

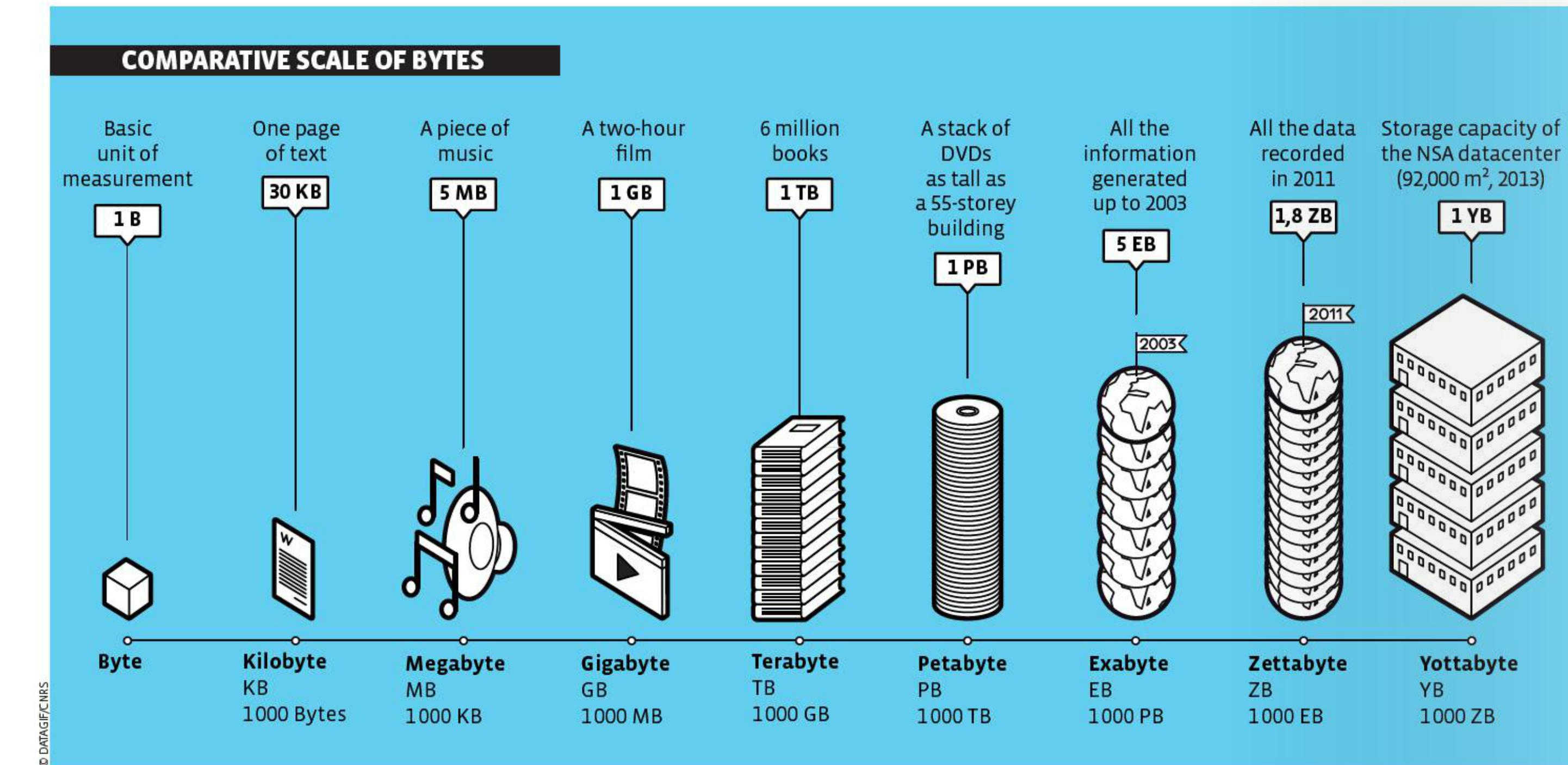
Pengantar Data Analisis

Bahan Kuliah SD2104 Pemrograman Lanjut

Sevi Nurafni

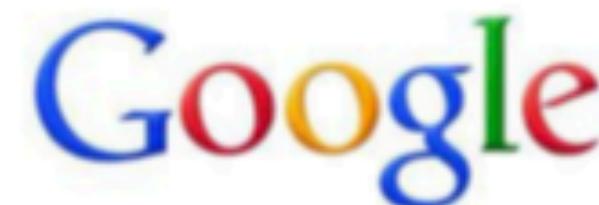
Fakultas Sains dan Teknologi
Universitas Koperasi Indonesia 2024

Big Data



- "Jika semua data yang digunakan di dunia saat ini ditulis ke dalam CD-ROM, dan CD-ROM tersebut ditumpuk dalam satu tumpukan, tumpukan itu akan menjulang dari Bumi ke Bulan dan kembali lagi hingga seperempat perjalanan."
 - Hilbert, M & Lopez, P. (2011), "The world's technological capacity to store, communicate and compute information", Science 332, 1 April 2011, 60-65
- "Pada tahun 2020, jumlah data digital di dunia akan mencapai 40 zettabyte (ZB), setara dengan 40 triliun GB data, atau sekitar 5.200 GB data untuk setiap orang di Bumi."
 - IDC (2010), "IDC Digital Universe Study, sponsored by EMC", May 2010

Big Data, IoT



24 PB/day
(2009)



2.5 PB of user data +
15 TB/day (2009)



6.5 PB of user data +
50 TB/day (2009)

Web, Social Media & Network



10^{18} bytes/day
(2024, est.)



22 PB (2012,
the Large Hadron
Collider)



*Scientific Data,
Scientific Instruments*



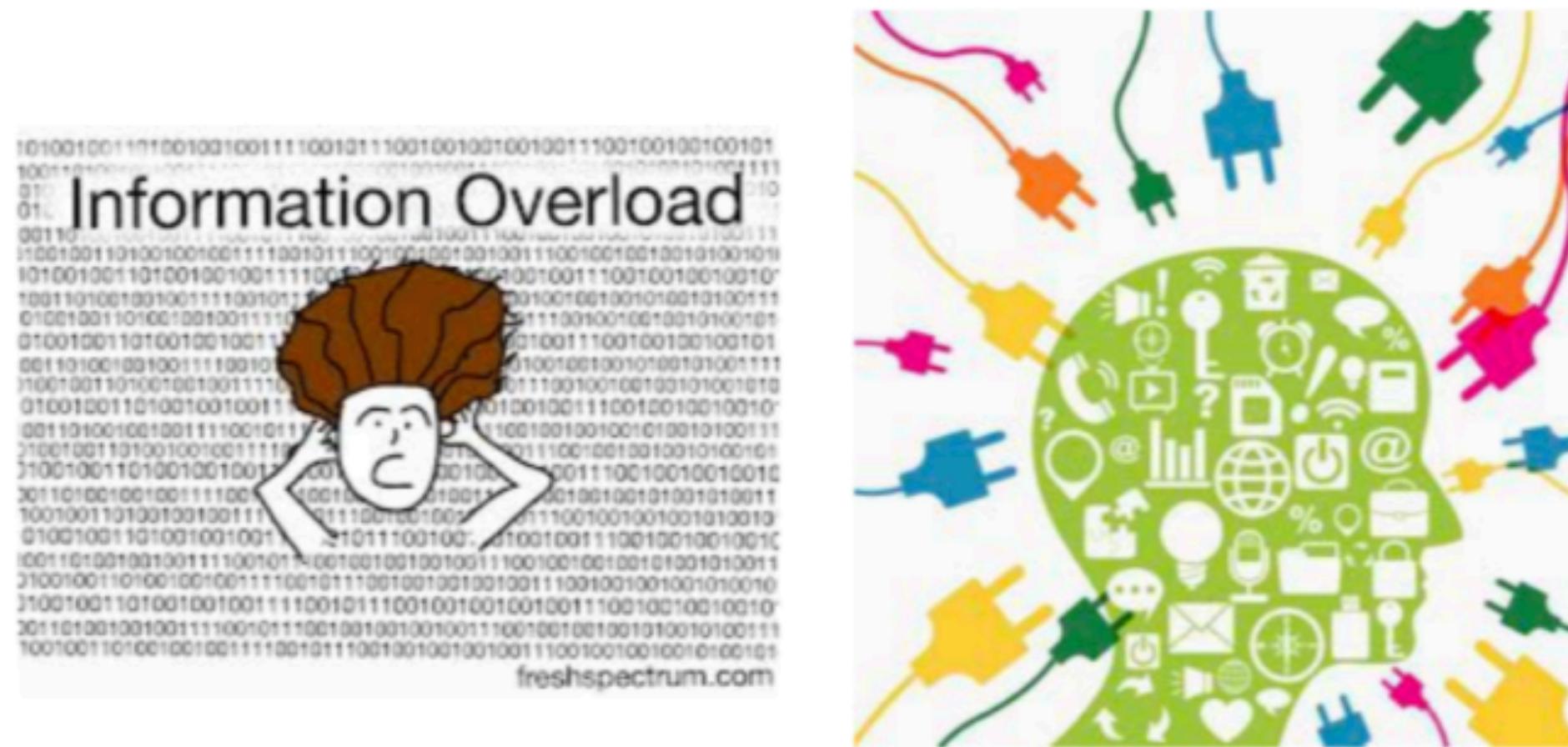
Mobile Devices, IoT, Sensor Technology & Networks

Informasi yang Berlebihan

Kesulitan dalam memahami sebuah isu dan membuat pilihan secara efektif ketika memiliki terlalu banyak informasi tentang isu tersebut. (Yang et al., 2003)



Overload informasi terjadi ketika jumlah masukan ke suatu sistem melebihi kapasitas pemrosesannya. Pembuat keputusan memiliki kapasitas pemrosesan kognitif yang cukup terbatas. Akibatnya, ketika overload informasi terjadi, kemungkinan besar kualitas keputusan akan menurun. (Speier et al., 1999)



People are getting stupider?



Adopted from: John Stasko <http://www.cc.gatech.edu/~stasko/7450/Notes/overview.pdf>

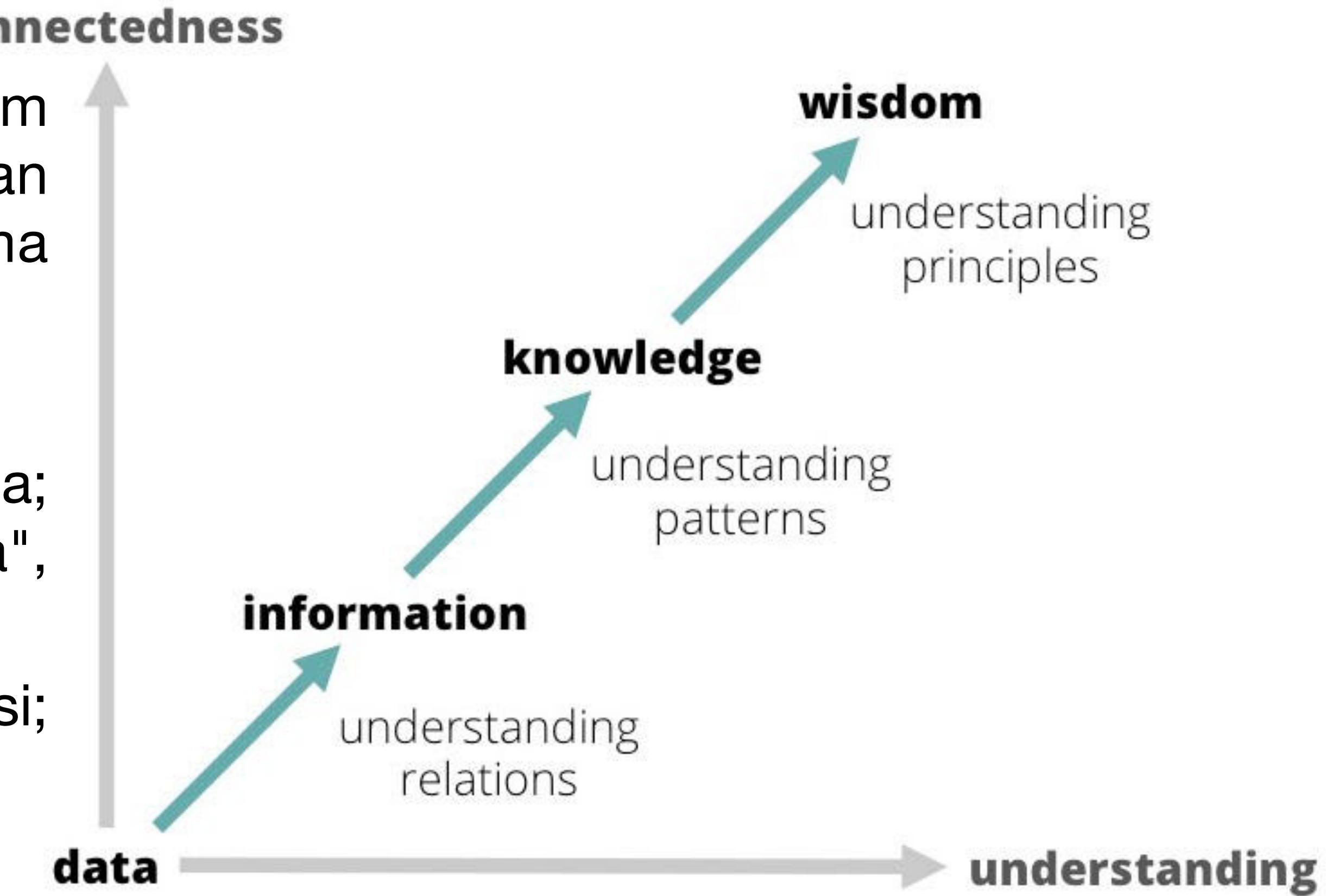
Data, Information, Knowledge, and Wisdom

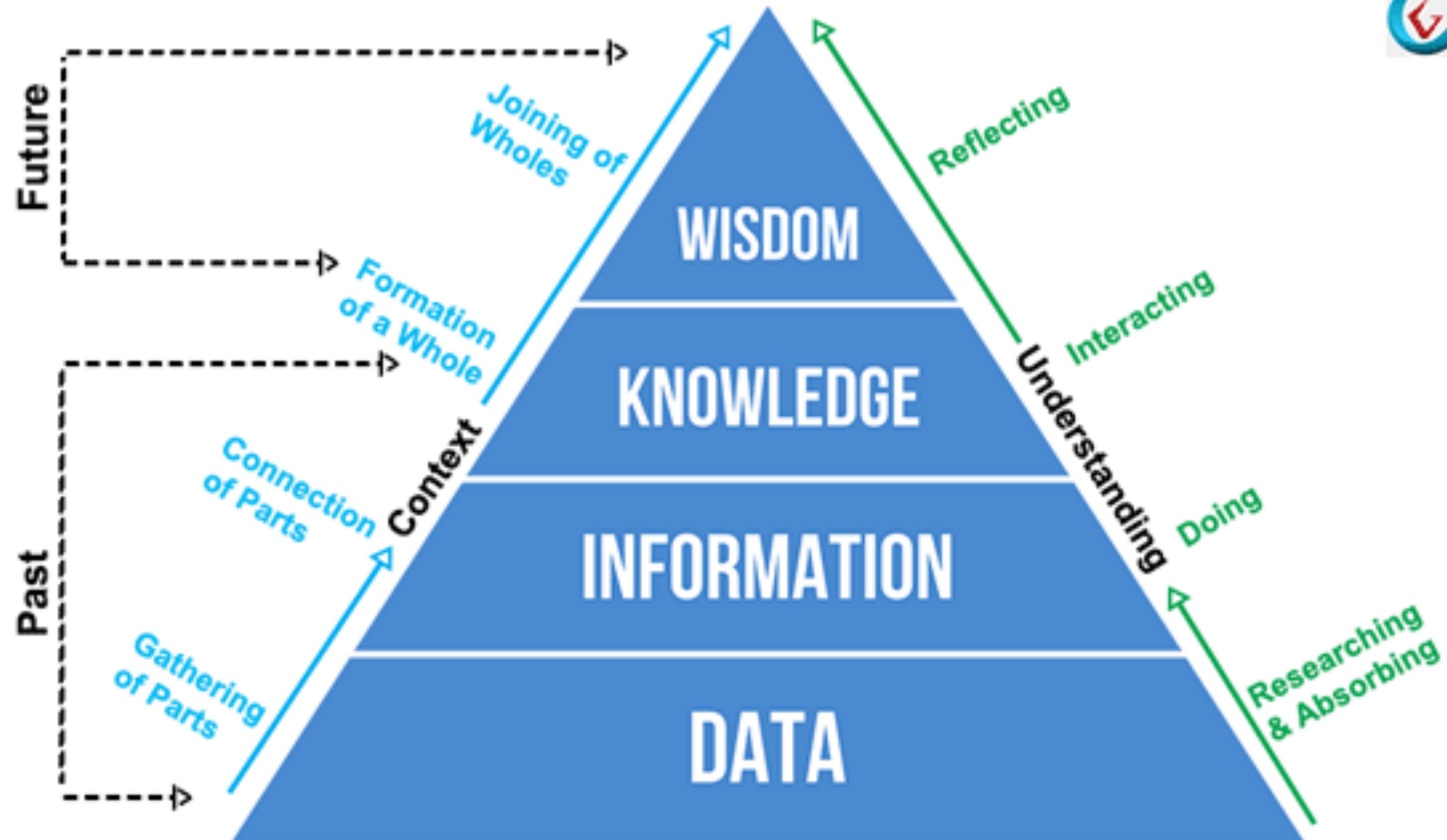
By Gene Bellinger, Durval, Castro, Anthony Mills



Menurut Russell Ackoff, seorang ahli teori sistem dan profesor perubahan organisasi, isi pikiran manusia dapat diklasifikasikan ke dalam lima kategori:

- **Data:** simbol-simbol.
- **Informasi:** data yang diproses agar berguna; memberikan jawaban atas pertanyaan "siapa", "apa", "di mana", dan "kapan".
- **Pengetahuan:** penerapan data dan informasi; menjawab pertanyaan "bagaimana".
- **Pemahaman:** apresiasi terhadap "mengapa".
- **Kebijaksanaan:** pemahaman yang dievaluasi.





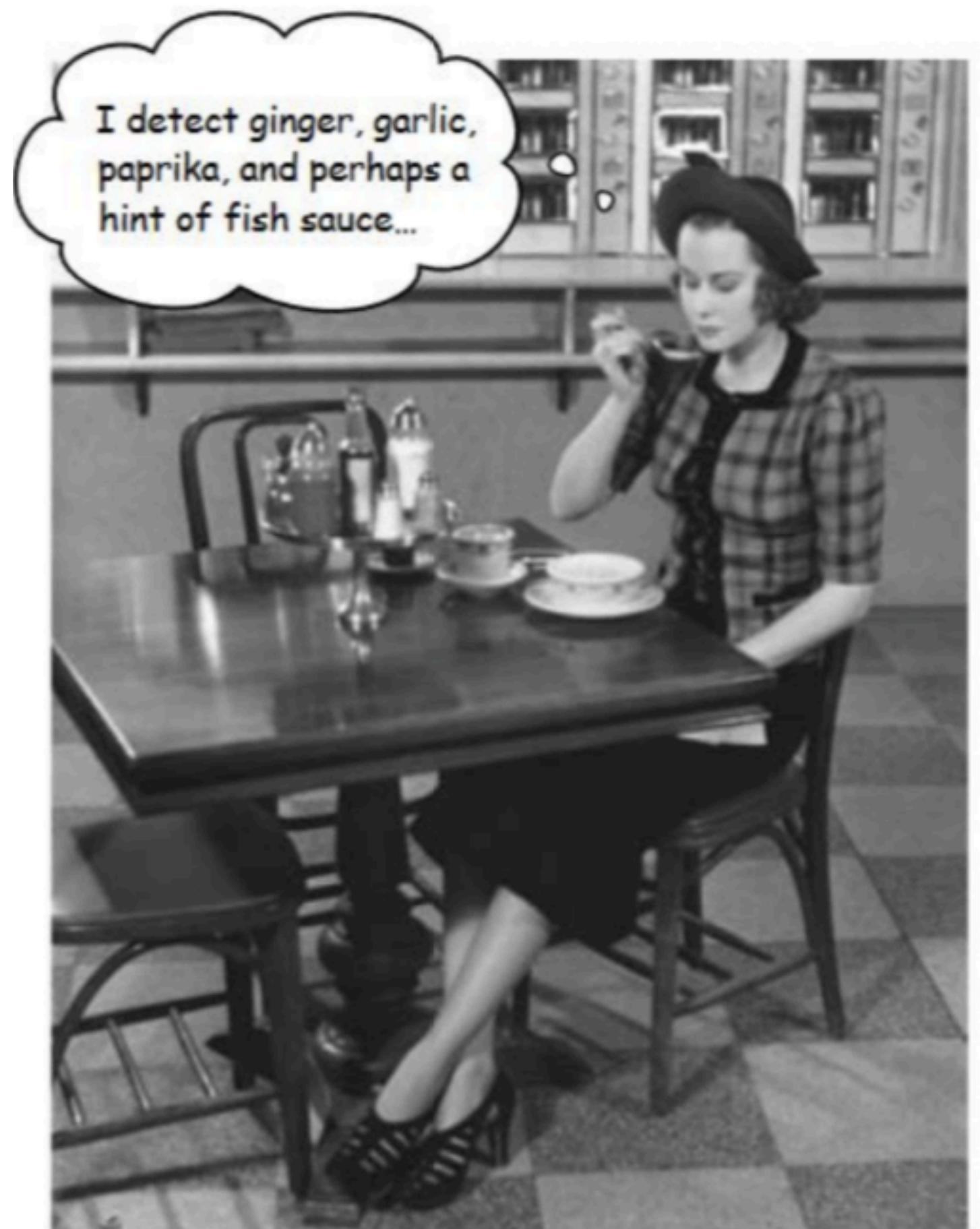
DIKW Pyramid
CertGuidance.com

Data Analysis

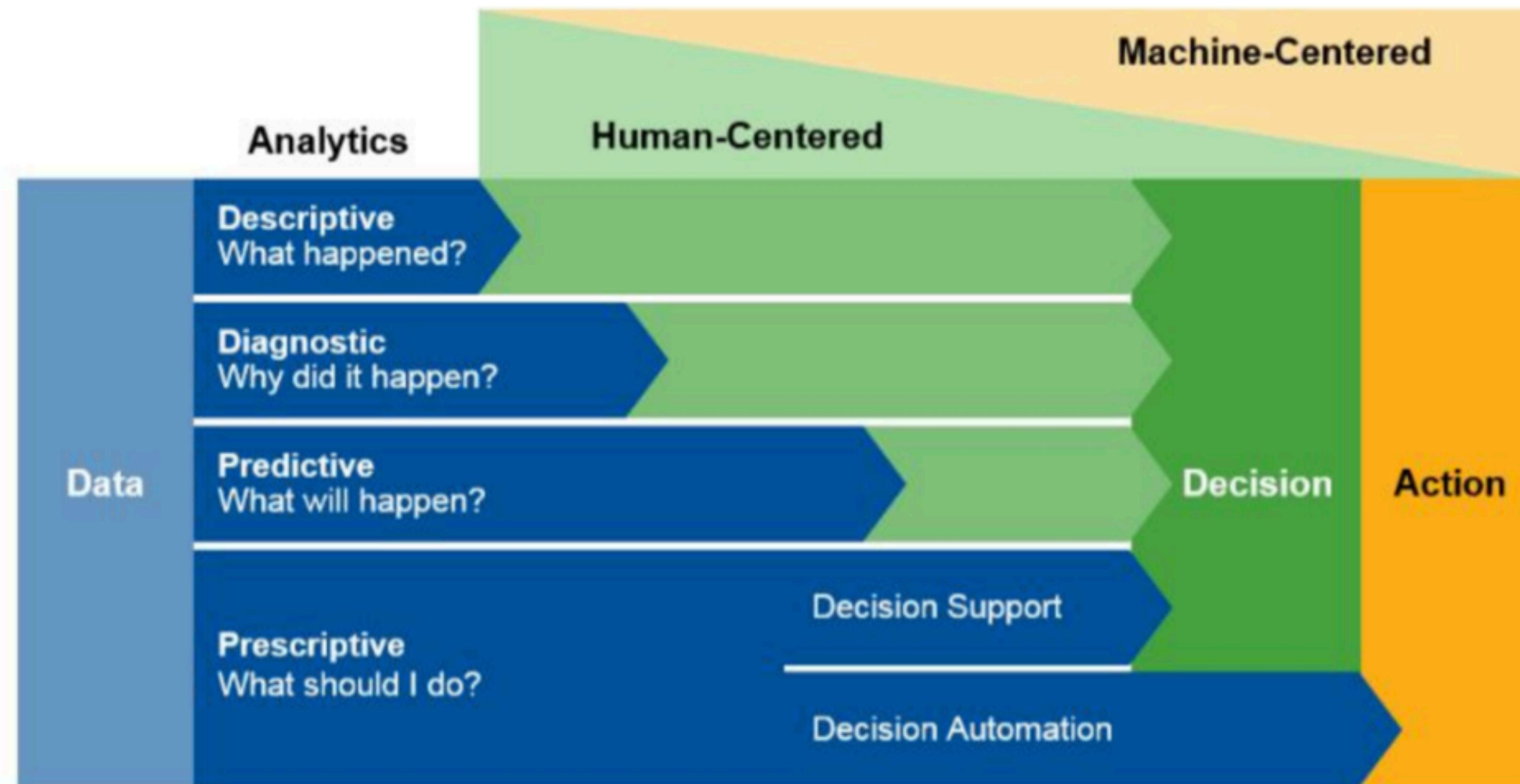
Data Analysis

- **Definisi Umum:** Bagaimana memecah dan menyusun masalah serta kumpulan data yang kompleks untuk menemukan inti permasalahan dalam bisnis mereka. [1]
- **Definisi Proses:** Proses yang meliputi inspeksi, pembersihan, transformasi, dan pemodelan data dengan tujuan menemukan informasi yang berguna, memberikan wawasan untuk kesimpulan, dan mendukung pengambilan keputusan. [2]

1. Michael Milton, Head First Data Analysis, O'Reilly 2009
2. https://en.wikipedia.org/wiki/Data_analysis



Klasifikasi Data Analytics



Source: Gartner (October 2016)

Sumber:
2017 Planning Guide for Data and Analytics

Descriptive Analytics

- Jenis paling sederhana dari Data Analytics
- Analisis terhadap data history untuk mendapatkan profil umum dalam bentuk summary dari data atau hubungan antar data untuk menjelaskan situasi yang telah terjadi.
- Contoh hasil analisis:
 - Banyaknya friend, mention, followers, page views
 - Banyaknya page views
 - Perbandingan banyaknya mahasiswa antar prodi di IKOPIN
 - Rata-rata nilai mahasiswa
 - Hubungan antara banyaknya jam belajar dengan prestasi akademik
 - Ada kecenderungan bahwa orang beli roti tawar bersamaan dengan butter/mentega
 - dll

Predictive Analytics

- Analisis terhadap data history untuk mendapatkan hubungan dan trend yang ada (yang direpresentasikan dalam bentuk model prediksi) dalam rangka untuk memperkirakan apa yang akan terjadi di masa yang akan datang.
- Contoh:
 - Memperkirakan nilai saham atau mata uang tertentu berdasarkan data nilai saham atau mata uang pada periode waktu sebelumnya
 - Memperkirakan apakah seseorang dengan karakteristik tertentu (usia, penghasilan, jumlah tanggungan, frekwensi sakit berat) layak/tidak diberikan kredit bank, berdasarkan data history pengambilan keputusan oleh ahli keuangan.
 - dll

Predictive Analytics

- Analisis terhadap data history untuk dapat menghasilkan kesimpulan berupa rekomendasi bagaimana sesuatu harus dilakukan
- Contoh:
 - Menentukan rute terbaik dari satu tempat ke tempat lain, berdasarkan data yang ada
 - Analisis oleh travel agent terhadap berbagai faktor terkait travel (customer, tujuan, waktu, dll) untuk optimasi harga tiket
- **Tidak dibahas detail di kuliah ini**

Tipe Data

- Categorical-Nominal
 - Nama negara, warna kulit, nama program studi, dll
- Categorical-Ordinal
 - Likert scale (“sangat setuju” s.d. “sangat tidak setuju”)
 - Indeks nilai A, B, C, D, E
- Categorical-Binary
 - Jenis kelamin, status mahasiswa (aktif, tidak aktif), dll
- Quantitative-Discrete
 - Banyaknya anak, banyaknya mahasiswa, banyaknya sks lulus
- Quantitative-Continues
 - Usia, berat badan, tinggi, suhu

Representasi Data berdasarkan Strukturnya

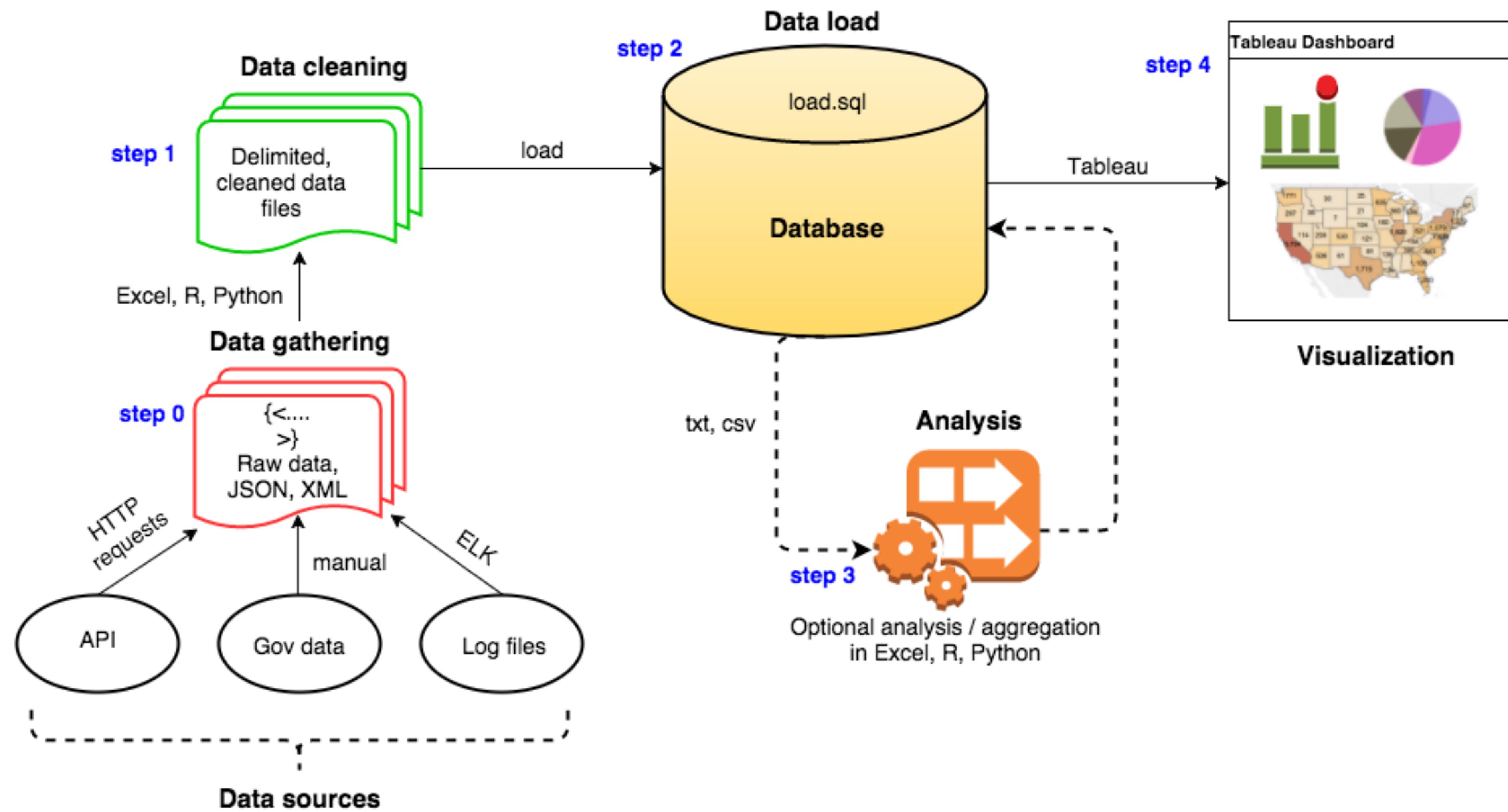
- Structured Data
 - Data dalam bentuk table/relational
 - Contoh: data dalam xls, data tersimpan dalam relational DBMS
- Semi-structured Data
 - Data yang tidak direpresentasikan dalam bentuk table, namun masih memiliki struktur/pengorganisasian yang memudahkan proses/analisis
 - Contoh: data dalam format xls, json, noSQL database
- Unstructured Data
 - Data yang tidak memiliki struktur yang memudahkan proses/analisis
 - Contoh: data teks, data video, data foto

Contoh-contoh kegiatan Data Analysis Descriptive Analytics dan Explanatory DA

- Retrieve Value (Selection)
- Filter
- Compute Derived Value
- Find Extremum
- Sort
- Determine Range
- Characterize Distribution
- Find Anomalies
- Correlation
- Clustering

https://en.wikipedia.org/wiki/Data_analysis

Data Analysis Workflow



Data Store

- Adalah sebuah repository untuk secara persisten (bersifat tetap) menyimpan dan mengelola kumpulan data
- Bentuk-bentuk data store:
 - File, semacam csv file, spreadsheet
 - Email
 - Database
 - Distributed Data Store
 - Directory Services
 - dll

Data Gathering

- Data dapat dikumpulkan dari berbagai sumber:
 - **Transactional data:** data yang berasal dari transaksi sebuah organisasi
 - Tersimpan dalam database organisasi, membutuhkan cara akses khusus
 - **Log files:** data mengenai event yang terjadi dalam suatu sistem, misalnya banyaknya klik atau page-request pada suatu website
 - **API** (Application Programming Interface): aplikasi khusus yang disediakan website untuk men-download data
 - Format data yang umum: XML, JSON, XLS
 - **Online-datasets:** data tersedia secara online dari berbagai website baik pemerintah maupun swasta, dapat di-download secara manual
 - Format data yang umum: CSV, TXT, PDF, JSON, XML, HTML, XLS
- Data yang tersedia untuk publik dapat bersifat gratis (free) atau berbayar (for-purchase)

Mengenal berbagai format data

CSV (Comma-Separated Values):

Text file, data dipisahkan comma atau separator lain

```
1 Year,Make,Model,Description,Price
2 1997,Ford,E350,"ac, abs, moon",3000.00
3 1999,Chevy,"Venture ""Extended Edition""","",4900.00
4 1999,Chevy,"Venture ""Extended Edition, Very Large""",,5000.00
5 1996,Jeep,Grand Cherokee,"MUST SELL!
6 air, moon roof, loaded",4799.00
```

XLS (excel spreadsheet):

Format khusus MS Excel, menyimpan data, chart, macro, dll.

	A	B	C	D	E
1	Year	Make	Model	Description	Price
2	1997	Ford	E350	ac, abs, moon	3000
3	1999	Chevy	Venture "Extended Edition"		4900
4	1999	Chevy	Venture "Extended Edition, Very Large"		5000
5	1996	Jeep	Grand Cherokee	MUST SELL!	4799
6					

Mengenal berbagai format data

JSON (JavaScript Object Notation)

XML (eXtensible Mark-up Language):

Data ditandai dengan menggunakan *tag*.

```
<?xml version="1.0" encoding="UTF-8"?>
<breakfast_menu>
<food>
  <name>Belgian Waffles</name>
  <price>$5.95</price>
  <description>
    Two of our famous Belgian Waffles with plenty of real maple syrup
  </description>
  <calories>650</calories>
</food>
```

```
{
  "empid": "SJ011MS",
  "personal": {
    "name": "Smith Jones",
    "gender": "Male",
    "age": 28,
    "address": {
      "streetaddress": "7 24th Street",
      "city": "New York",
      "state": "NY",
      "postalcode": "10038"
    },
    "profile": {
      "designation": "Deputy General",
      "department": "Finance"
    }
  }
}
```

Data Cleansing

Corrupted Data

At Spotless Data we estimate that 5% of overall data held by companies is corrupted and lacking in data integrity, though a recent report estimated that manually entered data could contain an error rate of anywhere between 2.3% and 26.9%.

What this may mean is that if I own a company with 500,000 clients or users and, like Google, I estimate that each customer is worth \$80 to my business, and if the primary contact I have with these customers is through a submitted email address, and 5% of those email addresses are badly formatted, then I will have lost 25,000 customers and \$2 million in income. This might represent all my profit or be the difference between turning a profit and making a loss.

Those customers might also end up consuming one of my competitor's services instead. This is especially so if, having given up on my company for not responding to their initial email submission, they then submit their faulty email to my competitor who, unlike me, is ensuring that they have Data Quality which they can trust. They can thus correctly identify, their email address, and start building a relationship with said customers.

Sumber: <https://web.archive.org/web/20171205042031/https://spotlessdata.com/blog/importance-data-cleaning-user-generated-content>

Data Cleansing

Proses mendeteksi dan memperbaiki (atau menghapus) catatan yang rusak atau tidak akurat dari kumpulan data, tabel, atau basis data. Proses ini mencakup identifikasi bagian data yang tidak lengkap, salah, tidak akurat, atau tidak relevan, kemudian mengganti, memodifikasi, atau menghapus data yang kotor atau kasar tersebut.

Wu, S. (2013), "A review on coarse warranty data and analysis"



Data yang salah atau tidak konsisten dapat menyebabkan kesimpulan yang keliru dan investasi yang salah arah.

Kualitas Data

- Hasil analisis sangat dipengaruhi oleh kualitas data
- Ada banyak dimensi kualitas data

Sumber:

https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf

<https://www.cdc.gov/ncbddd/hearingloss/documents/dataqualityworksheet.pdf>

<https://www.rfigroup.com/rfi-group/news/rfi-group-opinion-australia-why-Business-leaders-need-own-data-quality>



Problems with Data

- Duplicate entries
 - Data yang sama tercatat lebih dari 1 kali
 - Bisa jadi yang dibutuhkan sebenarnya hanya 1 data, tetapi dalam beberapa kasus entry lebih dari 1 kali menjadi penting
- Multiple entries of single entity
 - Data yang berhubungan dengan objek yang sama dientri lebih dari 1 kali, bisa dengan nilai yang berbeda. Contoh: Mahasiswa A, di satu data tercatat tingginya 150cm, di data lain tercatat tingginya 148cm
- Missing entries
 - Data yang seharusnya ada, tidak ditemukan. Data lengkap penting untuk analisis yang menyeluruh.
- Null values
 - Sebagian data terdefinisi nilainya, sebagian lagi null (unknown/tidak terdefinisi)

Problems with Data

- Huge outliers
 - Outlier: data point that differs significantly from other observations.
 - Outlier mungkin terjadi karena kesalahan pada proses pengambilan data atau eksperimen
 - Outlier dapat menyebabkan masalah pada data analysis à menyebabkan akurasi analysis menjadi rendah
- Out-of-date data
 - Data yang sudah tidak akurat pada saat analysis dilakukan
- Artificial entries
 - Banyak data buatan ditambahkan ke data asli dalam rangka kebutuhan testing (misalnya)
- Irregular Spacings
 - Pengukuran data sering dilakukan dalam jarak/jangka yang regular. Misalnya: lalu lintas suatu website diambil per jam, data temperature wilayah diambil per meter persegi
 - Jarak/jangka yang tidak regular akan menyebabkan masalah pada pemrosesan

Problems with Data - Formatting Issues

- Data berasal dari tabel-tabel dengan kolom-kolom yang berbeda
 - Kolom yang sama bisa memiliki data yang berbeda. Contoh: Jenis Kelamin bisa bernilai {"laki-laki", "perempuan"}, bisa {"M", "F"}, bisa {0,1}
- Extra whitespace: jumlah spasi yang berbeda dalam teks untuk data yang sama
 - Contoh: "ABC" dengan "A B C"
- Irregular capitalization: penggunaan huruf kapital (dan huruf kecil) yang berbeda
 - Contoh: "Bandung" atau "bandung"
- Inconsistent delimiter
 - Contoh: pada data CSV, digunakan semicolon atau comma atau tab untuk file data yang sama

Problems with Data - Formatting Issues

- Format data null (unknown) yang berbeda-beda
 - Bisa digunakan: null, N/A, atau hanya sekedar kosong
- Invalid characters: karakter-karakter tidak valid yang tidak dapat diproses oleh tools
- Data tanggal yang tidak compatible:
 - Perbedaan format: August 1, 2013 atau AUG 1, 2013 atau 2013-8-13
 - Perbedaan format berdasarkan negara: 10/9/2019 dalam format US adalah 9 Oktober 2019, sedangkan dalam format Indonesia adalah 10 September 2019

Data Explanatory: Understanding the Data



A few good generic questions to ask are as follows:

- How big is the dataset?
- Is this the entire dataset?
- Is this data representative enough? For example, maybe data was only collected for a subset of users.
- Are there likely to be gross outliers or extraordinary sources of noise? For example, 99% of the traffic from a web server might be a single denial-of-service attack.
- Might there be artificial data inserted into the dataset? This happens a lot in industrial settings.
- Are there any fields that are unique identifiers? These are the fields you might use for joining between datasets, etc.
- Are the supposedly unique identifiers actually unique? What does it mean if they aren't?
- If there are two datasets A and B that need to be joined, what does it mean if something in A doesn't match anything in B?
- When data entries are blank, where does that come from?
- How common are blank entries?

Sumber:
Cady (2017), “The Data Science Handbook”

Useful Statistics for Data Explanatory

- *Mean* (rata-rata)
- *Extreme values* (nilai ekstrim): Minimum, maximum
- *Standard Deviation* (simpangan baku)
- *Percentiles*
- *Correlation*

Mean

[Population] Mean:

- Merupakan ukuran tendensi sentral yang menggambarkan nilai tengah dari suatu distribusi probabilitas atau variabel acak yang dicirikan oleh distribusi tersebut.

$$\bar{x} = \frac{1}{n} \left(\sum_{i=1}^n x_i \right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

For example, the arithmetic mean of five values: 4, 36, 45, 50, 75 is:

$$\frac{4 + 36 + 45 + 50 + 75}{5} = \frac{210}{5} = 42.$$

Mean

Kapan Menggunakan Mean:

- Mean adalah ukuran rata-rata yang baik jika dataset memiliki nilai yang tersebar merata tanpa adanya nilai yang sangat tinggi atau sangat rendah.

Kapan Tidak Menggunakan Mean:

- Jika dataset memiliki satu atau dua nilai yang sangat tinggi atau sangat rendah, mean menjadi kurang representatif karena akan dipengaruhi oleh nilai-nilai ekstrem tersebut.
- Secara umum, mean tidak cocok digunakan sebagai ukuran rata-rata untuk data yang diukur pada skala ordinal.

Alternatif Lain Ukuran Tendensi Sentral:

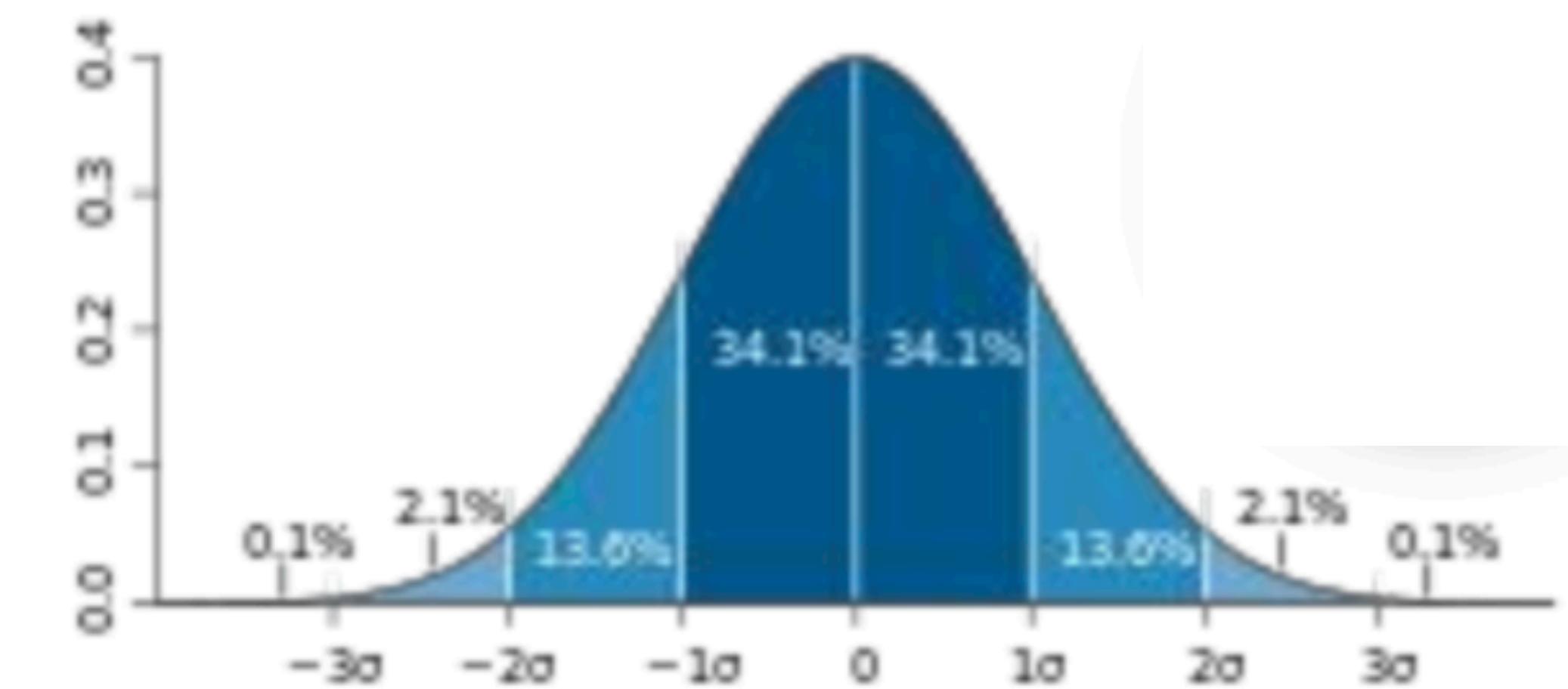
- Median: Nilai tengah dari dataset yang telah diurutkan; separuh data berada di bawahnya, dan separuh lainnya di atasnya. Median lebih tahan terhadap nilai ekstrem dibandingkan mean.
- Mode: Nilai yang paling sering muncul dalam dataset. Cocok digunakan untuk data kategori atau data dengan distribusi tidak normal.

Extreme Values

- **Minimum:** nilai terendah dari data
- **Maximum:** nilai tertinggi dari data
- Nilai ekstrem digunakan untuk menentukan *range* data [min..max]

Standar Deviation

- Merupakan ukuran seberapa besar variasi atau penyebaran dalam suatu kumpulan nilai.
 - **Standar deviasi rendah:** Menunjukkan bahwa sebagian besar nilai berada dekat dengan rata-rata.
 - **Standar deviasi tinggi:** Menunjukkan bahwa nilai-nilai lebih tersebar jauh dari rata-rata.



A plot of normal distribution (or bell-shaped curve) where each band has a width of 1 standard deviation

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2},$$

Percentiles

The near-est rank method:

$$n = \left\lceil \frac{P}{100} \times N \right\rceil$$

n: ordinal rank

P(-th): percentile ($0 \leq P \leq 100$)

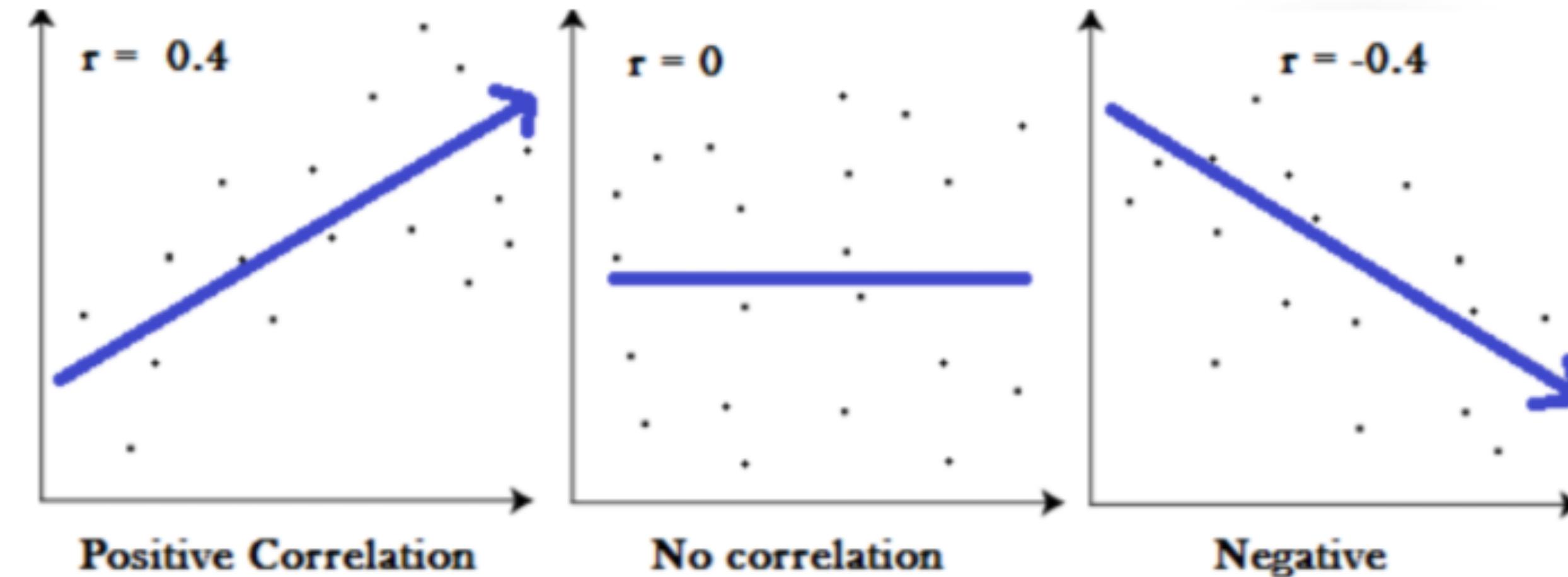
N: number of ordered values

- Merupakan ukuran dalam statistik yang menunjukkan nilai di bawah persentase tertentu dari pengamatan dalam suatu kelompok data.
- **Contoh:** Persentil ke-20 adalah nilai (atau skor) di bawahnya terdapat 20% dari pengamatan dalam kelompok data tersebut.

E.g.: The **50th** percentile of the data is **35** (from the ordered data, it ranks n = 3)

Percentile <i>P</i>	Number in list <i>N</i>	Ordinal rank <i>n</i>
5th	5	$\left\lceil \frac{5}{100} \times 5 \right\rceil = \lceil 0.25 \rceil = 1$
30th	5	$\left\lceil \frac{30}{100} \times 5 \right\rceil = \lceil 1.5 \rceil = 2$
40th	5	$\left\lceil \frac{40}{100} \times 5 \right\rceil = \lceil 2.0 \rceil = 2$
50th	5	$\left\lceil \frac{50}{100} \times 5 \right\rceil = \lceil 2.5 \rceil = 3$
100th	5	$\left\lceil \frac{100}{100} \times 5 \right\rceil = \lceil 5 \rceil = 5$

Correlation



- Teknik statistik yang digunakan untuk menunjukkan apakah dan seberapa kuat hubungan antara pasangan variabel.
- **Contoh:** Tinggi badan dan berat badan berkorelasi; orang yang lebih tinggi cenderung memiliki berat badan lebih besar dibandingkan orang yang lebih pendek.
- **Koefisien Korelasi (r):** Hasil utama dari analisis korelasi, dengan rentang nilai dari -1.0 hingga +1.0:
 - Jika r mendekati 0: Tidak ada hubungan antara variabel.
 - Jika r positif: Ketika satu variabel meningkat, variabel lainnya juga meningkat.
 - Jika r negatif: Ketika satu variabel meningkat, variabel lainnya menurun (sering disebut "korelasi invers").

**SELAMAT
BELAJAR**