**Say Hi, to Data Science**

*Oh ya!*Karena pada dasarnya data science adalah ilmu yang bergelut dengan data dan algoritma, pada submateri Data Science dan Matematika, Anda akan mempelajari dasar-dasar matematika guna memahami penerapan matematika di data science.

Selain itu, di submateri Metodologi Data Science, Anda akan dikenalkan dengan CRISP-DM atau *Cross Industry Standard Process for Data Mining*agar dapat memahami berbagai jenis metodologi yang digunakan pada data science.

Kemudian, di submateri Keterampilan Data Scientist, Anda akan mengeksplorasi keterampilan yang dapat Anda persiapkan untuk menjadi seorang data scientist andal.

**Apa itu Data Science**

Di sini Anda akan diajak berkenalan dengan Data Science sesuai topik pembahasan kita yaitu *Say Hi, to Data Science*. Sebenarnya apa *sih*data science itu? Mengapa data science begitu penting hingga dijuluki “*Sexiest job of the 21st century*” oleh Harvard Business Review? Daripada penasaran, simak penjelasannya di bawah ini.

**Amazon** mengemukakan bahwa data science adalah wawasan yang sangat berguna untuk kebutuhan bisnis. Ilmu ini merupakan penggabungan prinsip dan praktik dari bidang matematika, statistik, *artificial intelligence,*dan *computer engineering* untuk menganalisis data dalam jumlah besar.

Menurut **Jose** **Antonio Ribeiro** pada publikasinya di*Big Data for Executives and Professionals*, data science adalah kegiatan ilmiah yang mencakup pemahaman bisnis, persiapan data, analisis eksplorasi, pemodelan, aplikasi komputasi, evaluasi wawasan, dan tindakan untuk menghasilkan hasil serta menerapkan pengetahuan statistik, matematika, dan komputasi.

Adapun definisi lain, menurut IBM (*International Business Machine*) menyebutkan bahwa data science adalah menggabungkan matematika dan statistik, pemrograman, analitik, *artificial intelligence*, dan *machine learning*.

Dari ketiga definisi di atas dapat disimpulkan bahwa intinya, data science ini meliputi disiplin ilmu matematika, komputasi, dan sebagian strategi bisnis.

Disiplin ilmu ini berfokus pada penggunaan teknik matematika dan algoritma untuk memecahkan beberapa masalah bisnis sebuah organisasi atau perusahaan yang paling kompleks secara analitis, memanfaatkan kumpulan data mentah untuk mengetahui *insight*(wawasan) yang terdapat dalam data tersebut.

Ya! Jika menjadi ahli di bidang data science, harus memiliki keahlian dalam beberapa disiplin ilmu sekaligus. Cukup kompleks, tetapi tetap menarik, bukan? Itulah alasannya keahlian data scientist sangat diminati dan menjadi salah satu keputusan karier terbaik yang dapat Anda putuskan.

* **Data Science vs. Data Scientist**  
  Meskipun memiliki nama yang mirip, tetapi keduanya adalah dua hal yang berbeda. Data Science merupakan disiplin ilmu, sedangkan data scientist adalah praktisi dalam bidang data science. Sederhananya, data scientist adalah orang yang mengerjakan data science.
* **Data Scientist vs. Data Engineer**  
  Data engineer merupakan software engineer yang tugas utamanya adalah menyiapkan data yang dikoleksi dari berbagai sumber informasi dengan membangun sistem data atau infrastruktur agar dapat mudah untuk dianalisa dan mendukung kebutuhan perusahaan. Pada praktiknya, data scientist tidak meng-*handle*semua pekerjaan tentang data, tetapi terbagi menjadi beberapa bagian.
* **Data Scientist vs. Data Analyst**  
  Selain bekerja sama dengan data engineer, data scientist juga memiliki tanggung jawab yang tumpang tindih dengan data analyst. Data analyst adalah profesi pekerjaan yang bertugas untuk membaca serta menganalisis data yang ada dari perusahaan. Namun, keahlian seorang data scientist umumnya lebih luas dibandingkan data analyst. Secara komparatif, data science biasanya menggunakan R dan Python sebagai bahasa pemrograman. Nah, untuk mempelajari bahasa pemrograman tersebut seorang data scientist memerlukan ilmu komputer dan keterampilan sains di luar kemampuan analisis bisnis atau analisis data pada umumnya.

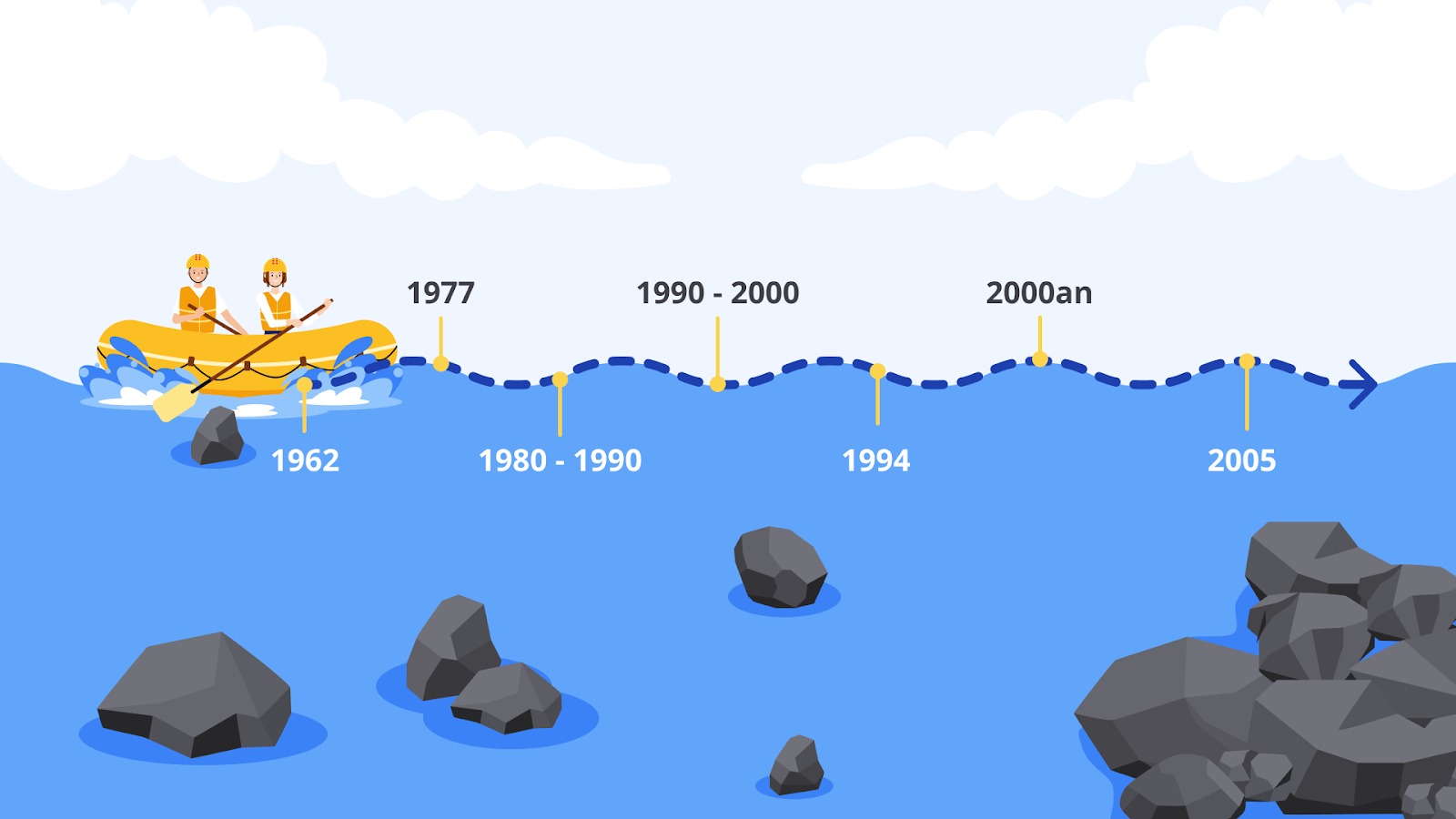
**Note:** Perbedaan terkait data analyst, data scientist, dan data engineer akan dibahas lebih mendalam lagi di modul lain pada kelas ini. *So*, *stay tune*sampai akhir, ya.

**Mengapa Data Science Begitu Penting?**

Sederhananya data science membantu organisasi atau perusahaan mendapatkan *insight*(wawasan) tentang bisnis mereka sendiri, tingkat keberhasilan strategi mereka, kinerja anggotanya, dan sebagainya.

Melalui data yang dikumpulkan, sebuah organisasi mampu memantau kinerja anggotanya dan mengambil langkah manajemen yang diperlukan.

**Evolusi Data Science di Dunia**



* **Tahun 1962**: **John W. Tukey**membuat artikel yang terkenal dengan judul “Masa Depan Analisis Data”, ia meramalkan munculnya bidang baru yang tak terelakkan hampir dua dekade sebelum lahirnya komputer atau PC (*Personal Computer*). Namun, saat Tukey meramalkan dan mengapresiasi yang sekarang kita sebut “Data Science”, ia tidak sendiri melainkan bersama rekannya yaitu Peter Naur, seorang insinyur komputer asal Denmark yang memiliki buku “*Concise Survey of Computer Methods”*yang berisi definisi data science pertama, yaitu “Ilmu yang berurusan dengan data, ketika ditetapkan, didelegasikan ke bidang dan ilmu lain”.
* **Tahun 1977**: Teori dan prediksi dari “sebelum” data scientist seperti Tukey dan Naur menjadi lebih konkret dengan adanya *The International Association for Statistical Computing* (IASC). Awalnya “mendelegasikan ke bidang dan ilmu lain” menjadi “informasi dan pengetahuan”.
* **Tahun 1980 - 1990**: Pada tahun ini, data science mulai mengambil langkah signifikan dengan munculnya *Knowledge Discovery in Databases*(KDD) pertama dan pendirian *International Federation of Classification Societies*(IFCS). KDD dan IFCS termasuk yang pertama berfokus pada mendidik dan melatih para profesional dalam teori dan metodologi ilmu data.
* **Tahun 1994**: *BusinessWeek*yaitu majalah bisnis terkemuka di Amerika, menerbitkan cerita tentang fenomena baru “*Database Marketing*”. Majalah tersebut menerbitkan proses bisnis mengumpulkan dan memanfaatkan big data untuk mempelajari lebih lanjut tentang pelanggan, persaingan, atau teknik periklanan.
* **Tahun 1990an - awal 2000an**: Tahun ini terlihat jelas bahwa data science telah muncul sebagai bidang yang diakui. Beberapa jurnal akademik data science mulai beredar, dan pendukung data science, seperti **Jeff Wu**dan **William S. Cleveland**terus membantu mengembangkan akan kebutuhan dan potensi data science.
* **Tahun 2000an**: Teknologi membuat batu loncatan dengan menyediakan akses yang universal seperti internet, komunikasi, dan pengumpulan data.
* **Tahun 2005**: Big Data mulai memiliki peran. Perusahaan besar seperti Google dan Facebook mengungkap data dalam jumlah besar, sehingga memerlukan teknologi baru yang mampu memprosesnya. Tantangan tersebut dijawab Hadoop dan kemudian Spark serta Cassandra tidak ingin kalah dalam debut big data.
* Tahun 2014: Semakin pentingnya data dan banyak organisasi/perusahaan yang berminat dalam menemukan pola serta membuat keputusan bisnis yang lebih baik, permintaan data scientist mulai terlihat pertumbuhannya di berbagai belahan dunia.
* Tahun 2015: Machine Learning, Deep Learning, dan Artificial Intelligence (AI) resmi memasuki ranah data science.
* Tahun 2018: Peraturan baru di lapangan menjadi salah satu aspek terbesar dalam evolusi data science.
* Tahun 2020an: Terdapat terobosan baru pada AI dan machine learning sehingga permintaan yang semakin meningkat untuk profesional yang berkualifikasi di Big Data.

**Dampak Data Science di Dunia**

Pada dasarnya, sebuah bisnis mempekerjakan data scientist untuk mengumpulkan, membersihkan, mengatur, dan menganalisis kumpulan data yang besar untuk memecahkan masalah yang terjadi pada organisasi atau perusahaan hingga mendapatkan *insight*(wawasan) untuk ditindak lanjuti. Namun, apakah sebenarnya Anda tahu dampak nyata dari adanya data science? Jika belum, simak penjelasan berikut.

1. **Promosi pelestarian lingkungan**. Tahukah Anda? Selain digunakan untuk membuat sebuah keputusan yang baik untuk organisasi atau perusahaan, data science memiliki peran penting dalam keberhasilan pengelolaan krisis iklim global, *lho*. PBB mengembangkan 17 tujuan pembangunan berkelanjutan atau *Sustainable Development Goals*(SDGs) untuk mengukur dan melacak bagaimana negara menyesuaikan faktor lingkungan, sosial, dan ekonomi untuk memperlambat perubahan iklim.

Contohnya, para ilmuwan di *The Ocean Cleanup* menggunakan data science untuk meneliti zona plastik lepas pantai terbesar di lautan. Mereka mengumpulkan data menggunakan metode komputasi dan matematis, menganalisis jutaan titik data untuk mengidentifikasi jenis limbah yang ada, dan memperkirakan total sampah plastik.

1. **Kemajuan bidang kesehatan**. Selama pandemi COVID-19, layanan kesehatan dan lembaga pemerintah menggunakan data analyst untuk melacak kasus, memproyeksikan penyebaran infeksi, menginformasikan keputusan kebijakan kesehatan, dan meningkatkan hasil layanan kesehatan.

Contohnya, *Johnson & Johnson*(J&J) mempekerjakan pakar data science untuk membuat dasbor pengawasan global yang memperkirakan titik persebaran yang memberi tahu pemerintah tentang cara penyakit itu menyebar.

1. **Pemberdayaan negara berkembang**. Setiap tahun, Bank Dunia menilai negara-negara dalam skala 1-100. Skala 1 mewakili kurangnya data statistik penting dan kapasitas analisis, serta skala 100 mewakili kapasitas statistik negara maju. Selama pandemi, 80% kantor statistik nasional di negara berpenghasilan rendah hingga menengah memerlukan dukungan tambahan untuk melakukan pengumpulan dan analisis data penting.  
     
   Misalnya, di Nigeria, data scientist menilai keakuratan, kelengkapan, konsistensi, dan kendala data untuk mengungkap cara memperbaiki proses pemilu dalam kebijakan pendaftaran memilih. Proyek tersebut mendemonstrasikan cara data scientist dapat memberdayakan pemerintah dan masyarakat untuk membangun sistem yang lebih baik menggunakan data.
2. **Peningkatan *customer insight.***Salah satu dampak paling menguntungkan yang dimiliki data science terhadap bisnis adalah mengevaluasi, memprediksi, dan memberikan rekomendasi yang akurat dan dipersonalisasi kepada pelanggan.

Contoh lainnya adalah Netflix. Siapa yang tidak kenal dengan platform penyedia layanan streaming terbesar di dunia? Perusahaan tersebut menggunakan data analytics dan machine learning untuk mengumpulkan insight secara real-time tentang kebiasaan penggunanya untuk menampilkan film dan acara TV berdasarkan minat para penggunanya.

1. **Mengembangkan pemimpin yang lebih cerdas.** Data scientist dan pemimpin bisnis telah bekerja sama untuk lebih memahami cara sebuah data dapat menjadi jembatan untuk menginformasikan strategi organisasi yang lebih baik, menginformasikan keputusan, dan mencapai hasil yang lebih baik.

Contoh pada kasus ini adalah Nancy Green, CEO dari Old Navy, ia menggunakan analisis prediktif untuk mengelola rantai pasokan (Supply Chain) agar lebih baik. CEO dari retailer pakaian tersebut berkata, “Kami melihat peluang besar di depan dengan menggunakan analisis prediktif untuk memperkirakan pasokan guna memenuhi permintaan pelanggan dengan lebih baik.”

Wah… Sungguh menarik, ya? Setelah mengetahui sekilas informasi tentang data science hingga dampak data science di dunia, Anda akan mempelajari hubungan antara data science dan matematika pada materi berikutnya. So, tunggu apa lagi? Mari kita menuju materi selanjutnya. See you there.

**Data Science dan Matematika**

Sebelum kita mempelajari hubungan data science dengan matematika, mari kita pahami dasar-dasar dari matematika dan kalkulus, lalu mempelajari korelasi antara data science dan matematika. Dengan mempelajari ini, diharapkan Anda mampu memahami konsep dari data science yang seyogyanya tidak luput dari data.

**Matematika Esensial Untuk Data Science**

*“Without mathematics, there’s nothing you can do. Everything around you is mathematics. Everything around you is numbers.”*

*-Shakuntala Devi, Indian writer and mental calculator-*

**Dasar Matematika**

Sebelum melangkah ke yang lebih besar, tentu Anda harus memiliki dasar pemahaman yang kuat. Pada materi ini Anda akan berkenalan dan belajar terkait beberapa materi di bidang matematika yang mencakup hal-hal dasar seperti variabel, fungsi, penjumlahan, persamaan, dan grafik yang tersusun dari materi berikut.

* Algoritma, eksponensial, fungsi polinomial, dan bilangan relasional
* Geometri dasar dan teorema, identitas trigonometri
* Bilangan riil dan kompleks
* Deret, penjumlahan, pertidaksamaan
* Grafik dan plotting, koordinat kartesius dan polar, serta irisan kerucut.

Namun, pada materi ini Anda hanya akan mempelajari beberapa hal-hal umum yang menjadi dasar dalam pemahaman matematika diantaranya variabel, fungsi, dan penjumlahan. Sudah tidak sabar? Yuk, mari simak penjelasannya di bawah ini.

**Variabel**

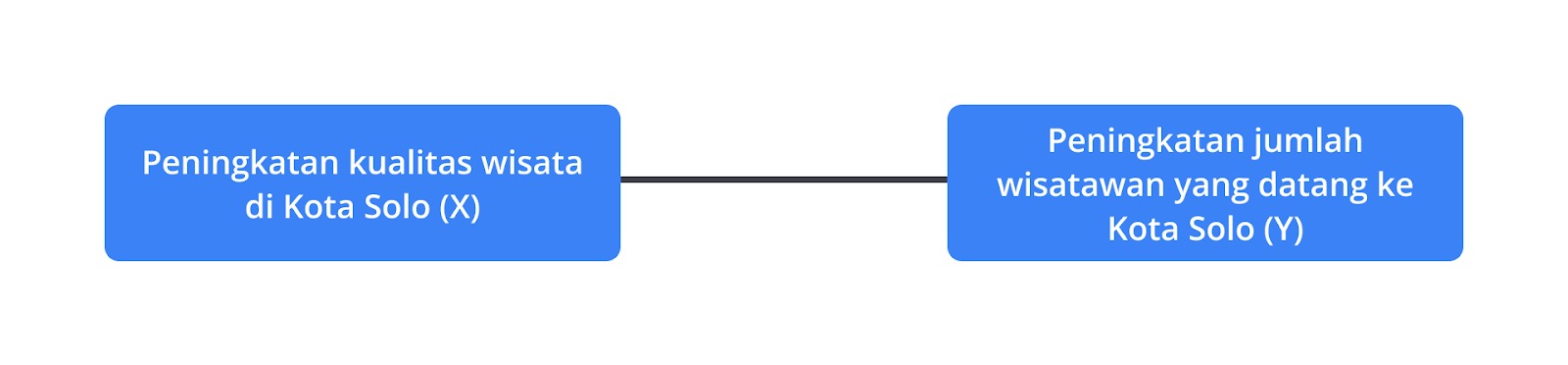
Menurut*Australian Bureau of Statistics*, variabel adalah karakteristik, angka, atau kuantitas apa pun yang dapat diukur atau dihitung. Variabel juga dapat disebut item data. Usia, jenis kelamin, pendapatan dan pengeluaran bisnis, negara kelahiran merupakan contoh variabel dan masih banyak lagi contoh-contoh lainnya.

Secara total, ada tiga jenis variabel di bidang data science yang tercantum di bawah ini.

* **Numerikal**  
  Apakah Anda masih ingat dengan jenis-jenis data? Sama seperti sebelumnya, numerikal merupakan bagian dari data kuantitatif, jenis data ini terbagi menjadi dua sub-kategori yaitu kontinu dan diskrit.
* **Kategorikal**  
  Masih ingatkah Anda dengan data kategorikal yang sudah pernah kita bahas sebelumnya? Jadi, kategorikal merupakan jenis data kualitatif. Jenis data ini terbagi menjadi dua sub-kategori yaitu nominal dan ordinal.
* **DateTime**  
  Kategori variabel ini berkaitan dengan aspek tanggal & waktu. Kategori ini dapat berisi jenis nilai seperti berikut.
  + Hanya memiliki tanggal.
  + Hanya memiliki waktu.
  + Memiliki keduanya.

Fungsi

Ekspresi yang mendefinisikan hubungan antara dua atau lebih variabel. Sebagian besar data science difokuskan untuk menemukan hubungan antara variabel dependen dan independen. Sederhananya, variabel independen adalah variabel bebas (x) dan variabel dependen adalah variabel terikat (y) yang dipengaruhi oleh variabel bebas.



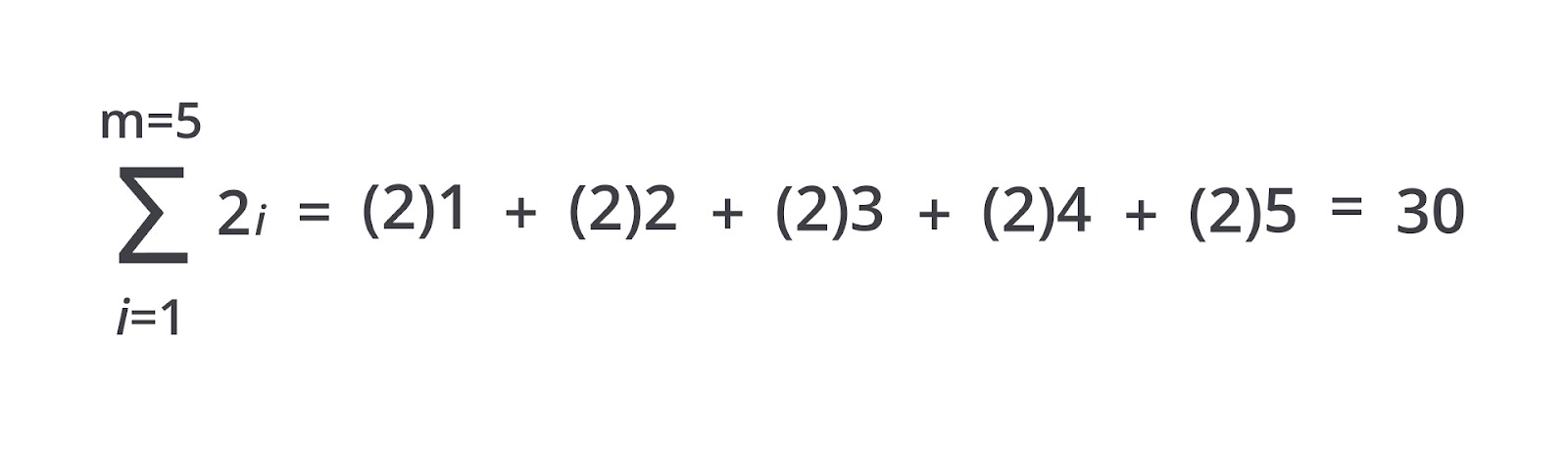
Gambar di atas menunjukan hubungan antara variabel X sebagai variabel bebas dan variabel Y sebagai variabel terikat. Dari gambar di atas dapat dibaca sebagai berikut.

**Peningkatan jumlah wisatawan yang datang ke Kota Solo (Y) dapat dipengaruhi oleh peningkatan kualitas wisata di Kota Solo (X)**.

Seperti definisinya bahwa variabel Y akan berubah apabila terdapat nilai atau perubahan dari variabel X.

**Penjumlahan**

Penjumlahan umumnya dinyatakan sebagai sigma (). Nah, agar Anda dapat lebih memahami terkait pembahasan penjumlahan, perhatikan contoh di bawah ini.



**Kalkulus**

Sebagian besar bidang data science tentunya memerlukan pemahaman tentang prinsip dasar kalkulus dan pengaruhnya terhadap model *machine learning.*Namun, perlu digaris bawahi bahwa kalkulus pada data science tidak seperti kalkulus pada kelas matematika di sekolah atau perguruan tinggi. Umumnya beberapa konsep kalkulus yang dapat digunakan data scientist adalah *Multivariate calculus*(Kalkulus multivariabel) dan Gradient descent (penurunan Gradien).

1. **Multivariate calculus**  
   *Multivariate calculus* adalah bidang kalkulus yang melibatkan banyak variabel. Jika output dari fungsi Anda z, bergantung pada satu variabel input (x). Catatan! Sebagian besar algoritma machine learning dilatih pada banyak fitur (variabel) oleh karena itu pemahaman tentang cara kerja multivariate calculus sangat penting untuk para pemula.

Berikut merupakan fungsi Multivariate calculus pada machine learning.

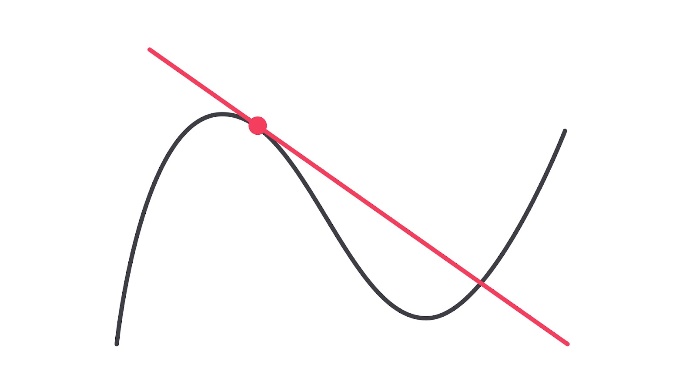
Dalam algoritma support vector (Support vector algorithm), multivariate calculus digunakan untuk menemukan margin maksimal.

Masalah pengoptimalan (optimization problems) bergantung pada multivariate calculus.

1. **Gradient descent**  
   *Gradient descent* digunakan untuk menemukan minimal atau maksimal dari suatu fungsi. Fungsi tersebut bisa menjadi *cost function*dari algoritma *machine learning.* Prinsip dari *gradient descent*apabila dibuat analoginya adalah Anda seperti melemparkan bola dari permukaan ke atas (kurva naik atau titik tertinggi) maka bola itu akan kembali lagi ke bawah (kurva turun atau titik minimum). *Gradient descent*(penurunan gradien) menemukan laju perubahan variabel dan menyesuaikan untuk bergerak menuju titik minimum. *Gradient descent*(penurunan gradien) menemukan laju perubahan variabel dan menyesuaikan untuk bergerak menuju titik minimum.

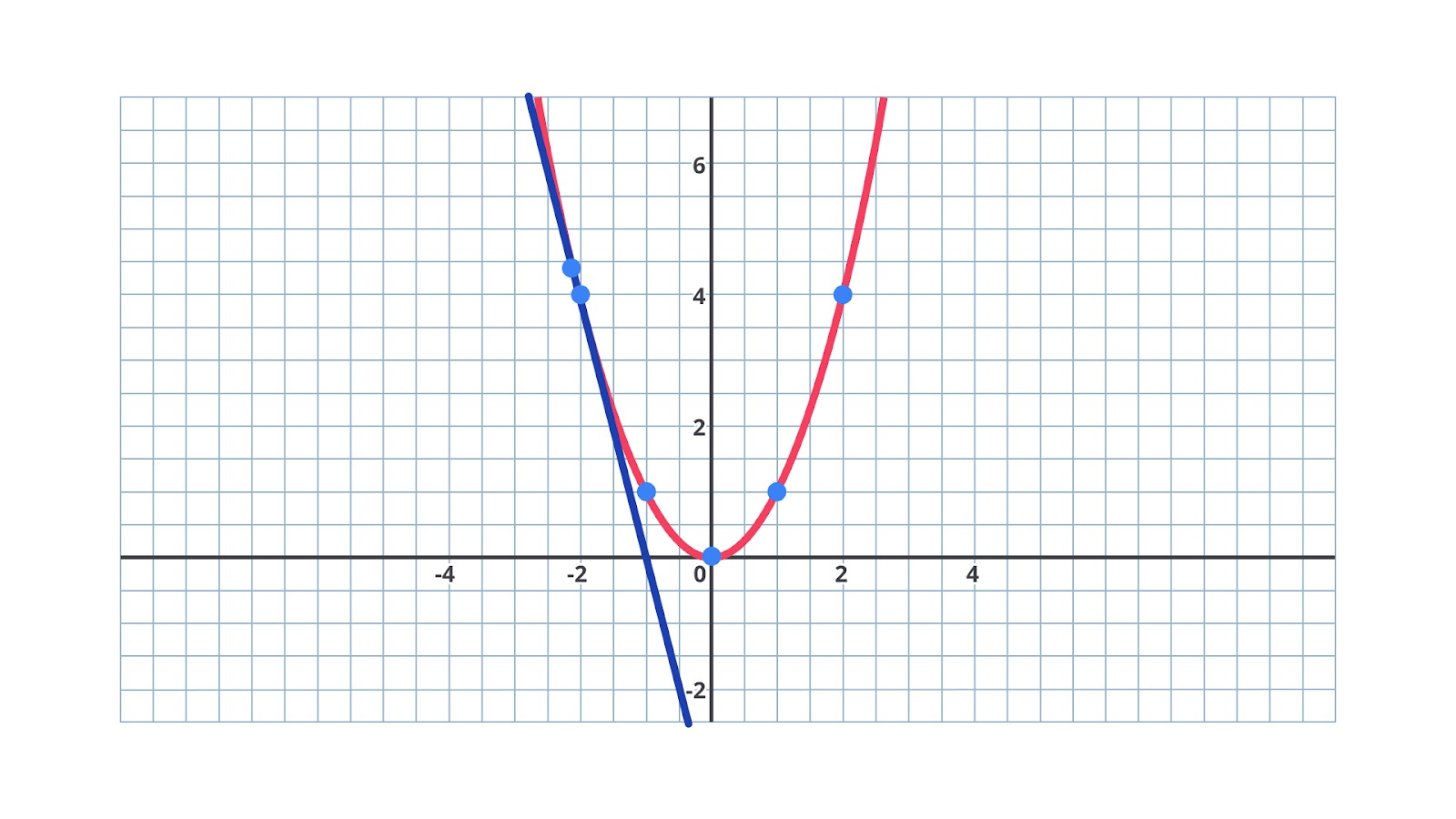
**Turunan**

Dalam matematika, derivatif atau turunan adalah laju perubahan fungsi terhadap variabel. Turunan berasal dari garis kemiringan yang bersinggungan dengan grafik fungsi.

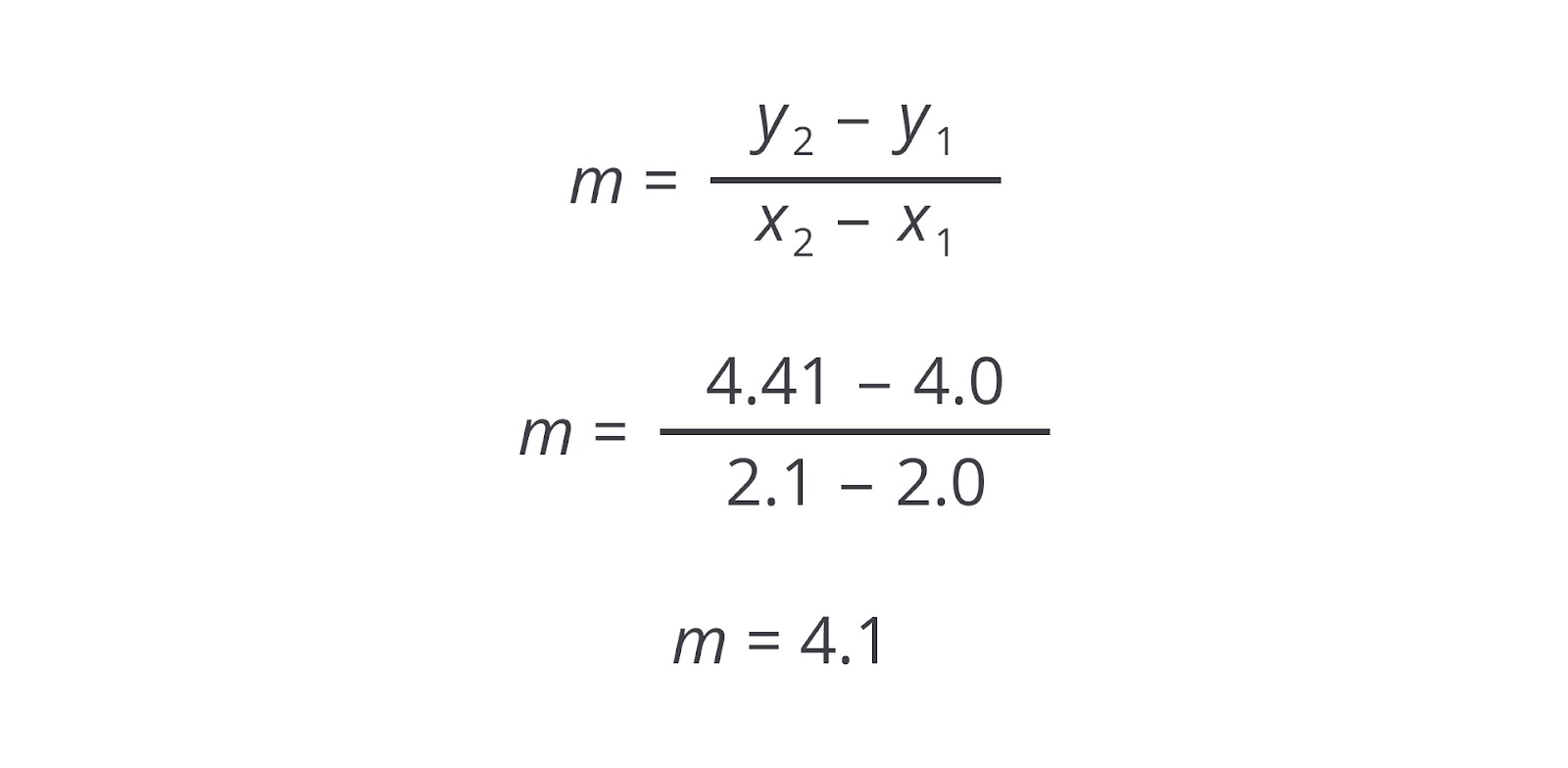


Grafik fungsi yang berwarna hitam dan garis tangen (garis singgung) yang berwarna merah. Kemiringan dari garis tangen sama dengan turunan fungsi pada titik tersebut.

Untuk mengetahui lebih lanjut penerapan fungsi untuk menjadi sebuah grafik, mari kita mulai dengan contoh fungsi sederhana. Umumnya, sebuah kurva suatu fungsi dapat digambar dengan menganalisis beberapa konsep turunan, yaitu fungsi naik atau turun, titik optimum (maksimum atau minimum), titik stasioner, dan titik belok. Seperti yang dijelaskan pada *gradient descent*di atas, bahwa prinsip dari fungsi tersebut adalah mencari laju perubahan variabel yang tertera seperti gambar di bawah ini.



Dari grafik di atas dapat dilihat bahwa saat kurva naik berada di titik maksimal 4.41 dan kurva turun berada di titik minimum 0. Setelah Anda mendapatkan nilai f(x), Anda dapat menghitung kemiringan dari nilai maksimum dengan cepat menggunakan persamaan berikut.



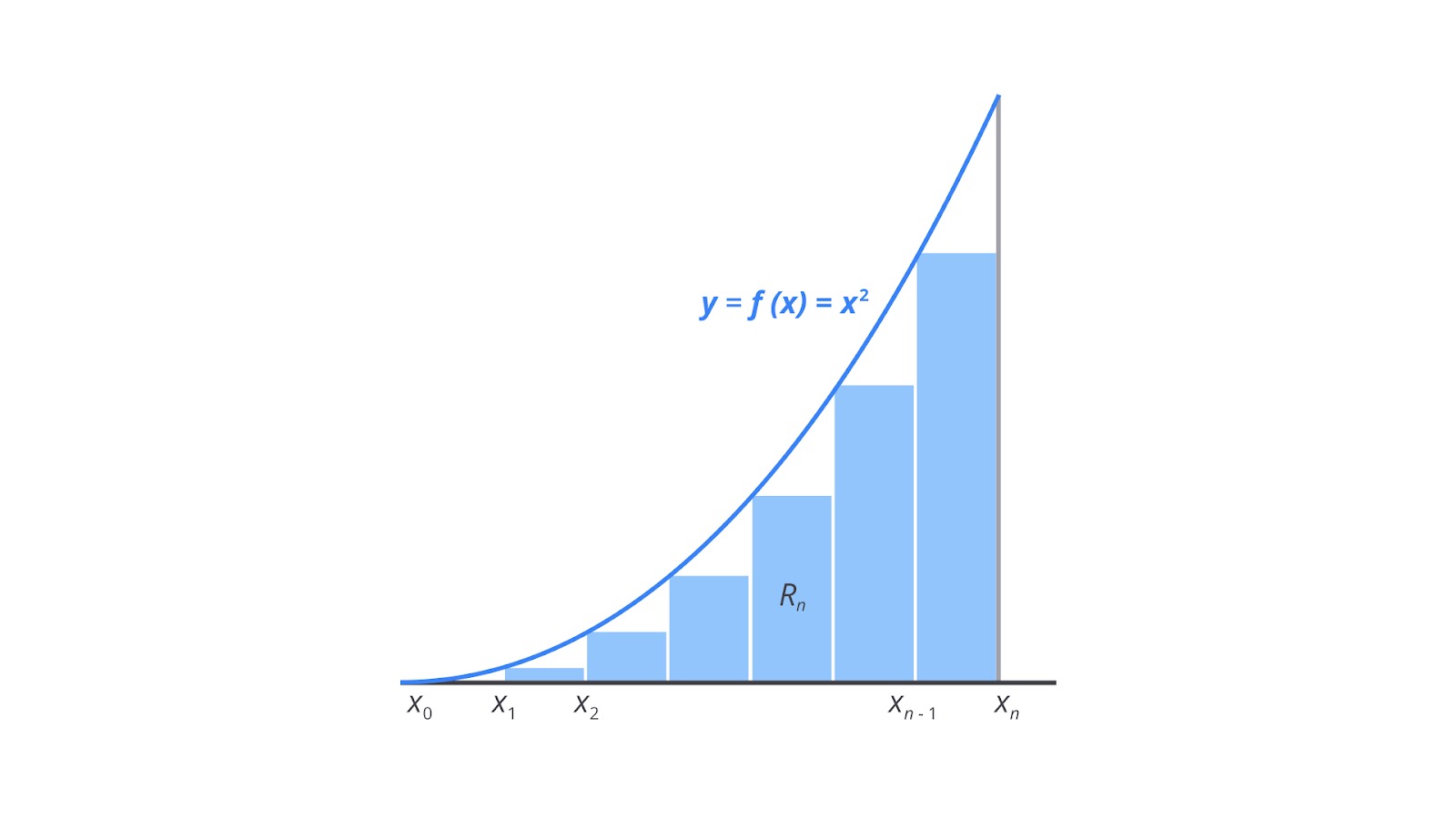
Turunan atau derivatif digunakan oleh *machine learning* untuk memecahkan masalah pengoptimalan. Algoritma pengoptimalan seperti gradient descent menggunakan turunan untuk memutuskan bertambah atau berkurangnya parameter training.

Untuk melatih pemahaman dan kemampuan Anda terkait materi turunan, berikut merupakan latihan variabel yang dapat Anda coba secara mandiri, lalu bandingkan hasilnya dengan jawaban yang telah tersedia. *Good luck!*

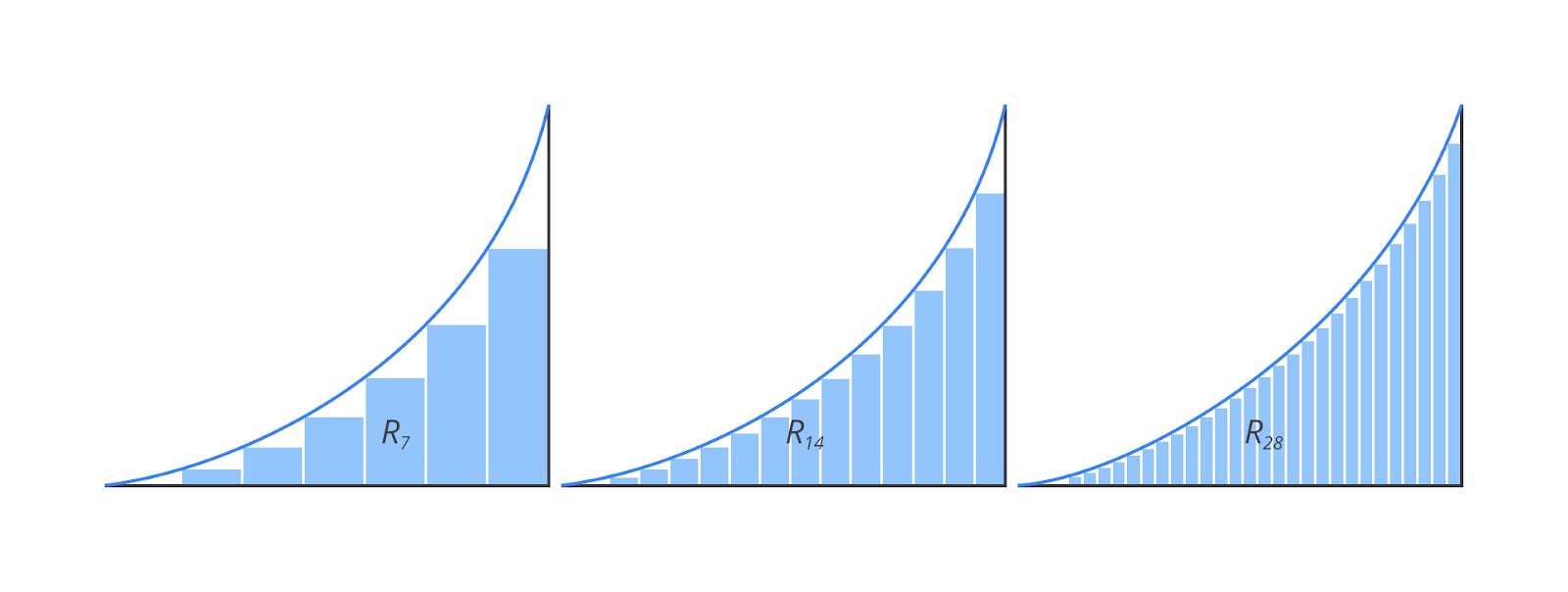
**Integral**

Apabila sebelumnya turunan merupakan suatu perhitungan terhadap perubahan nilai fungsi karena perubahan nilai input, maka secara sederhananya integral merupakan kebalikan dari operasi turunan.

Fungsi dari adanya integral adalah untuk mengetahui cara integral mencari luas di bawah kurva.



Luas area tersebut hampir mendekati luas dari total 6 persegi panjang. Jika jumlah persegi panjang tersebut terus diperbanyak hingga mendekati tak terhingga, luas dari seluruh persegi panjang akan sama dengan luas daerah R. Dengan kata lain, luas dari area R sama dengan luas dari seluruh persegi panjang yang banyaknya tak terhingga. Gambar di bawah ini merupakan contoh apabila Anda membagi integral menjadi bagian yang lebih banyak. Maka dari itu semakin banyak persegi panjang yang dibuat, semakin akurat pula luas dari bawah kurva tersebut.



Namun, pada modul ini, kita hanya akan membahas integral sederhana. Dari persamaan fungsi di atas, batas bawah x=a dan batas atas x=b dapat ditentukan dengan mengintegralkan fungsi tersebut pada interval a ≤ x ≤ b.

**Statistika**

Jika berbicara tentang statistika, Anda mungkin menyangka istilah tersebut sama dengan statistik, padahal keduanya adalah hal yang berbeda. Simak perbedaannya dalam tabel berikut.

* **Statistik :** Sebuah kumpulan data, angka, atau informasi.
* **Statistika :** Ilmu yang mempelajari data atau angka tersebut dikumpulkan, diolah, dan dianalisis untuk menghasilkan sebuah informasi yang bisa digunakan dalam pengambilan keputusan.

Sejauh ini, statistika adalah bagian penting dari matematika yang perlu Anda kuasai untuk data science. Statistika adalah cabang matematika yang mengumpulkan data dan menganalisis kumpulan data besar untuk menginterpretasikan wawasan yang bermakna. Data scientist menggunakan statistika sebagai berikut.

* Kumpulkan (*collect*), tinjau (*review*), analisis (*analyze*), wawasan dari data (*insight from data*).
* Identifikasi dan analisis pola data menjadi *insight*(wawasan) bisnis yang dapat ditindaklanjuti.
* Jawab pertanyaan dengan membuat eksperimen, menganalisis, dan menafsirkan kumpulan data.
* Memahami *machine learning* dan model prediksi (*predictive model*).

Pada kelas Dasar Data Science ini, Anda akan berkenalan dengan statistika deskriptif dan inferensial. Pernah mendengar kah sebelumnya? Jika belum, yuk kita langsung ke materinya.

**Deskriptif**

Statistika deskriptif adalah metode yang meringkas kumpulan data tertentu, dapat berupa representasi dari seluruh populasi atau sampel dari suatu populasi. Statistika deskriptif terbagi menjadi dua ukuran, yaitu ukuran tendensi sentral dan ukuran variabilitas.

* Tendensi sentral: Rata-rata, median, dan modus.
* Variabilitas: Range, varians,  standar deviasi, variabel minimum, dan maksimum, kurtosis, dan skewness.

Sederhananya, statistika deskriptif membantu menjelaskan dan memahami fitur kumpulan data tertentu dengan memberikan ringkasan singkat tentang sampel dan ukuran data.

**Tendensi Sentral :** Berfokus pada nilai rata-rata, median, dan modus

**Variabilitas :** Fokus pada penyebaran data.

**Tendensi sentral**

Seperti yang tertera pada tabel di atas, tendensi sentral lebih berfokus pada nilai rata-rata, median, dan modus.

**Variabilitas**

Variabilitas menjelaskan seberapa jauh titik data terpisah satu sama lain dan dari pusat distribusi karena contoh dari variabilitas sangat beragam. Oleh karena itu, pada kelas ini Anda hanya akan dikenalkan secara singkat beberapa jenis variabilitas, yaitu *range, varians,*dan*standar deviasi*.

* Range: jangkauan suatu nilai.
* Varians: penyebaran angka dalam kumpulan data.
* Standar deviasi: ukuran penyebaran dalam kumpulan data.

**Inferensial (*Inference*)**

Statistika inferensial merupakan teknik analisis data yang digunakan untuk menentukan kesamaan antara hasil yang diperoleh dari suatu sampel dengan hasil yang akan didapat pada populasi secara keseluruhan. Dengan demikian, statistika inferensial membantu peneliti untuk mencari tahu hasil yang diperoleh dari suatu sampel dapat digeneralisasi pada populasi.

Terdapat banyak sekali contoh aplikasi dan penerapan statistika inferensial dalam kehidupan. Namun, secara umum, statistika inferensial yang sering digunakan, yaitu.

1. **Analisis Regresi.**  
   Digunakan untuk memprediksi hubungan antara variabel independen dengan variabel dependen.  
   Contoh: Anda merupakan data scientist yang ingin mengetahui faktor yang memengaruhi minat wisatawan berlibur ke Bandung. Ia menguji hal tersebut dengan menggunakan variabel seperti keindahan, sarana dan prasarana, keindahan, sumber objek wisata, dll.
2. **Pengujian Hipotesis**  
   Pengujian ini merupakan pengujian statistika ketika kita ingin mengetahui kebenaran dari sebuah asumsi.  
   Contoh: Wanita cenderung lebih suka warna merah muda dibandingkan warna hitam. Benarkah demikian?  
     
   Nah, untuk membuktikan hal tersebut maka Anda bisa mengambil sebagian sampel dan melakukan pengujian terhadap sampel yang telah diambil. Terdapat beberapa metode testing yang umum digunakan seperti Z-test, T-test, dll.

**Aljabar Linier**

Aljabar Linier adalah salah satu keterampilan matematika yang banyak digunakan dalam beberapa bagian, diantaranya pada sebagian besar *machine learning* dapat diekspresikan dalam bentuk matriks, kemudian pada deep learning ketika membuat artificial neural network (jaringan syaraf tiruan). Data Scientist memanipulasi dan menganalisis data mentah melalui matriks, baris, dan kolom angka atau titik data.

Misalnya, aljabar linier adalah komponen inti dari pre-processing data. Maka, aljabar linier adalah proses mengatur data mentah sehingga dapat dibaca dan dipahami oleh mesin.

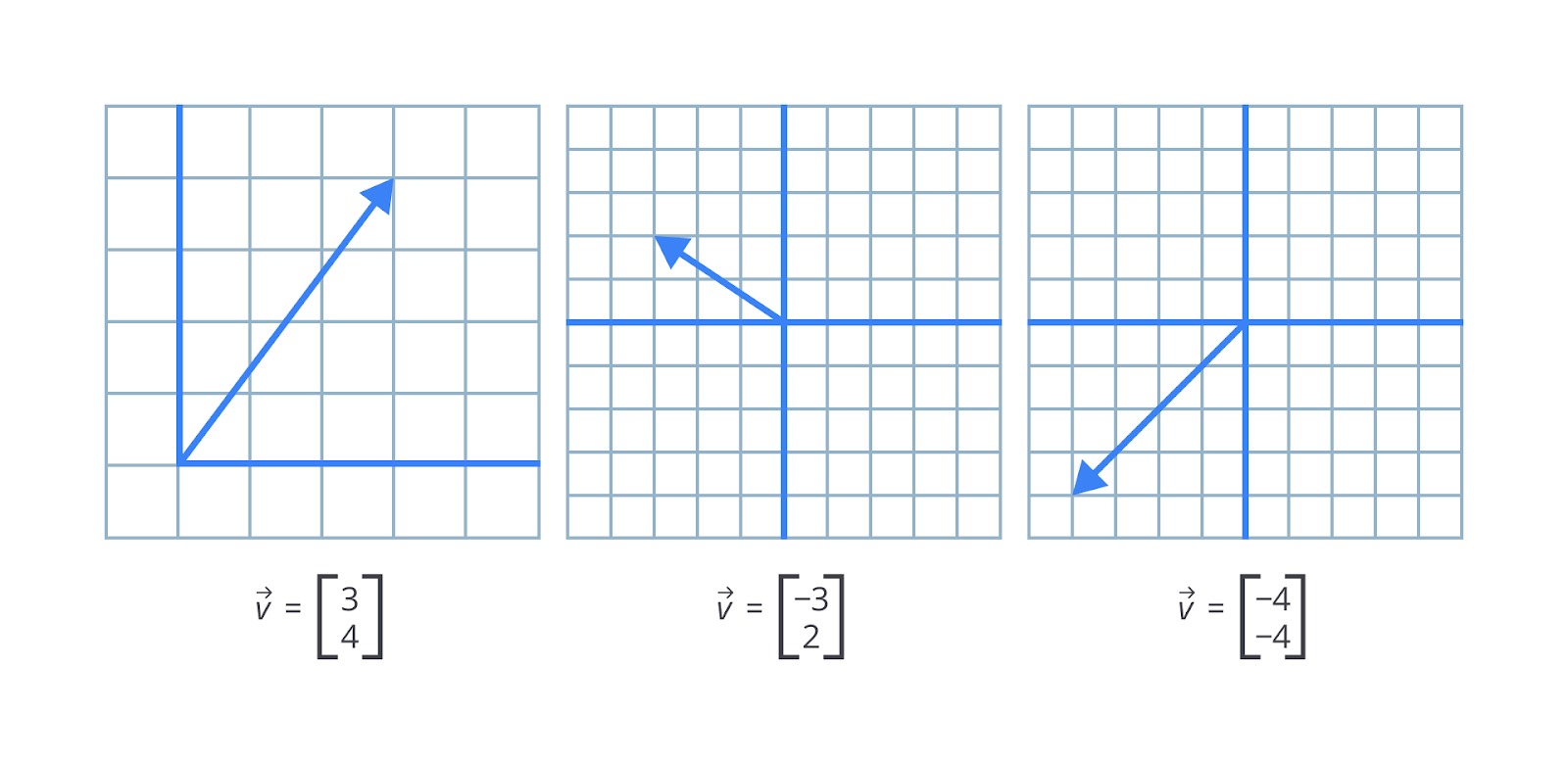
Minimal, seorang data scientist harus mengetahui vektor dan matriks, serta cara untuk menerapkan prinsip aljabar linier untuk menyelesaikan masalah data yang akan kita pelajari berikutnya. Yuk, simak penjelasannya di bawah ini.

**Vektor**

Sederhananya, vektor adalah besaran yang memiliki nilai dan arah. Vektor merupakan hal umum yang digunakan dalam *machine learning*untuk merepresentasikan data dengan cara yang paling optimal dan terorganisir. Tentunya, vektor memiliki beberapa dimensi dengan tujuan tertentu tetapi pada dasarnya, vektor memiliki dua dimensi yang mana hanya positif dan negatif untuk menunjukan sebuah arah. Pada pembelajaran ini, Anda akan dikenalkan dengan konsep dasar dari vektor yaitu dua dimensi dan tiga dimensi untuk mengenalkan bagaimana cara vektor bekerja.

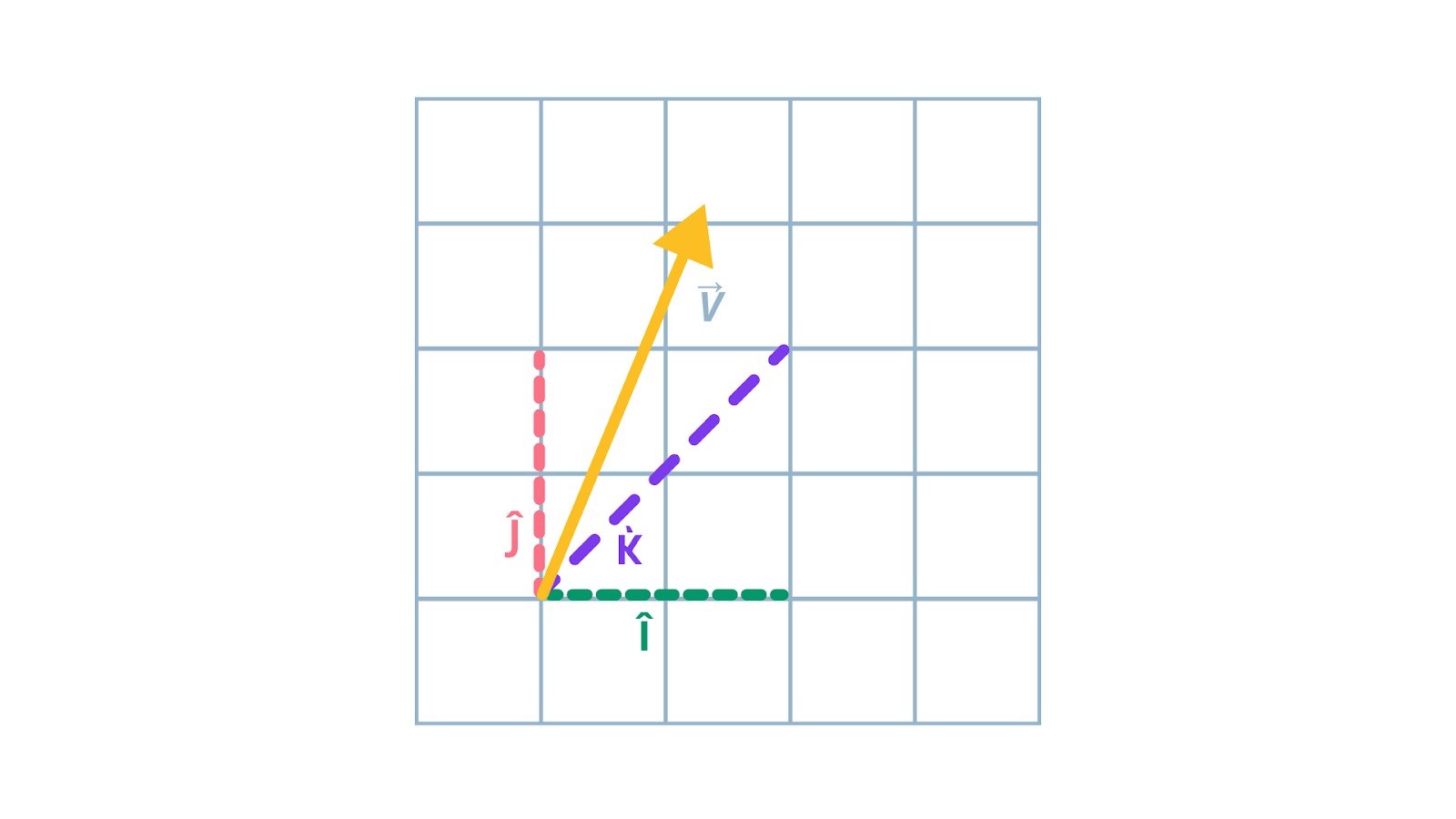
**Vektor Dua Dimensi**

Sederhananya, vektor adalah panah dalam ruang dengan arah dan panjang tertentu, seringkali mewakili sepotong data.



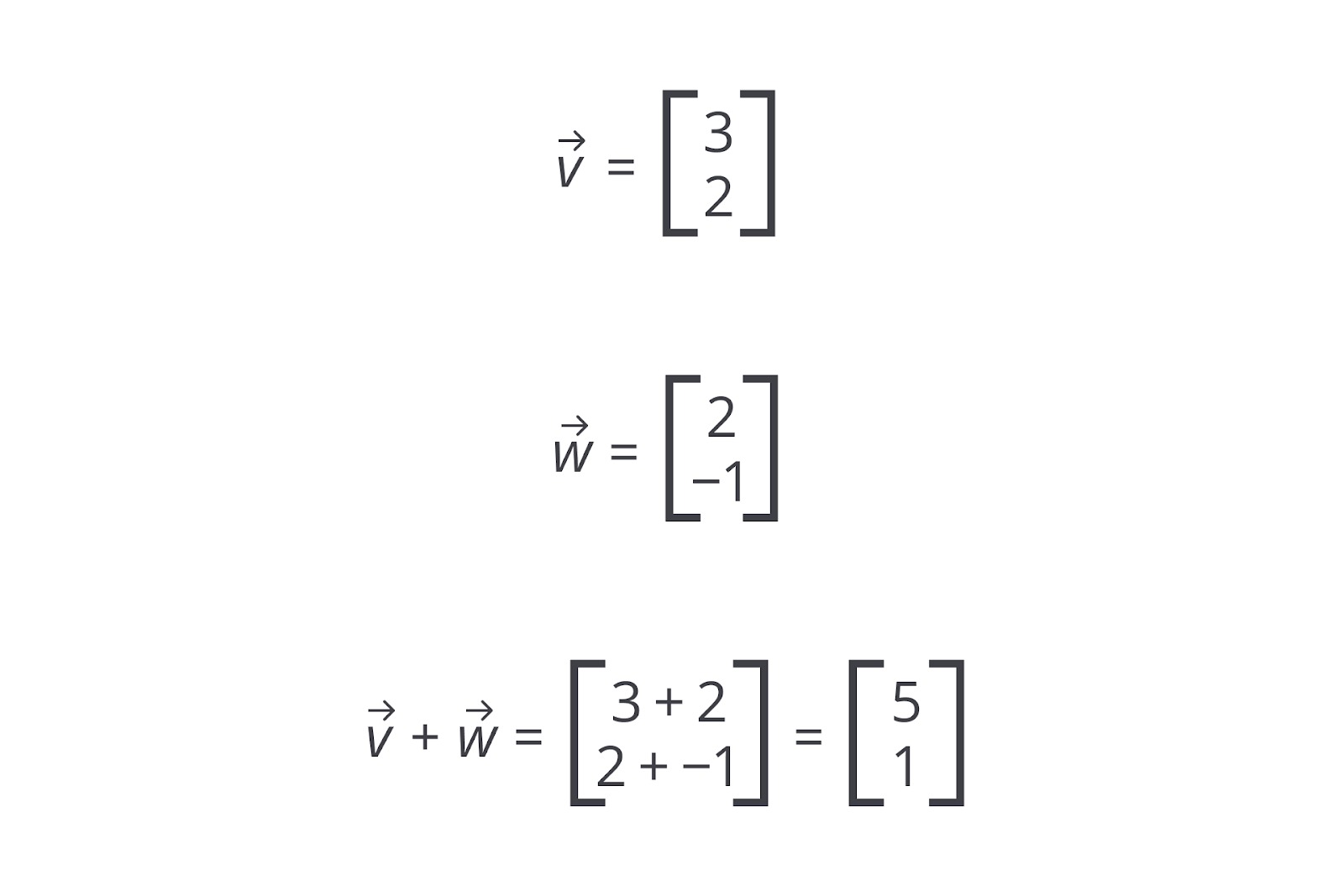
**Vektor Tiga Dimensi**

Selain vektor dua dimensi, ada pula vektor tiga dimensi seperti berikut.



**Menambahkan dan Mengombinasikan Vektor**

Dari dua dimensi dan tiga dimensi vektor yang telah Anda pelajari di atas, di bawah ini Anda akan mempelajari cara memenjumlahkan apabila terdapat dua atau lebih vektor dijumlahkan atau dikurangi.



Ketika terdapat dua jenis vektor yang menggunakan operasi penjumlahan dan pengurangan maka cara matematis berhitungnya adalah seperti yang tertera pada gambar di atas.

**Tahukah Anda?**

Mengapa Anda harus belajar vektor? Mengapa vektor sangat penting di data science? Pada dasarnya saat menggunakan statistik dan *machine learning,*data sering diimpor dan diubah menjadi vektor numerik sehingga Anda dapat bekerja dengannya.

**Matriks**

Setelah sebelumnya membahas vektor. Nah, bagian ini Anda akan mempelajari terkait matriks, yaitu vektor yang dibawa ke tingkat berikutnya. Ya, betul! Anda masih akan bergelut dengan (*another*) vektor. Perlu Anda ketahui bahwa matriks adalah struktur data dasar yang digunakan untuk menyimpan data dan menyelesaikan sistem persamaan linier. Perlu Anda ketahui bahwa matriks sangat berguna, *lho.*Ia dapat menyimpan kumpulan persamaan, posisi artikel dari waktu ke waktu, catatan keuangan, dan masih banyak hal lainnya. Dalam data science, matriks terkadang disebut tabel data yang terdiri dari baris dan kolom atau baris matriks dan kolom matriks. Matriks dapat berupa bilangan real, kompleks, bahkan boolean.

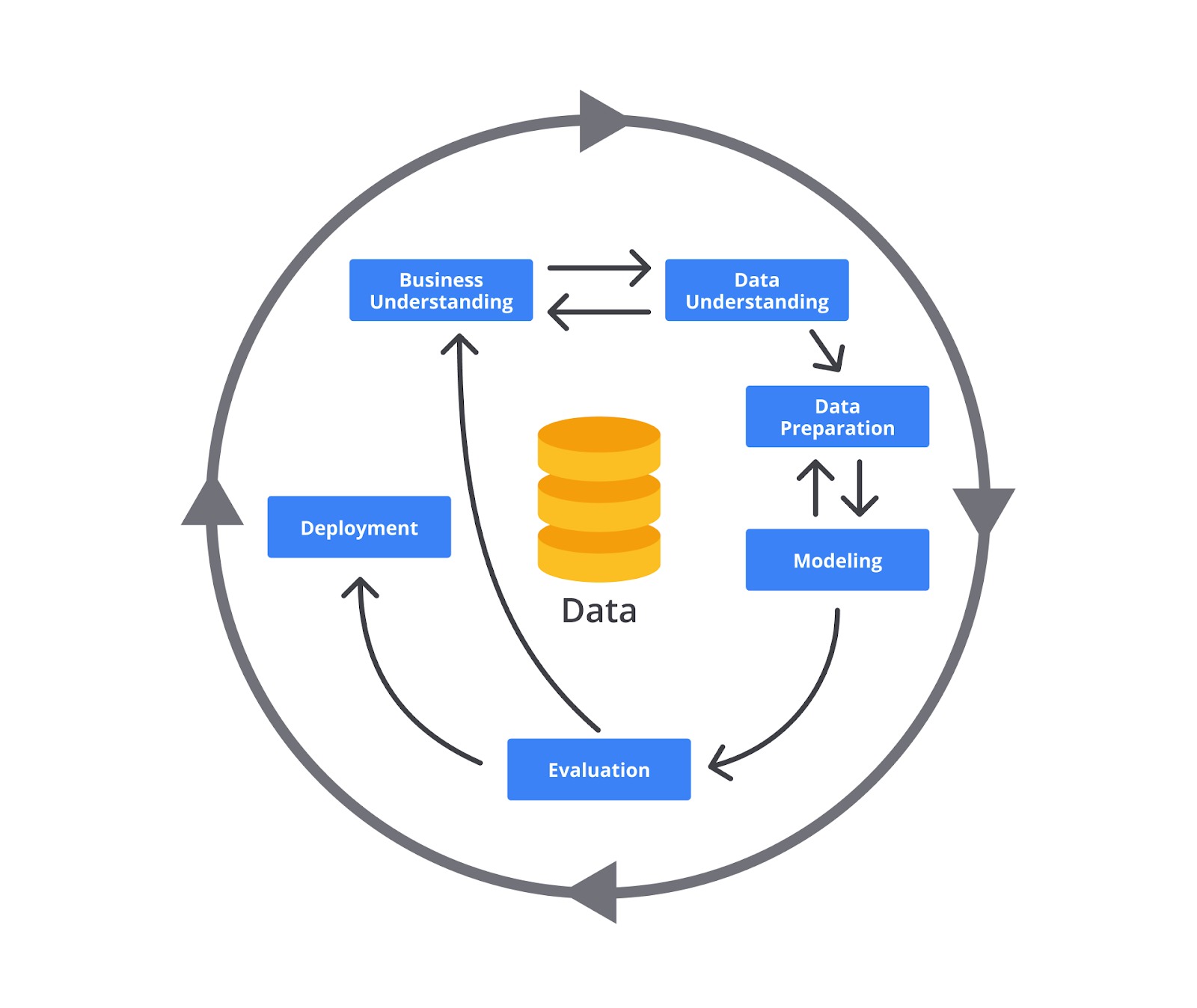
Secara sadar maupun tidak sadar, mungkin sebenarnya Anda telah bergelut dengan matriks di kehidupan sehari-hari. Mulai dari nilai rapot saat sekolah, pembagian daftar nama saat *study tour*, atau bahkan melihat daftar nama yang lolos saat seleksi masuk kerja, benar bukan?

Lalu, apa sebenarnya perbedaan antara vektor dan matriks? Mudahnya vektor adalah representasi dari satu dimensi array, dan matriks adalah representasi dari beberapa dimensi array.

Bagaimana terkait pembahasan vektor dan matriks? Anda pasti sudah paham, bukan? Nah, setelah Anda mempelajari seluruh rangkaian fundamental data science, hal terpenting lainnya yang perlu Anda ketahui adalah keterampilan yang perlu dimiliki untuk menjadi seorang data scientist maka dari itu yuk kita simak penjelasan di materi berikutnya.

**Metodologi Data Science**

*Last but not least* yang perlu menjadi dasar pengetahuan dalam kelas dasar data science adalah terkait metodologi data science. Pernahkah Anda sebelumnya mendengar CRISP-DM?



CRISP-DM atau *Cross Industry Standard Process for Data Mining*adalah proses model yang berfungsi sebagai dasar untuk proses data science. Proses model CRISP-DM memberikan gambaran tentang siklus hidup proyek data mining yang terdapat pada big data.

Sekilas info, *data mining*berarti proses pengumpulan informasi penting dari data dalam jumlah yang besar.

CRISP-DM memiliki enam tahapan yaitu *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation,*dan *Deployment*. Mari kita *breakdown*satu per satu.

1. ***Business Understanding***  
   Tahap ini merupakan pemahaman mendalam tentang kebutuhan pelanggan. Kegiatan yang dilakukan pada tahap ini adalah menentukan tujuan bisnis, menilai situasi ketersediaan sumber daya, tentukan tujuan pengumpulan data, dan menghasilkan rencana proyek.
2. ***Data Understanding***  
   Tahap pemahaman data yaitu mendorong fokus untuk mengidentifikasi, mengumpulkan, dan menganalisis kumpulan data yang dapat membantu Anda mencapai tujuan proyek. Kegiatan pada tahap ini adalah mengumpulkan data awal, menjelaskan data, jelajahi data, dan verifikasi kualitas data.
3. ***Data Preparation***  
   Fase ini sering disebut “data mining”, yaitu menyiapkan kumpulan data akhir untuk pemodelan. Kegiatan pada fase ini diantaranya memilih data yaitu menentukan set data yang akan digunakan, lalu bersihkan data, bangun data atau dengan kata lain mendapatkan atribut baru yang akan membantu, lalu integrasikan data, dan yang terakhir format data.
4. ***Modeling***  
   Secara garis besar pada tahap ini Anda akan membuat dan menilai berbagai model berdasarkan beberapa teknik pemodelan yang berbeda. Pada tahap ini terdapat empat tugas, yaitu memilih teknik pemodelan, menghasilkan desain pengujian, membangun model, dan yang terakhir menilai model.
5. ***Evaluation***  
   Fase evaluasi ini melihat lebih luas model yang paling sesuai dengan bisnis dan yang harus dilakukan selanjutnya. Terdapat tiga kegiatan yang mewakili fase evaluasi, yaitu evaluasi hasil, proses peninjauan, dan tentukan langkah selanjutnya.
6. ***Deployment***  
   Merupakan tahap terakhir dan tahap yang paling dihargai dari proses CRISP-DM. Perencanaan untuk *deployment* dimulai dari fase *business understanding*dan harus menggabungkan tidak hanya untuk menghasilkan nilai model, tetapi juga cara mengonversi skor keputusan, dan cara untuk menggabungkan keputusan tersebut.

**Keterampilan Data Scientist**



Menjadi seorang data scientist tentu bukan hanya keterampilan teknis yang diperlukan, tetapi perlu adanya keseimbangan antara soft skill dan hard skill untuk menunjang karier Anda di masa depan. Pada pembahasan kali ini, Anda akan mengeksplorasi keterampilan yang dibutuhkan agar menjadi seorang data scientist yang Andal. Simak penjelasannya dengan saksama, ya.

**Keterampilan Interpersonal**

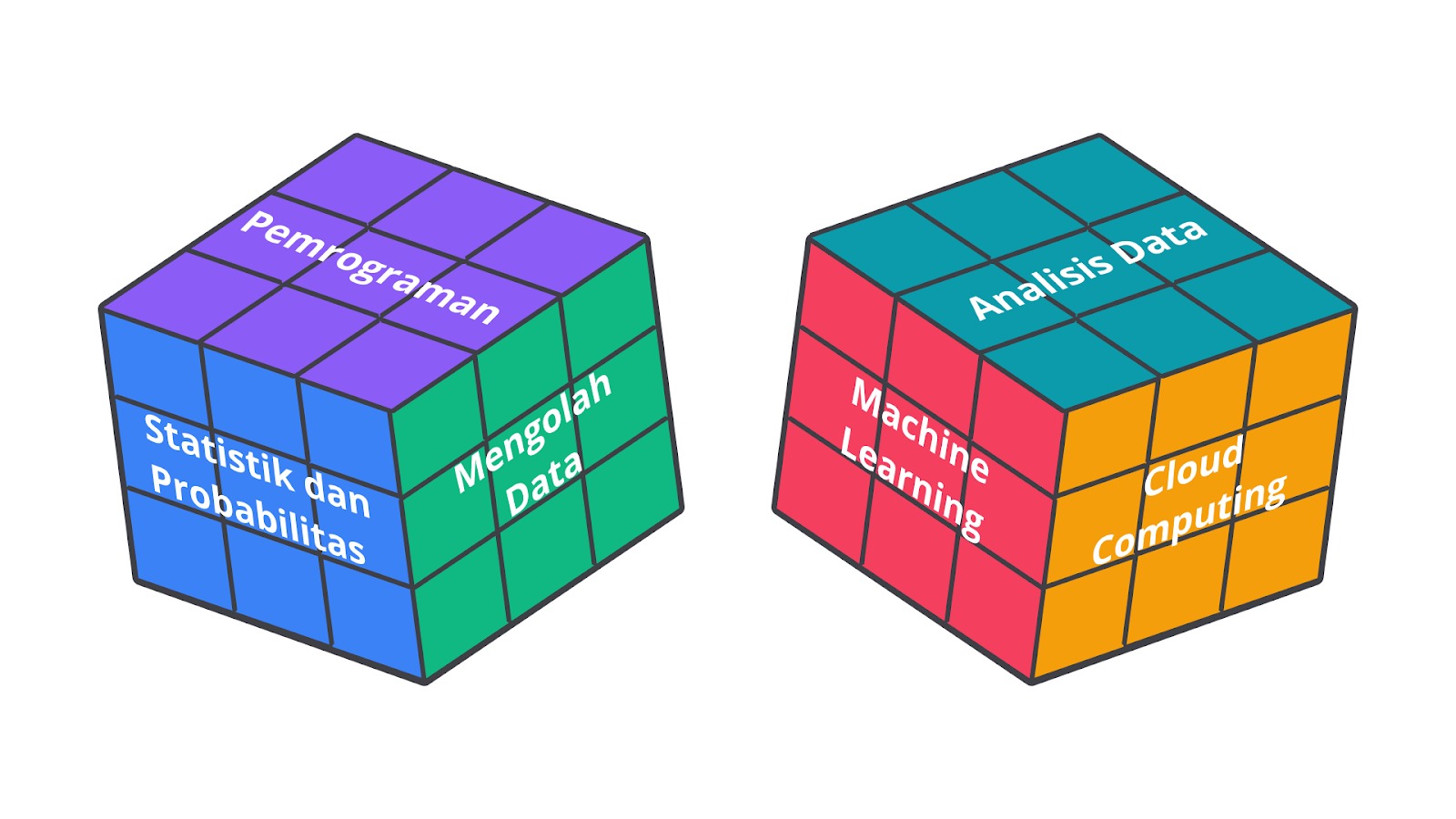
Menurut Investopedia, keterampilan interpersonal adalah kemampuan yang dimiliki seseorang dalam melakukan komunikasi maupun interaksi dengan orang lain.

Mari kita *breakdown*satu per satu lima hal esensial terkait keterampilan interpersonal yang harus dimiliki oleh seorang data scientist.

1. **Komunikasi Efektif**  
   Bagi seorang data scientist, komunikasi lisan atau tertulis sangatlah perlu diperhatikan. Bukan hanya untuk menunjang kualitas pekerjaan Anda, tetapi juga sebagai cara dalam berinteraksi dengan tim, klien, dan para senior Anda.  
     
   Mengembangkan keterampilan komunikasi yang baik sudah jelas akan membantu Anda dalam membagikan ide-ide yang kompleks dengan jelas agar semua orang mampu memahaminya dengan sederhana dan jelas. Selain itu, Anda juga akan lebih mudah ketika bertemu dengan klien yang memiliki kebutuhan berbeda dari Anda atau belum pernah melihat pekerjaan serupa sebelumnya.  
     
   Hal lain yang terpenting dari memiliki komunikasi efektif adalah saat melakukan presentasi lisan terkadang kita menggunakan slide presentasi sebagai penunjang, tetapi Anda harus tetap menunjukan informasi tambahan dan menyampaikan poin-poin penting untuk meningkatkan pemahaman audiens.
2. **Pengetahuan**  
   Seperti yang tertera pada gambar di atas, mungkin orang awam akan menganggap bahwa teknologi tersebut sangatlah canggih atau bahkan seperti sulap karena seolah-olah ajaib dapat dengan cepat mengeluarkan sebuah jawaban.  
     
   Pada kenyataannya, hal tersebut tidak luput dari peran seorang data scientist yang memiliki pengetahuan tentang mengidentifikasi suatu masalah dan relevansi dengan tujuan bisnis. Tanpa adanya keterampilan tersebut, akan sulit untuk mengidentifikasi metode yang tepat untuk diterapkan pada masalah.
3. **Memiliki Pemahaman Bisnis**  
   Memiliki pemahaman yang baik tentang bisnis adalah poin *plus*bagi seorang data scientist. Mengapa demikian? Ini memungkinkan Anda untuk memahami cara berbagai segmen dan proses dalam bisnis bekerja, serta dapat memahami masalah yang mungkin dihadapi perusahaan. Setelah Anda memiliki pemahaman tentang hal yang ingin dicapai oleh perusahaan, akan lebih mudah untuk menerjemahkan masalah bisnis dan memilih solusi yang paling tepat.  
     
   Tidak dapat dipungkiri apabila nanti Anda akan menemukan sebuah organisasi atau perusahaan yang tergoda untuk dapat membangun hal yang kompleks dengan seorang data scientist untuk menyelesaikan masalah perusahaan. Namun, sebelum menyelesaikan persoalan tersebut, data analyst bisa jadi sudah cukup maka dari itu sebelum benar-benar menyelesaikan persoalan sebuah perusahaan, pastikan bahwa pekerjaan tersebut adalah *scope*Anda.
4. **Kerja Sama**  
   Poin keempat yang perlu dimiliki oleh seorang data scientist yang lain, yaitu mampu bekerja dengan orang lain. Seperti yang kita tahu bahwa kerja sama tim adalah kunci keberhasilan untuk semua bidang. Dengan memiliki kemampuan kerja sama yang bagus, secara tidak langsung hal tersebut akan membantu Anda mencapai tujuan bersama.
5. **Penyelesaian Masalah (*Problem Solving*)**Hal esensial terakhir yang perlu dimiliki oleh seorang data scientist adalah *problem solving*. Pada poin nomor dua telah disebutkan bahwa seorang data scientist memiliki pengetahuan tentang mengidentifikasi suatu masalah dan relevansi dengan tujuan bisnis. Tidak cukup sampai di situ, setelah melaksanakan identifikasi terhadap suatu masalah maka tugas data scientist yang lain yaitu mampu menyelesaikan masalah hingga mencapai tujuan bisnis organisasi atau perusahaan.

**Keterampilan Teknis**

Setelah mengetahui dan menguasai keterampilan interpersonal untuk menjadi data scientist, Anda akan diajak berkenalan dengan keterampilan teknis yang diperlukan untuk menjadi seorang data scientist.



1. **Statistik dan Probabilitas**  
   Untuk dapat menulis model algoritma *machine learning*, data scientist perlu mempelajari statistika dan probabilitas terlebih dahulu sebelum mempelajari yang lain. Untuk *machine learning*, penting untuk menggunakan konsep analisis statistik seperti regresi linier, hal itu yang menjadi alasan bahwa langkah pertama dalam mempelajari data science yaitu memiliki ketertarikan terhadap statistika dan probabilitas. Seorang data scientist harus dapat mengumpulkan, menafsirkan, mengatur, dan menyajikan data, serta sepenuhnya memahami konsep rata-rata (*average*), median, modus, *varians,*dan standar deviasi.
2. **Mengolah Data (Data wrangling dan database management)**  
   Perlu digarisbawahi, bahwa kemampuan mengolah data ini sudah harus diiringi dengan kemampuan *data mining*yang telah dibahas di atas. Oke, setelah menguasai statistika dan probabilitas, langkah selanjutnya adalah mengolah data, tetapi pastikan bahwa Anda telah melakukan data mining sebelumnya. Arti dari mengolah data disini adalah proses *data wrangling*yaitu pembersihan dan pengorganisasian kumpulan data yang kompleks agar lebih mudah diakses dan dianalisis. Memanipulasi data untuk mengategorikannya berdasarkan pola dan tren untuk membuat keputusan berdasarkan data. Pada bagian ini juga berkaitan dengan pemahaman *database management* atau manajemen basis data yang berfungsi untuk mengekstrak data dari sumber yang berbeda dan mengubahnya menjadi format yang sesuai untuk kueri dan analisis, lalu memasukkan data tersebut ke *data warehouse system*.
3. **Pemrograman**  
   Setelah menguasai statistika dan probabilitas, lalu mampu mengolah data, langkah ketiga adalah Anda perlu memiliki ketertarikan terhadap pemrograman. Bahasa pemrograman yang sering digunakan pada data science umumnya seperti Python atau R. Hal tersebut diperlukan untuk menyortir, menganalisis, dan mengelola data dalam jumlah besar (*big data*). Sebagai data scientist yang baru memulai kariernya, Anda harus mengetahui konsep dasar data science dan mulai membiasakan diri dengan cara menggunakan Python. Adapun bahasa pemrograman populer yang digunakan pada data science antara lain sebagai berikut.
   * Python
   * R
   * SAS
   * SQL
4. **Machine Learning**  
   Perbedaan antara data science dengan disiplin ilmu data lainnya adalah terdapat *machine learning* di dalamnya. Menggabungkan beberapa disiplin ilmu dan teknik membantu Anda meningkatkan keahlian menjadi seorang data scientist karena dapat mengumpulkan dan menyintesis data dengan lebih efisien, serta mampu memprediksi kumpulan data tersebut untuk masa mendatang. Kita akan membahas lebih mendalam lagi terkait *machine learning*di modul lain pada kelas ini. *So… stay tune*terus, ya.
5. **Analisis Data (Data Visualisasi)**  
   Ibaratnya, setelah semua data dikumpulkan dan diolah, Anda diminta memiliki ketajaman dalam menganalisis data karena hasil dari analisis tersebut yang akan digunakan untuk membuat sebuah keputusan. Selain perlu mengetahui cara menganalisis, mengatur, dan mengategorikan data, tetapi Anda juga perlu membangun keterampilan dalam visualisasi data.  
     
   Menjadi seorang data scientist tentu perlu memiliki kemampuan dalam membuat *chart*atau grafik penting. *Tenang dulu,*materi terkait analisis data akan kita bahas pada modul berikutnya. *Jia you!* (baca: *ciayo!*)
6. **Cloud Computing**  
   Setelah semua rangkaian selesai, hal yang perlu diperhatikan adalah platform cloud. Sebagai data scientist, kemungkinan besar Anda perlu menggunakan *cloud*yang membantu menganalisis dan memvisualisasikan data yang tersimpan di platform *cloud*.

**Rangkuman Fundamental Data Science**

Amazon mengemukakan bahwa data science adalah wawasan yang sangat berguna untuk kebutuhan bisnis. Ilmu ini merupakan penggabungan prinsip dan praktik dari bidang matematika, statistik, *artificial intelligence,*dan *computer engineering* untuk menganalisis data dalam jumlah besar.

Menurut Jose Antonio Ribeiro pada publikasinya di*Big Data for Executives and Professionals*, data science adalah kegiatan ilmiah yang mencakup pemahaman bisnis, persiapan data, analisis eksplorasi, pemodelan, aplikasi komputasi, evaluasi wawasan, dan tindakan untuk menghasilkan hasil serta  menerapkan pengetahuan statistik, matematika, dan komputasi.

Adapun definisi lain, menurut IBM (*International Business Machine*) menyebutkan bahwa data science adalah menggabungkan matematika dan statistik, pemrograman, analitik, *artificial intelligence*, dan *machine learning*.

**Data Science :** Data Science merupakan disiplin ilmu.

**Data Scientist :** Data scientist adalah praktisi dalam bidang data science.

**Data Engineering :** Data engineer merupakan software engineer yang tugas utamanya adalah menyiapkan data yang dikoleksi dari berbagai sumber informasi dengan membangun sistem data atau infrastruktur agar dapat mudah untuk dianalisa dan mendukung kebutuhan perusahaan.

**Data Analyst :** Data analyst adalah profesi pekerjaan yang bertugas untuk membaca serta menganalisis data yang ada dari perusahaan.

**Mengapa Data Science Begitu Penting?**

Sederhananya data science membantu organisasi atau perusahaan mendapatkan *insight*(wawasan) tentang bisnis mereka sendiri, tingkat keberhasilan strategi mereka, kinerja anggotanya, dan sebagainya.

**Keterampilan Interpersonal**

1. Komunikasi Efektif
2. Pengetahuan
3. Memiliki Pemahaman Bisnis
4. Kerja Sama
5. Penyelesaian Masalah (*Problem Solving*)

**Keterampilan Teknis**

1. Statistik dan Probabilitas
2. Mengolah Data (Data wrangling dan database management)
3. Pemrograman
4. Machine Learning
5. Analisis Data (Data Analyst)
6. Cloud Computing