c. Proses implementasi model machine learning dalam lingkungan produksi
  - d. Proses evaluasi kinerja model machine learning\*
  - e. Proses identifikasi **masalah bisnis**
10. Apa yang harus dilakukan jika model machine learning tidak mencapai hasil yang diharapkan?
- a. Mengulangi proses Data Science Life Cycle
  - b. Mengubah parameter model machine learning
  - c. Menggunakan algoritma machine learning yang berbeda
  - d. Mengumpulkan lebih banyak data
  - e. Semua jawaban di atas\*

## BAB 6

### Tools Data Sains

Tools data Sains digunakan mulai dari proses data preparation, analisis hingga pengembangan. Dimana masing-masing memiliki kekurangan dan kelebihan. Berikut ini merupakan tools yang digunakan:

1. **Microsoft Excel**, Aplikasi ini biasanya digunakan untuk penolahan data numerik, fungsi statistika sampai dengan melakukan visualisasi data. Merupakan software standard yang biasa digunakan sebagai sumber database dengan file text atau file CSV.
2. **Python**, Bahasa pemrograman yang paling populer digunakan dalam Data Science, dengan library seperti NumPy, pandas, dan scikit-learn. Dibuat oleh Guido van Rossum dan pertama kali dirilis pada tahun 1991. Pandas, NumPy, dan Matplotlib yang memungkinkan manipulasi dan visualisasi data secara efektif digunakan dalam data science, serta **Scikit-learn** (untuk machine learning).
3. **SQL**, Digunakan untuk mengakses dan mengelola data yang tersimpan dalam database. SQL (Structured Query Language) adalah bahasa standar untuk query database relasional
4. R, Bahasa pemrograman yang banyak digunakan dalam statistik dan analisis data.
5. Julia: Bahasa pemrograman yang relatif baru, namun sudah menunjukkan kemampuan yang baik dalam Data Science.
6. RapidMiner, menggunakan platform AI dan struktur data dalam analisi data.
7. Knime, menawarkan platform lengkap untuk ilmu data menyeluruh, mulai dari membuat model analitik, hingga penerapannya dan berbagi wawasan dalam organisasi, hingga aplikasi dan layanan data.
8. Orange, perangkat lunak open source yang digunakan untuk melakukan data mining dan analisis data melalui visual programming. Dengan kata lain, Orange Data Mining memungkinkan pengguna untuk melakukan data mining tanpa perlu memiliki keterampilan pemrograman yang tinggi.

Ngolah big data pake



Ngolah big data pake



Ngolah big data pake



Ngolah big data pake



Ngolah big data pake



	A	B	C
1048571			
1048572			
1048573			
1048574			
1048575			
<b>1048576</b>			

The Last row in  
Excel

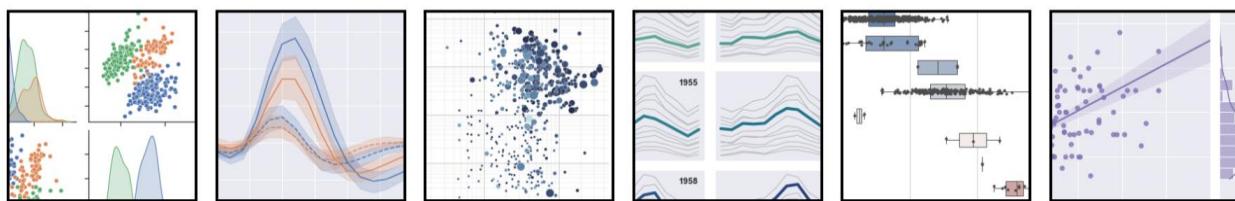


### Tools Visualisasi Data

Sebagian tools yang digunakan dalam visualisasi data :



Visualisasinya dapat berupa :



### 6.1 INFRASTRUKTUR BIG DATA

Pusat data (data center) adalah fasilitas yang digunakan oleh sebuah bisnis atau organisasi untuk menyimpan, memproses, dan menyerbarkan data dalam jumlah besar.

Hal-hal yang perlu diperhatikan dalam pusat data:

- ✓ Sumber daya listrik

- ✓ Keamanan
- ✓ Sistem pemrosesan dan penyimpanan
- ✓ Bandwidth
- ✓ Lokasi
- ✓ Keandalan
- ✓ Skalabilitas



Gambar Infrastruktur Big Data

## 6.2 Cloud Platforms

**Google Cloud Platform (GCP)**: Menyediakan berbagai layanan untuk analisis data besar, termasuk BigQuery (untuk analisis data besar), AutoML (untuk machine learning), dan banyak lagi.

- **Amazon Web Services (AWS)**: Platform cloud dari Amazon yang menawarkan berbagai layanan untuk penyimpanan data, pemrosesan data, dan machine learning, seperti **AWS S3** (untuk penyimpanan data), **AWS SageMaker** (untuk machine learning).
- **Microsoft Azure**: Platform cloud yang menawarkan berbagai alat analisis dan machine learning, termasuk **Azure Machine Learning** dan **Azure Databricks**

### Pengelola Data On-Promise VS Cloud



Jenis	Kelebihan	Kekurangan	Pengelola/ Penyedia
On-Premise	<ul style="list-style-type: none"> <li>• Tingkat keamanan tinggi</li> <li>• Biaya operasional rendah</li> <li>• Memberikan kontrol penuh</li> </ul>	<ul style="list-style-type: none"> <li>• Modal awal tinggi</li> <li>• Tanggung jawab sendiri</li> <li>• Skalabilitas terbatas</li> </ul>	Universitas, Pemerintah, Rumah Sakit, dll
Public Cloud	<ul style="list-style-type: none"> <li>• Skalabilitas</li> <li>• Modal yang rendah</li> <li>• Kemudahan akses</li> </ul>	<ul style="list-style-type: none"> <li>• Kontrol tidak penuh</li> <li>• Biaya operasional tinggi</li> <li>• SLA ditentukan penyedia</li> </ul>	Google, Amazon, Microsoft, dll



### Web Scraping vs API



## Web scraping

**Extracting data from a website using a web scraping software**

parsehub.com



## API

**Provide access to the data of an application, operating system or other services.**

**The goal for both is to access website data**

Akses tanpa batas

Lebih sulit

Bebas biaya

Isu etika

Akses terbatas

Lebih mudah

Berbayar

Legal

### 6.3 Jenis Penyimpanan Data

#### Unstructured data

The university has 5600 students.  
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.  
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.



#### Semi-structured data

```
<University>
<Student ID="1">
<Name>John</Name>
<Age>18</Age>
<Degree>B.Sc.</Degree>
</Student>
<Student ID="2">
<Name>David</Name>
<Age>31</Age>
<Degree>Ph.D. </Degree>
</Student>
...
</University>
```



#### Structured data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.



## **Soal Latihan**

**Manakah dari berikut ini yang termasuk tools populer dalam data science untuk analisis statistik dan visualisasi data?**

- A. AutoCAD
- B. Tableau
- C. Adobe XD
- D. WordPress
- E. Blender

**2. Python sering digunakan dalam data science karena alasan berikut, kecuali:**

- A. Banyak library untuk analisis data
- B. Gratis dan open-source
- C. Kurang dukungan untuk machine learning
- D. Komunitas yang besar
- E. Sintaks yang mudah dibaca

**3. Hadoop adalah bagian dari infrastruktur big data yang berfungsi untuk:**

- A. Mengedit video
- B. Mengelola email
- C. Menyimpan dan memproses data dalam jumlah besar secara terdistribusi
- D. Menjalankan aplikasi web
- E. Menyusun dokumen legal

**4. Di bawah ini adalah layanan cloud platform untuk penyimpanan dan komputasi data, kecuali:**

- A. Amazon Web Services (AWS)
- B. Google Cloud Platform (GCP)
- C. Microsoft Azure
- D. Dropbox
- E. Hadoop

**5. Salah satu keunggulan menggunakan cloud platform dalam big data adalah:**

- A. Memerlukan instalasi fisik perangkat keras besar
- B. Tidak bisa diakses secara remote
- C. Biaya tetap tanpa fleksibilitas
- D. Skalabilitas dan fleksibilitas tinggi
- E. Hanya cocok untuk data kecil

**6. Jenis penyimpanan data yang cocok untuk menyimpan data dalam bentuk tabel atau spreadsheet adalah:**

- A. Data warehouse
- B. Object storage
- C. Block storage
- D. NoSQL
- E. Blob storage

**7. Mana di bawah ini yang merupakan contoh tools Data Science untuk pemrograman dan analisis data?**

- A. PowerPoint
- B. Jupyter Notebook
- C. MS Paint
- D. Canva
- E. After Effects

**8. Salah satu kelebihan penyimpanan berbasis cloud dibanding penyimpanan lokal adalah:**

- A. Kapasitas terbatas
- B. Tidak bisa diakses dari perangkat lain
- C. Keamanan rendah
- D. Dapat diakses dari mana saja
- E. Hanya mendukung satu format file

**9. Berikut ini yang termasuk dalam jenis penyimpanan NoSQL adalah:**

- A. MySQL
- B. PostgreSQL
- C. MongoDB
- D. SQLite
- E. Oracle

**10. Platform berikut yang mendukung pengelolaan dan pemrosesan data besar (big data) secara cloud-native adalah:**

- A. TikTok
- B. Google BigQuery
- C. WhatsApp
- D. Figma
- E. Excel

## Pertemuan 7

### Teknik Pemrosesan Data

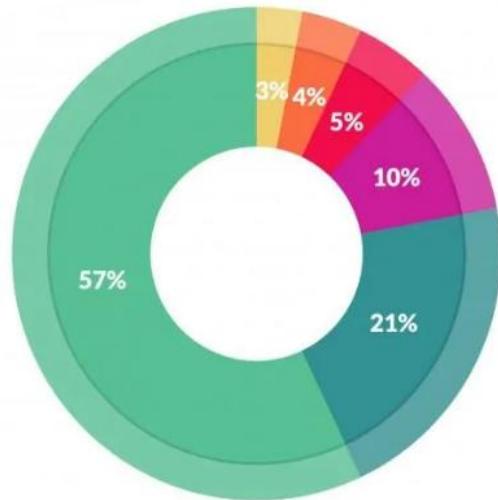
Teknik pemrosesan data dalam data science mencakup berbagai tahapan dan metode yang digunakan untuk mengubah data mentah menjadi informasi yang berguna dan siap dianalisis. Proses ini dimulai dari pengumpulan data, kemudian dilanjutkan dengan pembersihan, transformasi, dan analisis untuk menghasilkan wawasan dan pemahaman yang lebih baik. Penggunaan algoritma dan aspek matematika akan membantu pada saat pemrosesan data. Penggunaan teknik matematika digunakan untuk memahami *pattern* atau pola dalam dataset. Skill data science digunakan untuk membangun model dan menampilkan prediksi analisis dan pengujian pada testing.

#### 7.1 Data Preprocessing

Dalam data science, ada dua aspek penting yang perlu diperhatikan yaitu pre-processing. Data Preprocessing merupakan salah satu tahapan dalam melakukan mining data. Sebelum menuju ke tahap pemrosesan. Data mentah akan diolah terlebih dahulu.

#### Permasalahan Umum sebelum Processing Data

Permasalahan	Penjelasan Singkat
Data Hilang ( <i>Missing Values</i> )	Data tidak lengkap bisa mengganggu analisis atau membuat model bias
Outliers	Nilai ekstrem dapat memengaruhi rata-rata atau hasil model
Inkonsistensi Format	Data tidak memiliki format yang seragam (misalnya: tanggal, teks)
Data Duplikat	Mengganggu akurasi analisis atau mempengaruhi model secara negatif
Skala yang Berbeda	Perbedaan skala antar fitur menyebabkan dominasi fitur tertentu
Data Tidak Seimbang ( <i>Imbalanced</i> )	Terutama dalam klasifikasi, bisa membuat model bias terhadap kelas mayoritas
Noise	Kesalahan input atau data acak yang tidak relevan



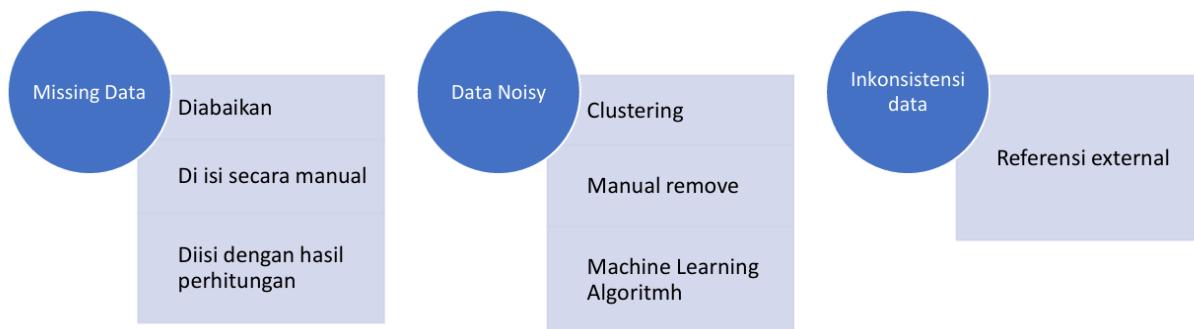
What's the least enjoyable part of data science?

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

Sumber : <https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/>

## 7.2 Data Preparation

Pembersihan yang dapat dilakukan pada data yang kotor sehingga diharapkan mendapat akurasi yang lebih baik, identifikasi dan memperbaiki error.



## 7.3 Tantangan Data Preparation

### 1. Volume dan Ragam Data yang Tersaji

Data yang dikumpulkan oleh organisasi berasal dari berbagai sumber. Ragam data yang tersaji dan volume data yang besar tentu menambah kompleksitas dalam proses pembersihan dan persiapan data. Seorang data scientist harus mampu menavigasi dan mengelola data yang kompleks untuk menghasilkan hasil analisis yang akurat.

## 2. Waktu dan Sumber Daya yang Dibutuhkan

Tugas data preparation sering memakan waktu yang signifikan dalam proyek data science. Hampir 45% dari keseluruhan tahapan proyek data science dihabiskan untuk tugas-tugas seperti loading dan cleaning data.

## 3. Kesulitan dalam Identifikasi Pola

Sebelum data dapat dianalisis, data scientist harus memastikan bahwa pola dan tren yang terkandung dalam data tidak terganggu oleh kesalahan atau ketidakakuratan.

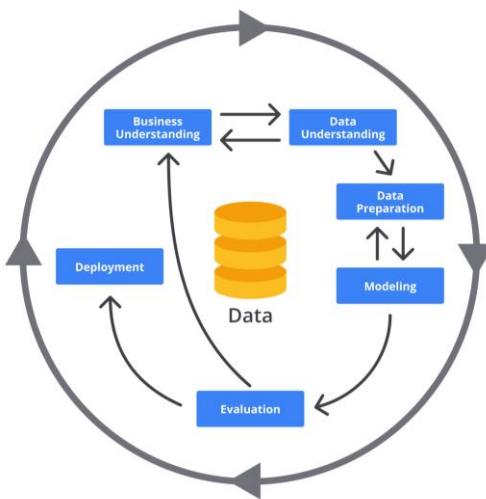
## 4. Kualitas Data yang Buruk

Seringkali, data yang diterima data scientist masih mentah dan memiliki kualitas yang buruk. Hal ini dapat disebabkan oleh beberapa faktor, termasuk proses pengumpulan data yang tidak terstruktur, kesalahan dalam pengambilan data, atau perbedaan format data antar sumber.

## 7.4 CRISP-DM

Cross Industry Standard Process for Data Mining (CRISP-DM). metodologi yang terstruktur dan terbukti untuk memandu proyek penambangan data, yang mencakup pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan penerapan

### Tahapan CRIPS-DM



### Soal Latihan

#### 1. Apa yang dimaksud dengan data preprocessing?

- A. Proses visualisasi data akhir

- B. Tahap membuat laporan hasil data mining
- C. Proses membersihkan dan menyiapkan data mentah
- D. Tahap interpretasi model prediktif
- E. Proses enkripsi data sebelum digunakan

**2. Manakah berikut ini yang termasuk langkah-langkah utama dalam data preprocessing?**

- A. Data labelling, data encryption, data modeling, data testing
- B. Data sampling, data collection, data analytics, data storage
- C. Data cleaning, data integration, data transformation, data reduction
- D. Data filtering, data partitioning, data optimization, data generation
- E. Data analysis, data mining, data visualization, data reporting

**3. Salah satu tantangan utama dalam data preprocessing adalah menangani data yang tidak lengkap. Apa pendekatan yang paling umum digunakan untuk menangani masalah ini?**

- A. Menghapus seluruh kolom data
- B. Menambahkan data fiktif secara acak
- C. Mengisi nilai yang hilang dengan estimasi seperti mean atau median
- D. Mengabaikan nilai yang hilang dan lanjut ke pemodelan
- E. Mengganti data hilang dengan nilai maksimum

**4. Bagaimana cara yang umum digunakan untuk menangani data yang hilang (missing values)?**

- A. Menghapus seluruh dataset
- B. Menambahkan data palsu
- C. Mengisi nilai hilang dengan mean atau median
- D. Mengabaikan nilai yang hilang sepenuhnya
- E. Mengacak nilai dari data lain

**5. Apa perbedaan utama antara data cleaning dan data transformation?**

- A. Cleaning berfokus pada kompresi, transformation pada backup
- B. Cleaning menyaring data, transformation menampilkan grafik
- C. Cleaning memperbaiki data kotor, transformation mengubah format atau struktur
- D. Cleaning menghapus data penting, transformation menggandakan data
- E. Cleaning hanya untuk gambar, transformation hanya untuk angka

**6. Salah satu tantangan dalam preprocessing adalah menangani data yang sangat besar (big data). Apa solusi yang tepat untuk mengatasi tantangan ini?**

- A. Menggunakan software spreadsheet seperti Excel
- B. Menyimpan data di hard disk eksternal

- C. Mengabaikan sebagian data secara acak
- D. Menggunakan teknik sampling atau distributed computing
- E. Menghapus data yang tidak dipahami

**7. Apa kepanjangan dari CRISP-DM?**

- A. Common Research In Standard Process for Data Modeling
- B. Cross-Industry Standard Process for Data Mining
- C. Certified Reporting In Secure Programming - Data Model
- D. Customer Rating In Strategic Prediction - Data Mining
- E. Core Rule Integration System Protocol - Data Metrics

**8. Manakah yang BUKAN termasuk fase dalam CRISP-DM?**

- A. Business Understanding
- B. Modeling
- C. Evaluation
- D. Deployment
- E. Visualization

**9. Mengapa tahap Business Understanding dalam CRISP-DM penting?**

- A. Untuk memilih algoritma data mining terbaik
- B. Untuk memahami struktur tabel data
- C. Untuk memastikan tujuan proyek sesuai kebutuhan bisnis
- D. Untuk membuat dashboard interaktif
- E. Untuk menghindari biaya proses data

**10. Apa pernyataan yang paling tepat mengenai perbedaan data preprocessing dan data preparation dalam CRISP-DM?**

- A. Data preprocessing dilakukan setelah deployment
- B. Data preparation hanya dilakukan oleh user interface designer
- C. Data preprocessing adalah bagian dari data preparation
- D. Data preparation tidak memerlukan data preprocessing sama sekali
- E. Data preprocessing hanya mencakup visualisasi data

## **BAB 8**

### **Penelaah Data**

Penelaahan data dalam analisis data adalah proses memeriksa, mengevaluasi, dan memahami data untuk mengidentifikasi pola, tren, wawasan, dan kesimpulan yang berharga. hal ini dilakukan setelah mendapatkan data mentah atau *raw data*. Ini adalah langkah penting dalam mengubah data mentah menjadi informasi yang bermanfaat dan dapat ditindaklanjut. Kegiatan ini bertujuan :

1. Mengidentifikasi masalah: Menemukan area yang perlu diperbaiki atau ditingkatkan.
2. Mengambil keputusan: Mendukung pengambilan keputusan yang lebih tepat dan berbasis fakta.
3. Meningkatkan efisiensi: Mengoptimalkan proses dan operasional berdasarkan temuan data.

#### **8.1 Analisis dan Penafsiran data**

Analisis dan penafsiran data dalam penelitian kualitatif memiliki ciri diantaranya :

Natural setting (latar alamiah),

1. Pengungkapan makna dari sudut pandang subyek penelitian,
2. holistik dan tidak dapat diisolasi sehingga terlepas dari konteknya, Peneliti sebagai instrumen utama untuk mengungkapkan makna yang terikat nilai dan konteks,
3. Data kualitatif diungkapkan melalui hubungan alamiah antara peneliti dengan informan.
4. Sample dipilih didasarkan oleh tujuan penelitian (purposive sampling) dan bukan menggunakan sampel random, Analisis Data data dilakukan secara induktif, serta Mengarahkan penyusunan teori dari data lapangan.

**Pada saat penelaah data dapat dilakukan dengan memperhatikan :**

1. Data Statistik: Nilai Mean, nilai modus dan nilai rata-rata
2. Distribusi Diagram : Box Plot, Histogram, korelasi data
3. Visualisasi Data : Scatter plot, heatmap,
4. Pencarian

## **8.2 Model Analisis Data**

### **1. Analisis Kawasan**

Analisis ranah (kawasan) merupakan proses menemukan bagian-bagian, unsur-unsur, kawasan kawasan dari makna kultural yang mengandung kategori-kategori lebih kecil. Berkaitan dengan hal itu, Spradley (1972:88-91) mengemukakan bahwa suatu kawasan kultural adalah suatu kategori dari makna kultural yang meliputi kategorir-ketegori yang lebih keci outlier

### **2. Observasi Terfokus**

Prasyarat untuk mememilih fokus adalah daftar ranah secara lengkap. Melalui daftar peneliti tersebut peneliti dapat memilih satu atau sejumlah ranah untuk dilakukan studi terfokus

### **3. Analisis Taksonomi**

Setelah analisis kawasan (ranah) dan observasi terfokus peneliti sudah dapat mengidentifikasi ranah ranah yang akan dipelajari secara mendalam.

### **4. Analisis Komponensial**

Analisis komponensial merupakan suatu usaha mencari secara sistematis atribut-atribut yang berhubungan dengan kategori budaya

### **Soal Latihan**

#### **1. Apa tujuan utama dari penelaahan data dalam analisis data?**

- A. Menghapus semua data mentah
- B. Menyederhanakan laporan keuangan
- C. Mengubah data mentah menjadi informasi yang dapat ditindaklanjuti
- D. Menambahkan data palsu ke sistem
- E. Menghindari pengambilan keputusan

#### **2. Berikut ini adalah tujuan penelaahan data, kecuali...**

- A. Mengidentifikasi masalah
- B. Mengambil keputusan

- C. Meningkatkan efisiensi
- D. Menyembunyikan kesalahan data
- E. Menemukan pola dan tren

**3. Dalam penelitian kualitatif, data biasanya dikumpulkan di...**

- A. Laboratorium tertutup
- B. Natural setting (latar alamiah)
- C. Kelas komputer
- D. Ruang rapat formal
- E. Simulasi tertutup

**4. Ciri analisis kualitatif adalah...**

- A. Berdasarkan angka dan grafik
- B. Menggunakan statistik inferensial
- C. Bersifat holistik dan terikat konteks
- D. Selalu memakai uji hipotesis
- E. Berdasarkan pengambilan sampel random

**5. Dalam analisis kualitatif, peneliti berperan sebagai...**

- A. Alat pembanding statistik
- B. Pemrogram utama
- C. Instrumen utama yang mengungkap makna
- D. Operator SPSS
- E. Pengamat pasif

**6. Salah satu cara analisis statistik deskriptif adalah...**

- A. Melakukan clustering
- B. Menjalankan decision tree
- C. Menghitung mean, median, modus
- D. Menyusun sintesis teori
- E. Menentukan coding kategori

**7. Visualisasi *box plot* digunakan untuk...**

- A. Menunjukkan frekuensi kategori
- B. Menghitung korelasi
- C. Menampilkan distribusi dan outlier
- D. Menentukan relasi antar teks
- E. Menentukan ukuran gambar

**8. Visualisasi *scatter plot* paling tepat digunakan untuk...**

- A. Mencari nilai minimum
- B. Mengukur variasi antar kategori
- C. Melihat hubungan antara dua variabel numerik
- D. Mengukur modus data kategorikal
- E. Menentukan ukuran sampel

**9. Heatmap digunakan untuk menggambarkan...**

- A. Persebaran nilai dalam diagram lingkaran
- B. Hubungan atau korelasi antar variabel dalam bentuk warna
- C. Nilai tengah data
- D. Variasi teks kualitatif
- E. Ukuran kolom data

**10. Dalam model analisis kawasan menurut Spradley, kawasan adalah...**

- A. Wilayah administratif
- B. Unit observasi geografis
- C. Kategori makna kultural besar yang memuat subkategori
- D. Blok kode numerik
- E. Dataset utama yang diproses

**11. Tujuan analisis kawasan adalah...**

- A. Menghapus nilai ekstrim
- B. Mengubah data numerik menjadi kategorikal
- C. Mengidentifikasi bagian-bagian dari makna kultural

- D. Mengganti kategori utama
- E. Menentukan urutan logis pengolahan data

**12. Observasi terfokus dilakukan setelah...**

- A. Uji normalitas
- B. Analisis komponensial
- C. Daftar ranah disusun lengkap
- D. Data statistik dihitung
- E. Visualisasi dibuat

**13. Analisis taksonomi bertujuan untuk...**

- A. Mengurangi jumlah variabel
- B. Memvalidasi outlier
- C. Menganalisis secara mendalam ranah terpilih
- D. Mengubah ranah menjadi angka
- E. Menetapkan alat ukur

**14. Analisis komponensial difokuskan pada...**

- A. Mengelompokkan data berdasarkan nilai mean
- B. Mencari atribut dalam kategori budaya
- C. Menyesuaikan data dengan teori statistik
- D. Menyusun grafik batang
- E. Menguji hubungan antar populasi

**15. Teknik purposive sampling digunakan dalam penelitian kualitatif untuk...**

- A. Menghasilkan sampel acak
- B. Memilih sampel secara sistematis
- C. Memilih sampel berdasarkan tujuan penelitian
- D. Menentukan ukuran populasi
- E. Menghindari bias pengambilan data

## BAB 9

### Visualisasi Data

#### 9.1 Perlunya Visualisasi Data

Visualisasi data sangat penting karena membantu menyederhanakan informasi kompleks menjadi bentuk yang mudah dipahami, memungkinkan pengambilan keputusan yang lebih cepat dan tepat. Karena :

1. Setengah dari otak manusia, (secara langsung, maupun tidak langsung) dikhususkan untuk memproses informasi visual.



Sumber : MIT Research – Brain Processing of Visual Information

2. Kemampuan untuk mengidentifikasi gambar yang dilihat secara singkat dapat membantu otak memutuskan dimana harus menentukan fokus.

**Otak manusia dapat mengidentifikasi gambar yang terlihat hanya dalam 13 milidetik.\***



3. Manusia mudah merepresentasikan apa yang mereka lihat.



## 9.2 Manfaat Visualisasi Data:

- **Memudahkan Pemahaman:**

Data yang kompleks dapat diubah menjadi representasi visual yang lebih mudah dicerna, seperti grafik atau diagram.

- **Mempercepat Pengambilan Keputusan:**

Dengan visualisasi, informasi yang relevan dapat diakses dengan cepat, memungkinkan pengambilan keputusan yang lebih tepat dan efisien.

- **Identifikasi Pola dan Tren:**

Visualisasi data membantu dalam menemukan pola, tren, dan hubungan antar variabel yang mungkin tidak terlihat dalam data mentah.

- **Meningkatkan Komunikasi dan Kolaborasi:**

Visualisasi data memungkinkan anggota tim untuk memahami data secara bersama-sama, memfasilitasi diskusi dan kolaborasi yang lebih efektif.

- **Mendukung Analisis Data:**

Visualisasi data menjadi alat yang penting dalam proses analisis data, membantu dalam mengidentifikasi masalah, mengevaluasi kinerja, dan membuat prediksi.

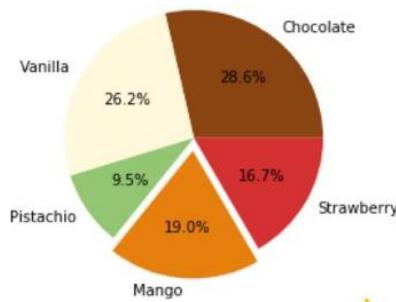
- **Meningkatkan Keterlibatan Pengguna:**

Visualisasi data yang menarik dapat meningkatkan keterlibatan pengguna dalam berinteraksi dengan data, baik itu dalam konteks bisnis, pendidikan, atau penelitian.

## 9.3 Grafik

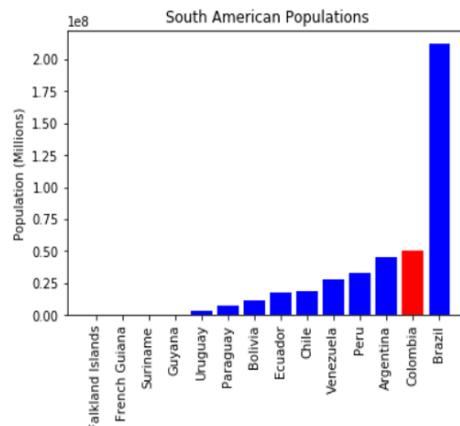
### Pie Chart Pie

Pie Chart Pie chart digunakan untuk menunjukkan seberapa banyak dari setiap jenis kategori dalam dataset dengan berbanding keseluruhan. Variabel label berisi tupel rasa es krim Variabel voting berisi tupel voting. Data tersebut mewakili jumlah voting rase es krim favorit.



### BarChart

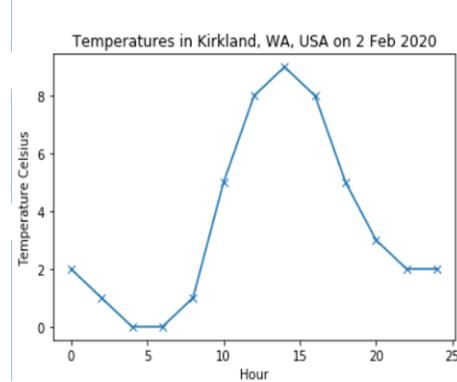
Bar Chart adalah merupakan tools visualisasi yang dapat digunakan untuk membandingkan kategorikal. Mirip dengan diagram lingkaran, diagram ini dapat digunakan untuk membandingkan kategori data satu sama lain. Diagram batang dapat menampilkan lebih banyak kategori data daripada diagram lingkaran.



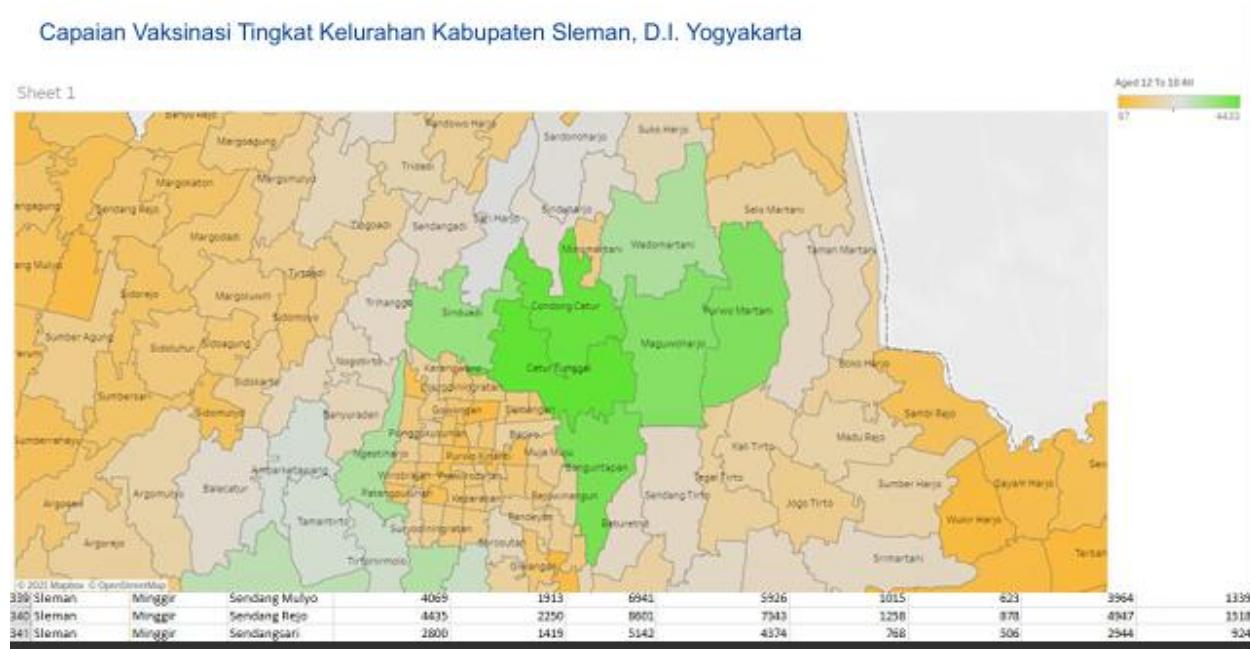
### Line Graph

Line Graph adalah bentuk visualisasi lainnya selain diagram lingkaran dan diagram batang.

- Diagram garis lebih berguna untuk menunjukkan bagaimana kemajuan data selama beberapa periode. • Misalnya, grafik garis dapat berguna dalam membuat grafik temperatur dari waktu ke waktu, harga saham dari waktu ke waktu, berat menurut hari, atau metrik berkelanjutan lainnya.

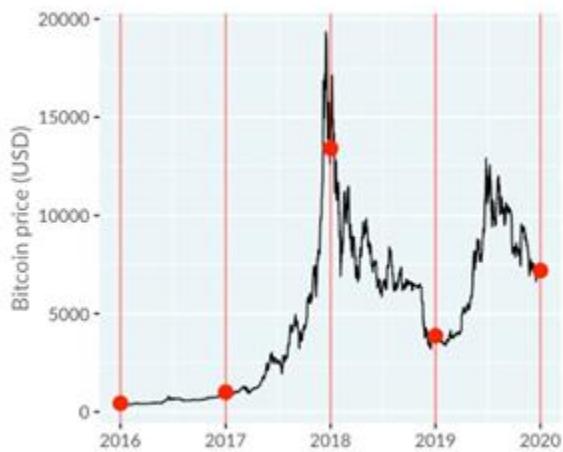


## Contoh Penggunaan:



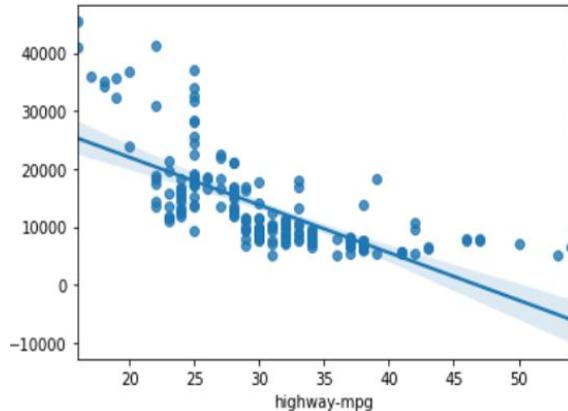
## Pola harga Bitcoin

Pada tahun berapa Bitcoin dimulai dengan harga tertinggi?



### Correlation & Causation

pengukuran sejauh mana nilai saling ketergantungan antar variabel. Causation merupakan hubungan antara sebab dan akibat antara dua variable. Penting untuk mengetahui perbedaan antara keduanya dan bahwa korelasi tidak mendeskripsikan sebab akibat. Menentukan korelasi jauh lebih sederhana menentukan sebab memerlukan analisis lebih lanjut.



### Soal Latihan

#### 1. Apa tujuan utama dari visualisasi data?

- A. Menyembunyikan informasi dari data
- B. Mengubah data menjadi teks panjang
- C. Mempercepat proses pengambilan keputusan melalui penyajian data yang mudah dipahami
- D. Menambahkan warna agar data menarik
- E. Menyederhanakan perhitungan data

**2. Jenis grafik manakah yang paling sesuai untuk menunjukkan perubahan data dari waktu ke waktu?**

- A. Diagram batang
- B. Pie chart
- C. Scatter plot
- D. Line chart
- E. Box plot

**3. Mengapa pie chart kurang tepat digunakan untuk membandingkan banyak kategori?**

- A. Karena warnanya terlalu banyak
- B. Karena sulit membandingkan ukuran sudut secara akurat
- C. Karena tidak bisa digunakan di Excel
- D. Karena data menjadi tidak valid
- E. Karena hanya menampilkan data numerik

**4. Jenis visualisasi apa yang paling cocok untuk menunjukkan distribusi data?**

- A. Line chart
- B. Pie chart
- C. Histogram
- D. Bar chart
- E. Radar chart

**5. Salah satu manfaat visualisasi data dalam bisnis adalah:**

- A. Mengurangi biaya produksi secara langsung
- B. Menghindari kesalahan dalam entri data
- C. Menyampaikan informasi kompleks dengan cara yang mudah dipahami
- D. Menghilangkan kebutuhan akan data mentah
- E. Mengganti kebutuhan pelatihan karyawan

**6. Dalam visualisasi data, “dashboard” merujuk pada:**

- A. File data mentah
- B. Tampilan visual interaktif untuk memantau data secara real-time
- C. Grafik 3D yang diputar
- D. Kumpulan tabel dalam bentuk PDF
- E. Halaman depan laporan tahunan

**7. Grafik apa yang paling tepat digunakan untuk menunjukkan hubungan antara dua variabel numerik?**

- A. Pie chart
- B. Line chart
- C. Scatter plot
- D. Area chart
- E. Treemap

**8. Dalam bidang pendidikan, visualisasi data sering digunakan untuk:**

- A. Mengganti peran guru

- B. Membingungkan siswa dengan data statistik
- C. Menunjukkan perkembangan hasil belajar siswa secara visual
- D. Menghapus data ujian
- E. Menambah jumlah siswa

**9. Berikut ini adalah contoh alat visualisasi data KECUALI:**

- A. Tableau
- B. Microsoft Excel
- C. Power BI
- D. SQL Server
- E. Google Data Studio

**10. Apa yang dimaksud dengan heatmap dalam visualisasi data?**

- A. Grafik batang yang diberi warna panas
- B. Grafik yang menunjukkan hubungan dua kategori dengan warna
- C. Grafik lingkaran dengan warna merah
- D. Peta suhu wilayah
- E. Bar chart dengan efek 3D

## Bab 10

### Teknik Sampling

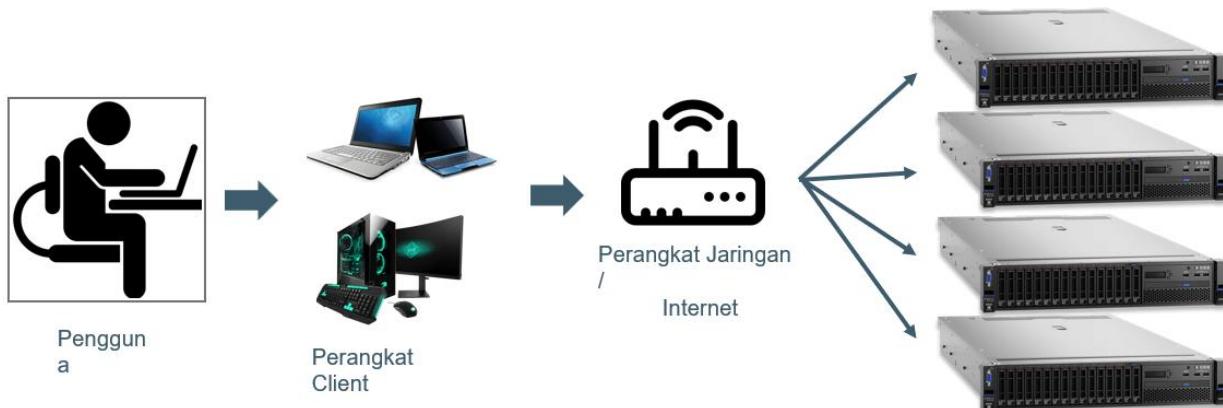
#### 10.1 Big Data

Definisi : Data yang sangat banyak (volume), beragam (variety) dan bertambah dengan sangat cepat (velocity) yang tidak cukup diolah dengan perangkat keras dan perangkat lunak biasa.



#### INFRASTRUKTUR BIG DATA

Meskipun server computer lebih powerful untuk mengolah data, namun dibutuhkan banyak server computer untuk mengolah big data secara parallel dan terdistribusi dalam suatu cluster. Misal: Pengguna melakukan pencarian di Google



Komputasi terdistribusi adalah penggunaan terkoordinasi dari computer yang secara fisik terpisah atau terdistribusi.

Pusat data (data center) adalah fasilitas yang digunakan oleh sebuah bisnis atau organisasi untuk menyimpan, memproses, dan menyerbarkan data dalam jumlah besar.

Hal-hal yang perlu diperhatikan dalam pusat data:

1. Sumber daya listrik
2. Keamanan
3. Sistem pemrosesan dan penyimpanan

4. Bandwidth
5. Lokasi
6. Keandalan
7. Skalabilitas

### On Promise VS Cloud

Jenis	Kelebihan	Kekurangan	Pengelola/ Penyedia
On-Premise	<ul style="list-style-type: none"> <li>• Tingkat keamanan tinggi</li> <li>• Biaya operasional rendah</li> <li>• Memberikan kontrol penuh</li> </ul>	<ul style="list-style-type: none"> <li>• Modal awal tinggi</li> <li>• Tanggung jawab sendiri</li> <li>• Skalabilitas terbatas</li> </ul>	Universitas, Pemerintah, Rumah Sakit, dll
Public Cloud	<ul style="list-style-type: none"> <li>• Skalabilitas</li> <li>• Modal yang rendah</li> <li>• Kemudahan akses</li> </ul>	<ul style="list-style-type: none"> <li>• Kontrol tidak penuh</li> <li>• Biaya operasional tinggi</li> <li>• SLA ditentukan penyedia</li> </ul>	Google, Amazon, Microsoft, dll

### Contoh Public Cloud



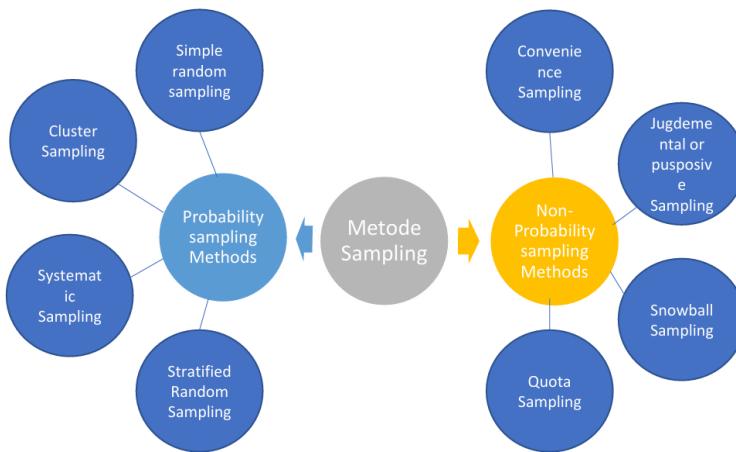
### 10.2 Sampling Data

Pengertian Sampling Sebelum preparation, melakukan terlebih tahapan dahulu dalam data adalah pemilihan/penentuan objek yang dapat dilakukan dengan menggunakan Penentuan: Populasi Sampel Populasi.



### Metode Sampling

Terbagi menjadi 2 yaitu probability sampling dan non probability sampling:



## 1. Probability Sampling (Sampling Probabilitas)

Setiap elemen dalam populasi memiliki **peluang yang sama** untuk dipilih sebagai sampel. Cocok untuk menghasilkan sampel yang **representatif dan objektif**.

**Jenis-jenisnya:**

1. **Simple Random Sampling**
  - Setiap anggota populasi memiliki peluang yang sama, biasanya dengan undian atau acak komputer.
2. **Stratified Sampling**
  - Populasi dibagi menjadi strata (kelompok) berdasarkan karakteristik tertentu, lalu sampel diambil dari tiap strata.
3. **Cluster Sampling**
  - Populasi dibagi menjadi kelompok (cluster), lalu beberapa cluster dipilih secara acak dan semua anggota dalam cluster tersebut disurvei.
4. **Systematic Sampling**
  - Memilih elemen sampel dari populasi berdasarkan interval tetap, misalnya setiap orang ke-10 dari daftar.

## 2. Non-Probability Sampling (Sampling Non-Probabilitas)

Tidak semua elemen populasi memiliki peluang yang sama untuk dipilih. Sering digunakan saat sulit mengakses seluruh populasi.

**Jenis-jenisnya:**

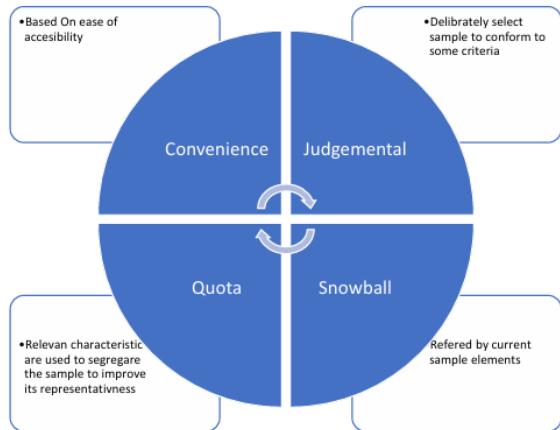
1. **Convenience Sampling**  
Mengambil sampel dari yang paling mudah dijangkau.
2. **Purposive (Judgmental) Sampling**  
Sampel dipilih berdasarkan pertimbangan atau tujuan tertentu.

### 3. Quota Sampling

Sampel dipilih sampai kuota tertentu tercapai, sering berdasarkan proporsi karakteristik.

### 4. Snowball Sampling

Digunakan untuk populasi yang sulit diakses; responden awal diminta merekomendasikan responden lain.



## Tahapan sampling

### 1. Menentukan Tujuan Sampling

Apa yang ingin dicapai dari pengambilan sampel?

Contoh: Mengukur kepuasan pelanggan, mengetahui preferensi masyarakat, dll.

### 2. Menentukan Populasi Target (Target Population)

Siapa atau apa yang menjadi keseluruhan objek yang akan diteliti?

Contoh: Semua mahasiswa di sebuah universitas, seluruh penduduk desa X, dll.

### 3. Menentukan Kerangka Sampel (Sampling Frame)

Daftar lengkap dari semua elemen dalam populasi yang dapat dipilih sebagai sampel.

Contoh: Daftar nama mahasiswa, daftar pelanggan, data pemilih, dll.

### 4. Menentukan Metode Sampling

Pilih apakah akan menggunakan:

Probability sampling (acak, sistematis, stratifikasi, cluster)

Non-probability sampling (kuota, purposive, convenience, snowball)

## **Resampling**

**Resampling** adalah teknik dalam statistik dan data science yang digunakan untuk **mengambil sampel ulang dari data yang sudah ada** untuk mengevaluasi, mengestimasi, atau meningkatkan model statistik tanpa perlu mengumpulkan data baru.

Berikut adalah bbrp cara utk mengatasi imbalance dataset:

Gunakan pengukuran (metrik) yang tepat, misal dengan menggunakan:

- Precision/Spesifikasi: berapa banyak instance yang relevan • Recall/Sensitifitas: berapa banyak instance yang dipilih
- F1 score: harmonisasi mean dari precision dan recall
- MCC: koefisien korelasi antara klasifikasi biner antara observasi vs prediksi
- AUC: relasi antara tingkat true-positive vs false-positive
- Resample data training, dengan dua metode:
  - Undersampling: menyeimbangkan dataset dengan mereduksi ukuran kelas yang melimpah. Dilakukan jika kuantitas data mencukupi
  - Oversampling: Kebalikan dari undersampling, dilakukan jika kuantitas data tidak mencukupi

## **Manfaat dan jenis seleksi Fitur Manfaat:**

Reduksi Overfitting: semakin kecil data redundant maka keputusan berdasarkan noise semakin berkurang

Meningkatkan Akurasi: semakin kecil data misleading maka akurasi model lebih baik  
Waktu Training: semakin kecil titik data (data point) maka kompleksitas algoritma berkurang dan latih algoritma lebih cepat

## **Jenis:**

Unsupervised: metode yang mengabaikan variabel target, seperti menghapus variabel yang berlebihan menggunakan korelasi

Supervised: metode yang menggunakan variabel target, seperti menghapus variabel yang tidak relevan

## **Latihan Soal**

1. Mengapa teknik sampling penting dalam analisis Big Data?
  - A. Untuk memperbesar ukuran data
  - B. Untuk membuat data menjadi lebih kompleks
  - C. Untuk mengurangi biaya dan waktu pemrosesan data
  - D. Untuk menghapus data yang tidak penting
  - E. Untuk menyimpan data dalam format visual
2. Sampling data dilakukan karena alasan berikut, kecuali:
  - A. Tidak semua data bisa diproses sekaligus
  - B. Dapat meningkatkan efisiensi komputasi
  - C. Memungkinkan prediksi dari populasi besar
  - D. Selalu memberikan hasil yang 100% akurat
  - E. Mengurangi waktu analisis
3. Manakah dari berikut ini yang merupakan metode sampling acak sederhana (simple random sampling)?
  - A. Memilih setiap anggota ke-n dari populasi
  - B. Membagi populasi berdasarkan kelompok
  - C. Menggunakan pengocokan acak atau generator angka acak
  - D. Memilih berdasarkan kategori
  - E. Memilih berdasarkan peringkat
4. Metode sampling yang membagi populasi menjadi kelompok homogen sebelum dilakukan pengambilan sampel disebut:
  - A. Sistematik
  - B. Random
  - C. Cluster
  - D. Stratified
  - E. Snowball
5. Apa yang dimaksud dengan resampling dalam konteks statistik dan data science?

- A. Menghapus sebagian data dari populasi
  - B. Menggabungkan dua populasi berbeda
  - C. Mengambil ulang sampel dari sampel yang sudah ada
  - D. Menentukan ukuran sampel awal
  - E. Mengukur volume data
6. Teknik resampling yang sering digunakan untuk mengestimasi akurasi model prediktif adalah:
- A. Stratified Sampling
  - B. Bootstrapping
  - C. Cluster Sampling
  - D. Systematic Sampling
  - E. Quota Sampling
7. Dalam Big Data, sampling sangat berguna karena:
- A. Data yang sedikit lebih akurat
  - B. Menurunkan resolusi data
  - C. Volume data terlalu besar untuk diproses secara langsung
  - D. Semua data memiliki bobot sama
  - E. Hanya data penting yang disimpan
- 8. Metode *cluster sampling* cocok digunakan ketika:**
- A. Populasi homogen
  - B. Akses ke seluruh populasi terbatas
  - C. Semua data harus dicatat
  - D. Ukuran populasi sangat kecil
  - E. Hanya satu kelompok yang dianalisis
9. Teknik sampling sistematis dilakukan dengan cara:
- A. Memilih secara acak dari setiap strata
  - B. Mengambil setiap elemen ke-n dari daftar populasi
  - C. Mengambil seluruh data tanpa seleksi
  - D. Membagi populasi berdasarkan karakteristik

E. Menentukan bobot untuk setiap elemen

10. Salah satu risiko dalam menggunakan sampling data pada big data analytics adalah:

- A. Semua hasil menjadi 100% akurat
- B. Data bisa menjadi lebih besar dari populasi
- C. Kesalahan estimasi jika sampel tidak representatif
- D. Komputasi menjadi terlalu cepat
- E. Semua data menjadi bersifat kategorikal

## Daftar Pustaka

Ardilla, Y., Guntoro, Afnarius, S., Santoso, A. B., Azdy, R. A., Putra, R., . . . Arnita. (2022). *DATA SCIENCE*. Bandung: Widina Bhakti Persada.

Cambell, A. (2021). *Data Science for Beginners*. Independently published.

Gupta, P. (2024). Securing Tomorrow: The Intersection of AI, Data, and Analytics in Fraud Prevention. Asian Journal of Research in Computer Science, 17(3), 75–92.  
<https://doi.org/10.9734/ajrcos/2024/v17i3425>

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining : Concepts and Techniques* Third Edition. Elsevier Inc.

Kuncoro, B. A. (2023). *Lima Dasar Data Science untuk Pemula*. Pustaka Media Guru.

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*. O'Reilly Media

Rasmussen,C,E. Williams C,K,I,, (2006). Gaussian Processes for Machine Learning. MIT Press. Cambridge, 2006.

Ramdani, F., & Utami, I. Q. (2024). *Pengantar Data Science*. Bumi Aksara.

Santoso,Joseph. (2023). Ilmu Data : Data Scaince. Yayasan Prima Agus Teknik bekerja sama dengan Universitas Sains & Teknologi Komputer (Universitas STEKOM), Semarang

Veronika S. Moertini, Mariskha T. Adithia. Pengantar Data Science Dan Aplikasinya Bagi Pemula. Bandung: Unpar Press, Bandung Indonesia,2020

Wede,(2020). Data Science Adalah: Sejarah Data Science, Ilmu Baru yang Saat Ini Sedang Jadi Primadona. <https://dqlab.id/data-scientist-profesini-primadona-era-transformasi-digital>

What is Big Data, <https://cloud.google.com/learn/what-is-big-data?hl=en>