Institut Shanti Bhuana
Wisdom, Faith and Integrity

No : 184/ISB/KAPRODITI/EX04/2022
Hal : Permohonan menjadi Narasumber

Kepada Yth.
Bapak Hairani, S.Kom.,M.Eng
Di tempat

Dengan Hormat,

Bersama dengan surat ini, kami dari pihak Institut Shanti Bhuana mengundang **Bapak Bapak Hairani, S.Kom.,M.Eng.** sebagai narasumber untuk memberikan kuliah umum mengenai IoT dan Data Mining di Institut Shanti Bhuana, yang akan diselenggarakan pada :

Hari/Tanggal	: Kamis, 21 April 2022
Waktu	: 10.45 Wib – 12.00 Wib
Media	: Daring (Zoom)
Tema	: “ IoT dan Data Mining”

Atas kesediaan Bapak untuk dapat menghadiri kegiatan ini, kami mengucapkan terima kasih.

Bengkayang, 20 April 2022

Kepala Program Studi
Teknologi Informasi

Azriel Christian Nurcahyo, S.Kom., M.Kom.
NIDN 1122019301

KULIAH UMUM

Kampus Merdeka
INDONESIA JAYA

"IoT dan Data Mining"
Program S1
Teknologi Informasi

Gogor Chrismass Setyawan S.Si.,M.Cs.

Pemateri Peminatan
Jaringan dan IoT

Hairani S.Kom.,M.Eng.

Pemateri Peminatan
Teknologi Web
dan Basis Data

Waktu Pelaksanaan

Hari : Kamis
Tanggal : 21 April 2022
Pukul : 10:45 - 12:00
dilanjut 14:00 -16:00 WIB
Tempat : Live Di Zoom

Wajib Bagi Mahasiswa TI Institut Shanti Bhuana

Gratis E - Certificate
Ilmu Baru yang bermanfaat

Link Pendaftaran :
s.id/KuliahUmum_TI_ISB

KULIAH UMUM
DATA MINING

HAIRANI, S.Kom., M.Eng.

21 APRIL 2022

Personal Information

Name : Hairani, S.Kom., M.Eng.
Address : Suralaga, Lombok Timur.
Hp : 087839793970
Email : Hairani@stmikbumigora.ac.id

Education

S1 : Universitas Islam Indonesia
S2 : Universitas Gadjah Mada

Research Area of Interests

Expert System, Artificial Intelligent, Data Mining, Software Testing, Decision Support System, Machine Learning, dll.

Research and Publication

1. Handling Class Imbalance Using K-Means-SMOTE and Data Mining Classification for Classification Diabetes (2020).
2. An Expert System for Diagnosis of Rheumatic Disease Types Using Forward Chaining Inference and Certainty Factor Method (2020).
3. Perancangan Sistem Pakar Diagnosis Penyakit Rematik Menggunakan Inferensi Forward Chaining Berbasis Prolog (2019).
4. Komparasi Akurasi Metode Correlated Naive Bayes Classifier dan Naive Bayes Classifier untuk Diagnosis Penyakit Diabetes (2018).
5. Aplikasi Pemetaan Kualitas Pendidikan di Indonesia Menggunakan Metode K-Means (2018).

Achievement

1. **Session Chair** The International Conference on Sustainable Information Engineering and Technology (SIET 2019)
2. **Peneliti Terbaik** pada Seminar Hasil Penelitian Dosen Pemula 2019

SERTIFIKAT PROFESI

ARTIFICIAL INTELLIGENT

Microsoft Certified
Azure AI Fundamentals

HAIRANI

Has successfully completed the requirements to be recognized as a Microsoft Certified: Azure AI Fundamentals.

Date of achievement: June 28, 2021

Microsoft

Satya Nadella
Chief Executive Officer

Microsoft CERTIFIED FUNDAMENTALS

Certification number: H877-5329

DATA SCIENCE

6464602

SERTIFIKAT KOMPETENSI
CERTIFICATE OF COMPETENCE

No. 63111 2511 6 0012648 2021

Dengan ini menyatakan bahwa,
This is to certify that

Hairani

No. Reg. TIK 1568 10794 2021

Telah kompeten pada bidang:
has been competence in the area of:

Sains Data
Data Science

Dengan Kualifikasi / Kompetensi:
With Qualification / Competency:

Associate Data Scientist

sertifikat ini berlaku untuk: 3 (tiga) tahun
this certificate is valid for: 3 (three) years

Yogyakarta, 20 November 2021

Atas Nama Badan Nasional Sertifikasi Profesi
On Behalf of Indonesian Professional Certification Authority

Lembaga Sertifikasi Profesi Teknologi Digital
Professional Certification Body Technology of Digital

Ir. Gunawan Ramli, M.Kom.
(Direktur / Director)


DATA MINING
collection of 24 icons

Manusia Memproduksi Data

- Manusia memproduksi beragam data yang **jumlah dan ukurannya sangat besar**
  - Astronomi
  - Bisnis
  - Kedokteran
  - Ekonomi
  - Olahraga
  - Cuaca
  - Financial
  - …

Pertumbuhan Data

- **Astronomi**
- Sloan Digital Sky Survey
  - New Mexico, 2000
  - 140TB over 10 years
- Large Synoptic Survey Telescope
  - Chile, 2016
  - Will acquire 140TB every five days

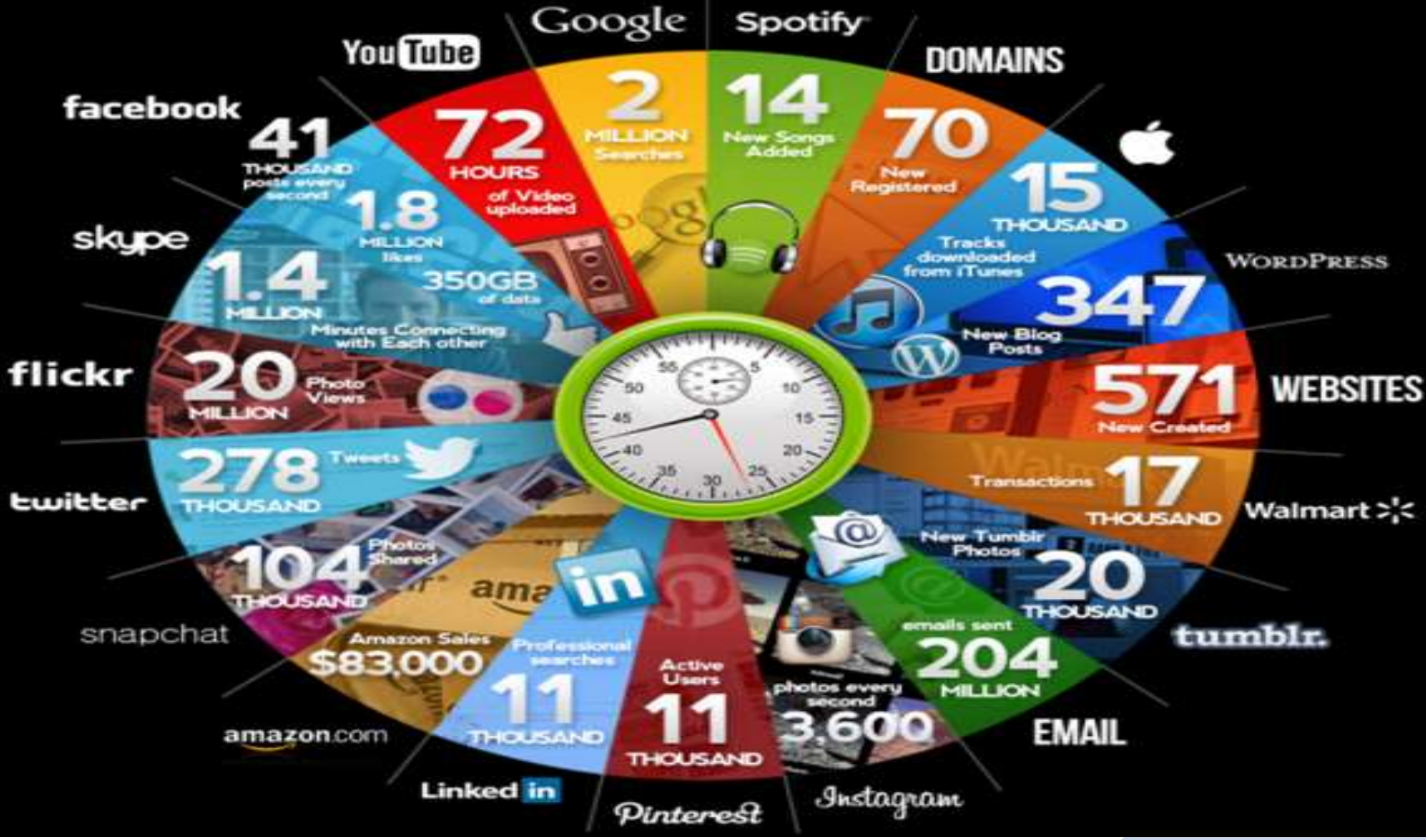
kilobyte (**kB**)	10^3
megabyte (**MB**)	10^6
gigabyte (**GB**)	10^9
terabyte (**TB**)	10^{12}
petabyte (**PB**)	10^{15}
exabyte (**EB**)	10^{18}
zettabyte (**ZB**)	10^{21}
yottabyte (**YB**)	10^{24}

Biologi dan Kedokteran

European Bioinformatics Institute (EBI)

20PB of data (genomic data doubles in size each year)

A single sequenced human genome can be around 140GB in size


Google
Spotify
YouTube
DOMAINS
facebook
41
THOUSAND
posts every second
72
HOURS
of Video uploaded
2
MILLION
Searches
14
New Songs Added
70
New Registered
15
THOUSAND
Tracks downloaded from iTunes
skype
1.8
MILLION
likes
1.4
MILLION
Minutes Connecting with Each other
350GB
of data
WordPress
347
New Blog Posts
flickr
20
MILLION
Photo Views
571
New Created
WEBSITES
278
THOUSAND
Tweets
twitter
Transactions
17
THOUSAND
Walmart
104
THOUSAND
Photos Shared
New Tumblr Photos
20
THOUSAND
snapchat
Amazon Sales
$83,000
Professional searches
11
THOUSAND
Active Users
11
THOUSAND
emails sent
204
MILLION
tumblr.
photos every second
3,600
amazon.com
EMAIL
LinkedIn
Pinterest
Instagram

Datangnya Tsunami 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}

Mobile Electronics market
7B smartphone subscriptions in 2015

Web & Social Networks generates amount of data
Google processes 100 PB per day, 3 million servers
Facebook has 300 PB of user data per day
Youtube has 1000PB video storage

Why **Need** of **Data Data Mining ?**

Data Rich

But

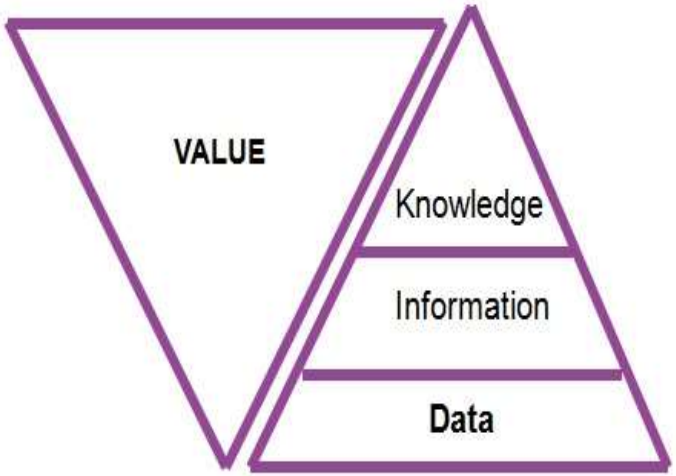
Information Poor.

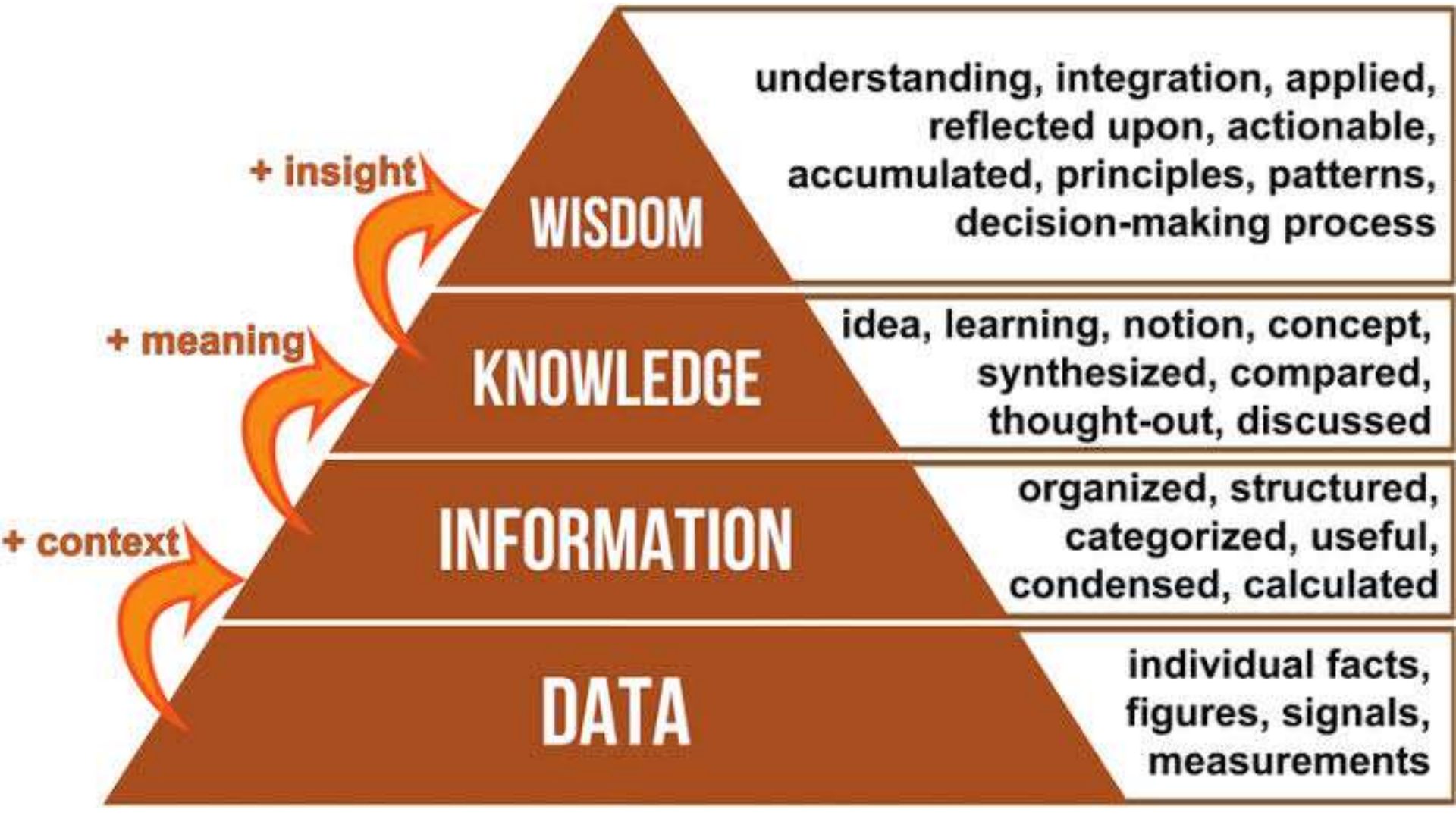
Data is pure gold in the right hands

Mengubah Data Menjadi Pengetahuan

Data harus kita olah menjadi pengetahuan supaya bisa bermanfaat bagi manusia

Dengan pengetahuan
tersebut, manusia dapat:
Melakukan estimasi dan prediksi
apa yang terjadi di depan
Melakukan analisis tentang asosiasi, korelasi dan pengelompokan antar data dan atribut
Membantu pengambilan keputusan dan pembuatan kebijakan




+ insight
+ meaning
+ context
WISDOM
KNOWLEDGE
INFORMATION
DATA
understanding, integration, applied, reflected upon, actionable, accumulated, principles, patterns, decision-making process
idea, learning, notion, concept, synthesized, compared, thought-out, discussed
organized, structured, categorized, useful, condensed, calculated
individual facts, figures, signals, measurements

Data - Informasi – Pengetahuan - Kebijakan

NIP	TGL	DATANG	PULANG
1103	02/12/2004	07:20	15:40
1142	02/12/2004	07:45	15:33
1156	02/12/2004	07:51	16:00
1173	02/12/2004	08:00	15:15
1180	02/12/2004	07:01	16:31
1183	02/12/2004	07:49	17:00

Data Kehadiran Pegawai

Data - Informasi – Pengetahuan - Kebijakan

NIP	Masuk	Alpa	Cuti	Sakit	Telat
1103	22				
1142	18	2		2	
1156	10	1	11		
1173	12	5			5
1180	10			12	

Informasi Akumulasi Bulanan Kehadiran Pegawai

Data - Informasi – Pengetahuan - Kebijakan

	Senin	Selasa	Rabu	Kamis	Jumat
Terlambat	7	0	1	0	5
Pulang Cepat	0	1	1	1	8
Izin	3	0	0	1	4
Alpa	1	0	2	0	2

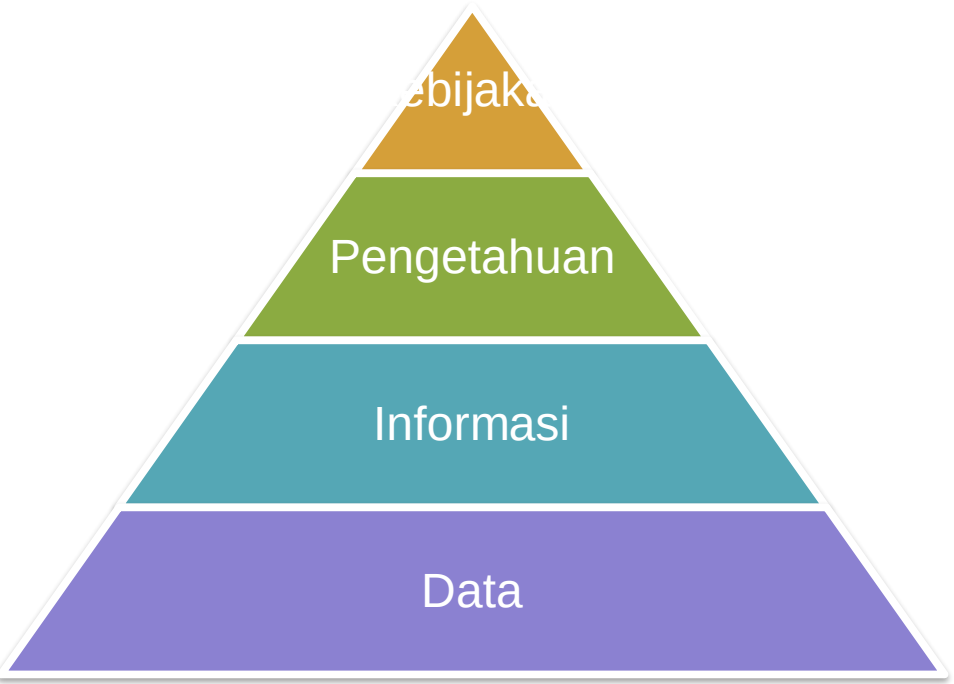
Pola Kebiasaan Kehadiran Mingguan Pegawai

Data - Informasi – Pengetahuan - Kebijakan

Kebijakan penataan jam kerja karyawan khusus untuk hari senin dan jumat

Peraturan jam kerja:
Hari Senin dimulai jam 10:00
Hari Jumat diakhiri jam 14:00
Sisa jam kerja dikompensasi ke hari lain

Data - Informasi – Pengetahuan - Kebijakan



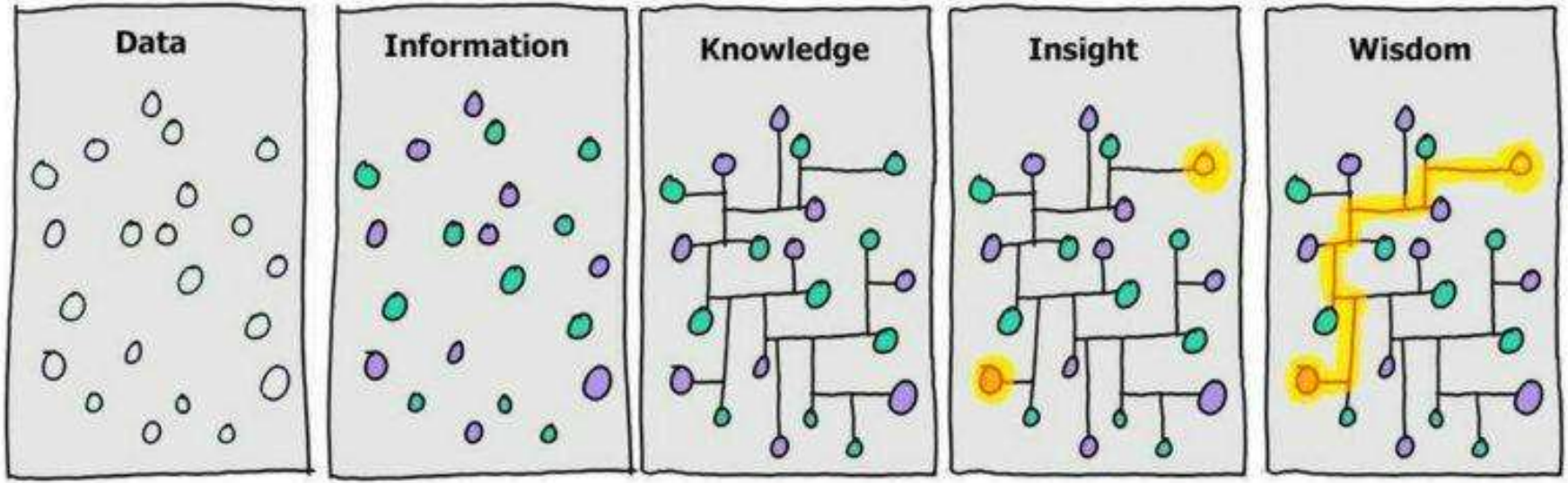
Kebijakan Penataan Jam Kerja Pegawai

Pola Kebiasaan Datang-Pulang Pegawai

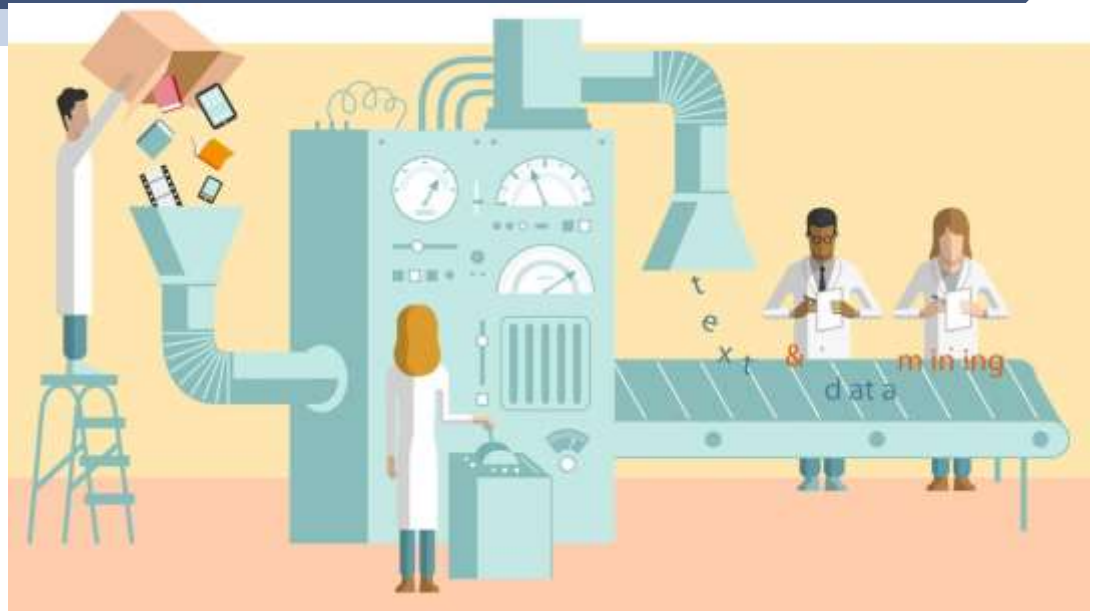
Informasi Rekap Kehadiran Pegawai

Data Absensi Pegawai

Data - Informasi – Pengetahuan - Kebijakan



Apa itu Data Mining?



Disiplin ilmu yang mempelajari **metode** untuk **mengekstrak pengetahuan** atau **menemukan pola** dari suatu data yang besar

Apa itu Data Mining?

Disiplin ilmu yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data yang besar

Ekstraksi dari data ke pengetahuan:

1. Data: fakta yang terekam dan tidak membawa arti
2. Informasi: Rekap, rangkuman, penjelasan dan statistik dari data
3. Pengetahuan: pola, rumus, aturan atau model yang muncul dari data

Nama lain data mining:

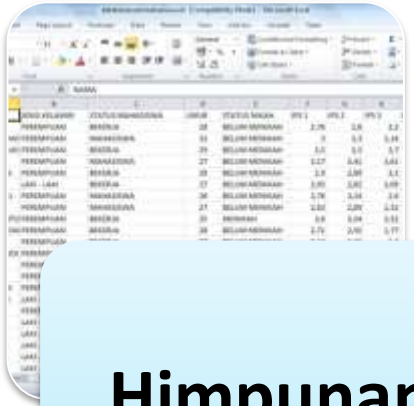
Knowledge Discovery in Database (KDD)

Big data

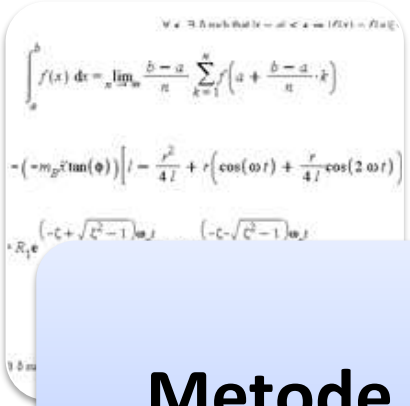
Business intelligence

Knowledge extraction

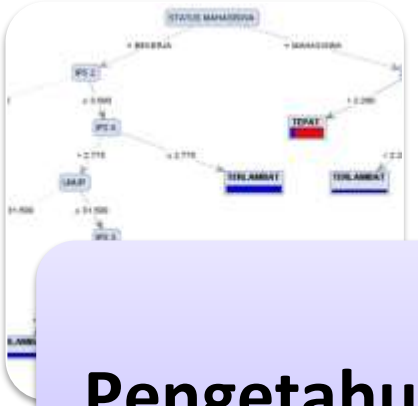
Konsep Proses Data Mining



Himpunan Data



Metode Data Mining



Pengetahuan

Definisi Data Mining

- Melakukan ekstraksi untuk mendapatkan informasi penting yang sifatnya implisit dan sebelumnya tidak diketahui, dari suatu data *(Witten et al., 2011)*
- Kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola dan hubungan dalam set data berukuran besar *(Santosa, 2007)*
- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data *(Han et al., 2011)*

From Stupid Apps to Smart Apps

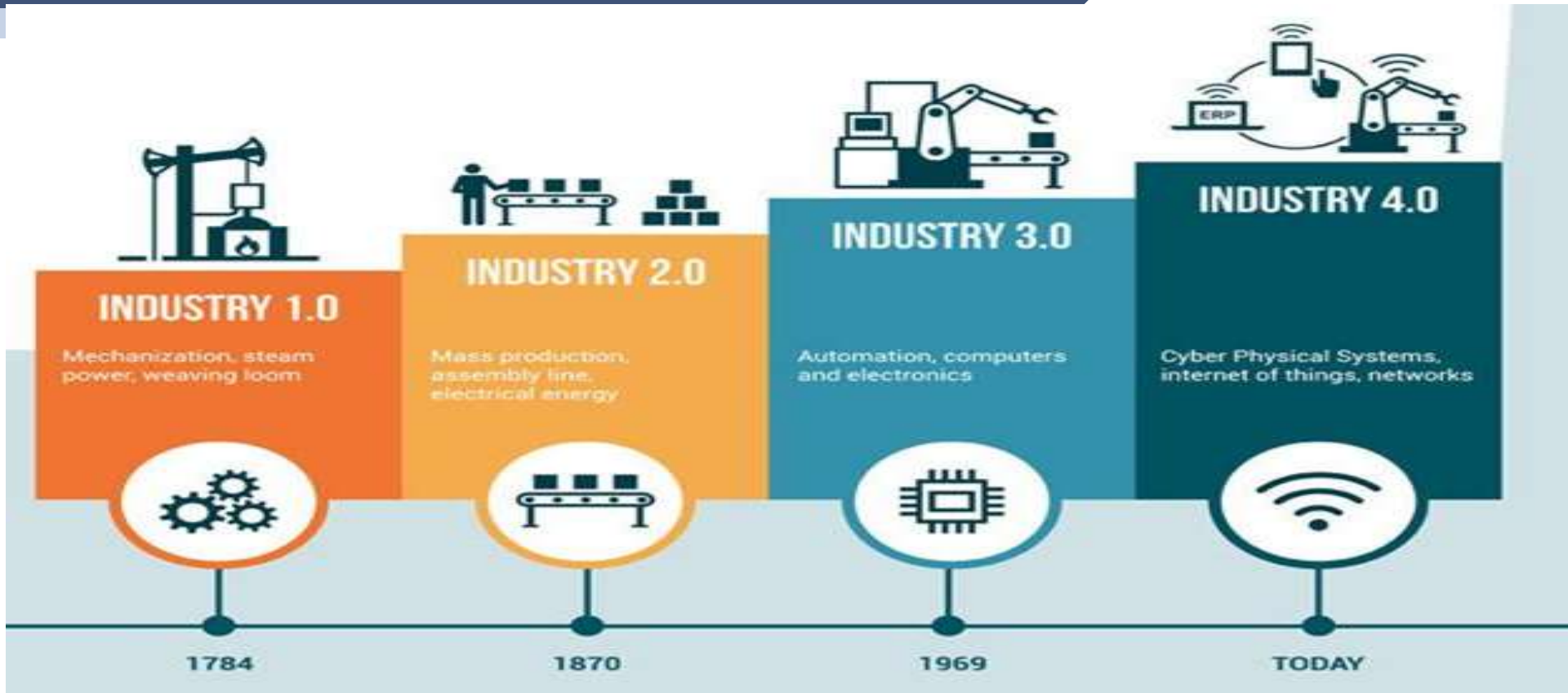
Stupid Applications

- Sistem Informasi Akademik
- Sistem Pencatatan Pemilu
- Sistem Laporan Kekayaan Pejabat
- Sistem Pencatatan Kredit

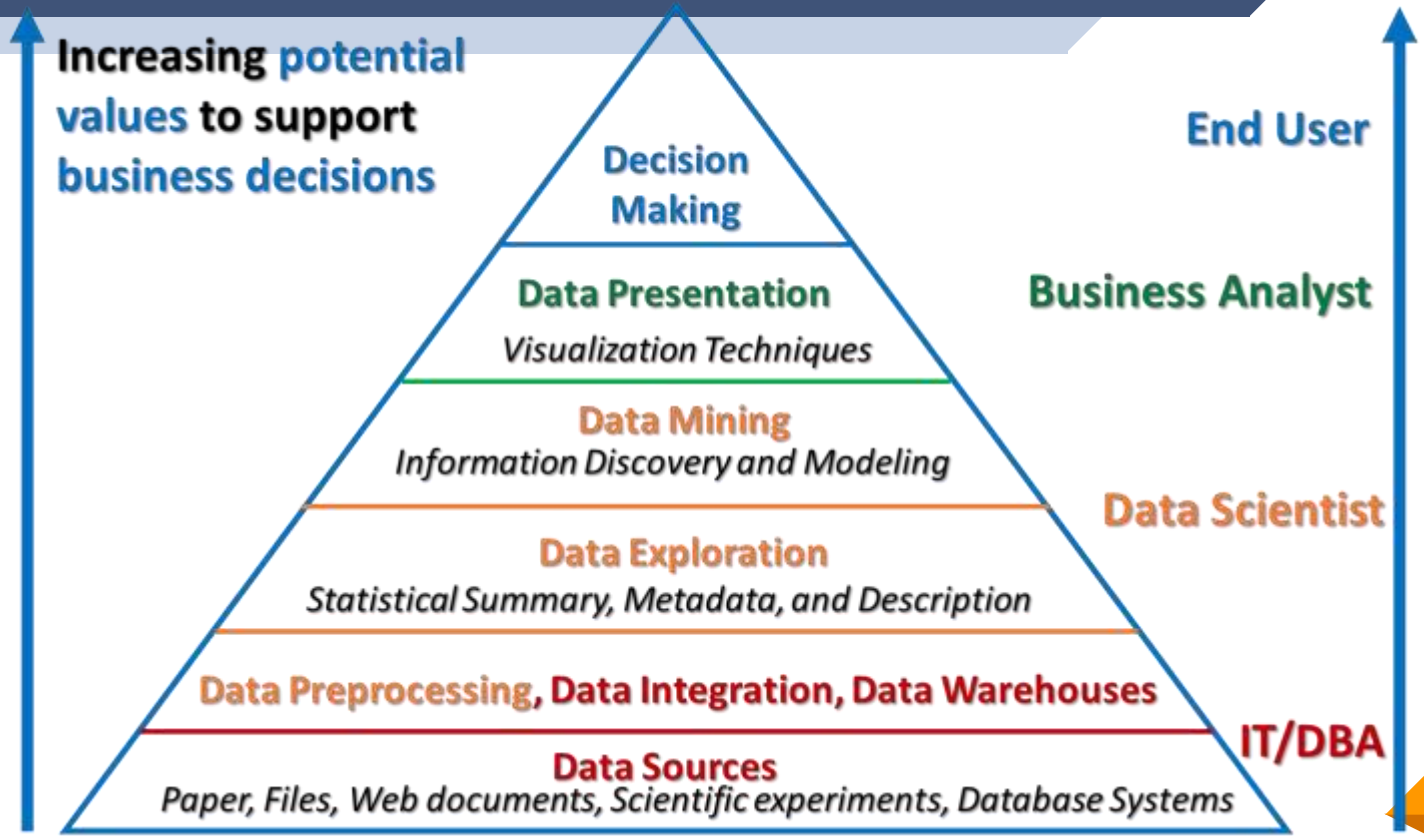
Smart Applications

- Sistem Prediksi Kelulusan Mahasiswa
- Sistem Prediksi Hasil Pemilu
- Sistem Prediksi Koruptor
- Sistem Penentu Kelayakan Kredit

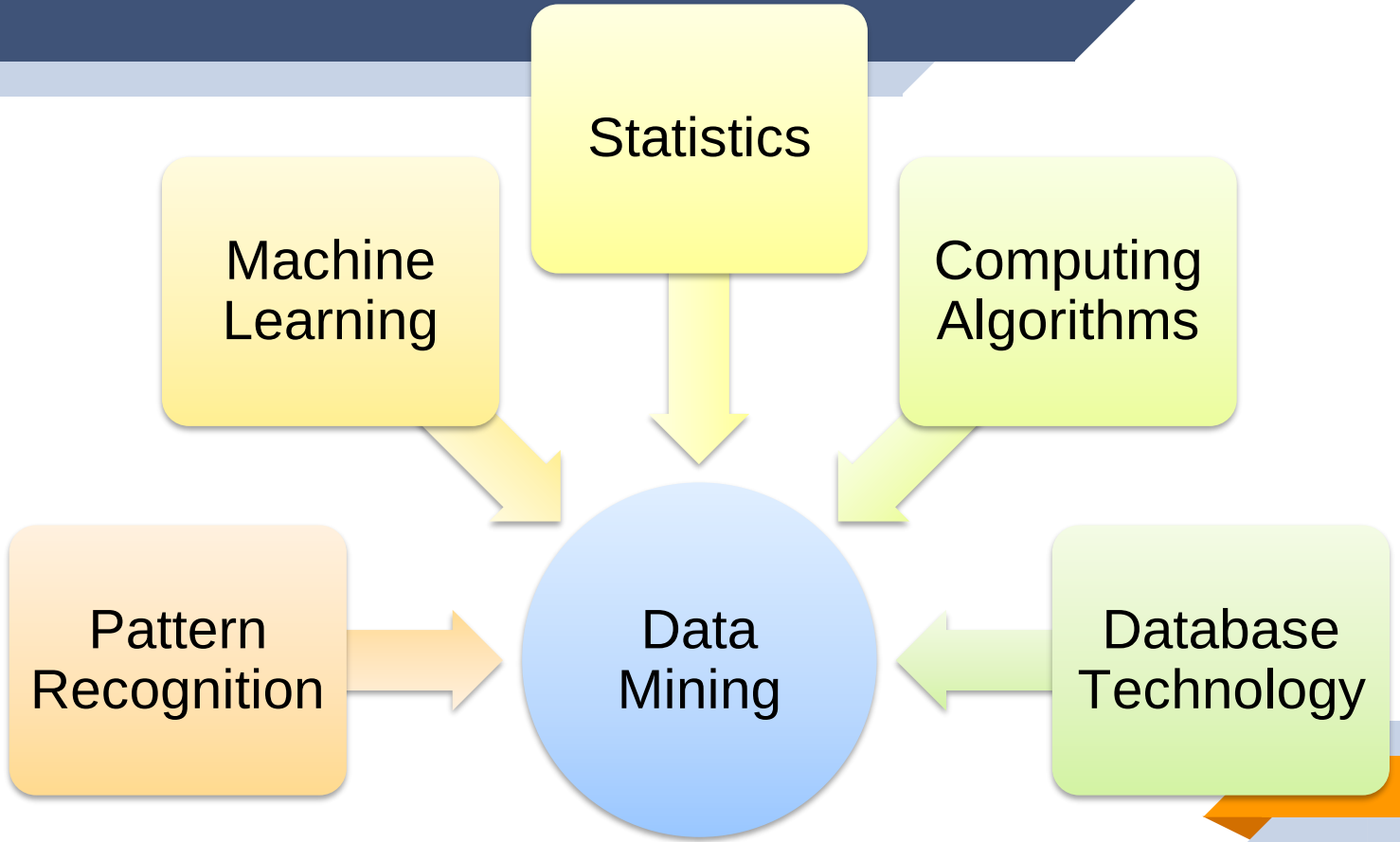
Revolusi Industri 4.0



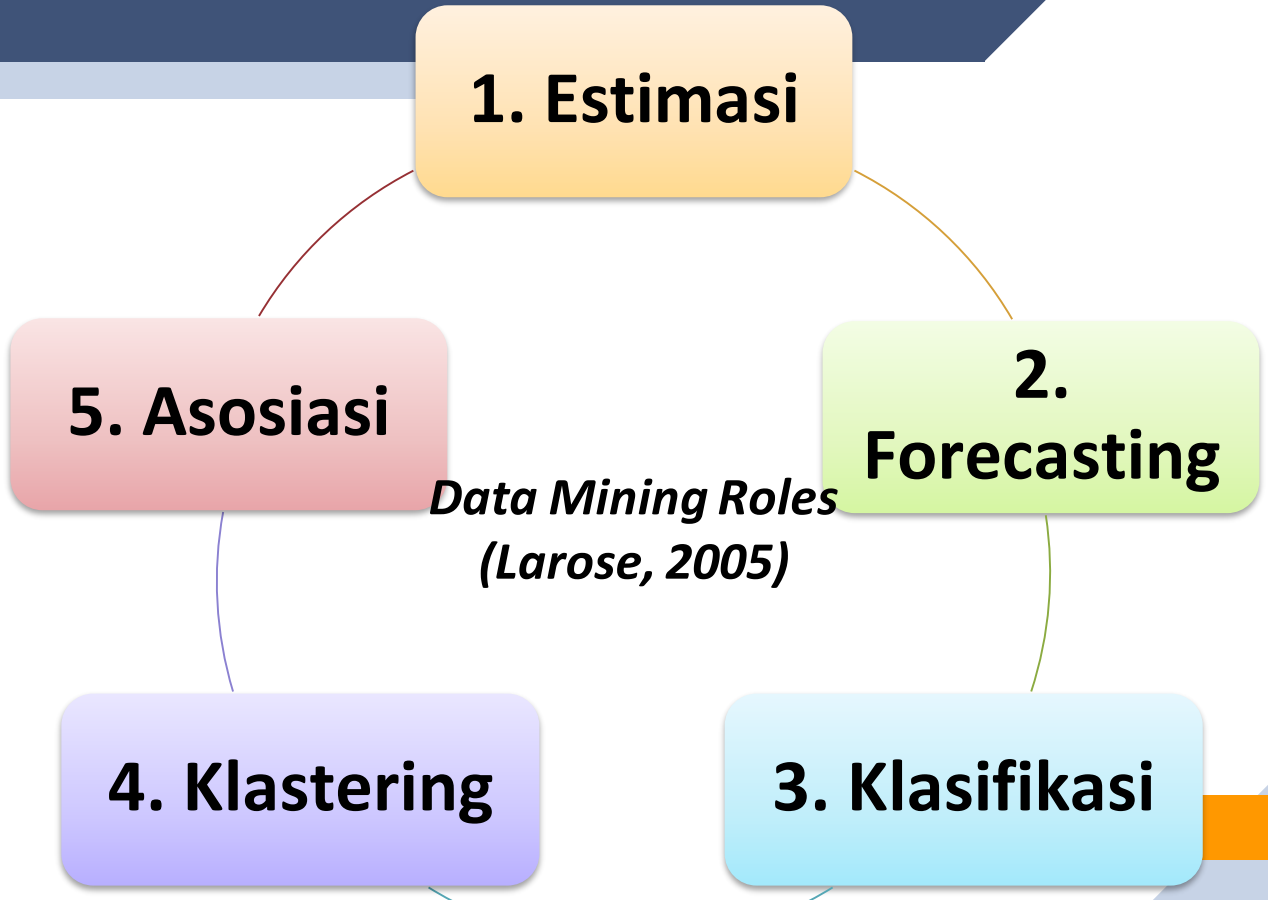
Data Mining Tasks and Roles



Hubungan Data Mining dan Bidang Lain



Peran Utama dan Metode Data Mining



1. Estimasi Waktu Pengiriman Pizza

Customer	Jumlah Pesanan (P)	Jumlah Traffic Light (TL)	Jarak (J)	Waktu Tempuh (T)
1	3	3	3	16
2	1	7	4	20
3	2	4	6	18
4	4	6	8	36
...				
1000	2	4	2	12

Pembelajaran dengan **Metode Estimasi (*Regresi Linier*)**

Waktu Tempuh (T) = 0.48P + 0.23TL + 0.5J

Pengetahuan

Contoh: Estimasi Performansi CPU

Example: 209 different computer configurations

	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performance
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
...							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

Linear regression function

PRP = -55.9 + 0.0489 MYCT + 0.0153 MMIN + 0.0056 MMAX
+ 0.6410 CACH - 0.2700 CHMIN + 1.480 CHMAX

2. Forecasting Harga Saham

Label Time Series

Row No.	Close	Date	Open	High	Low	Volume
1	1286.570	Apr 11, 2006	1296.600	1300.710	1282.960	223288000(
2	1288.120	Apr 12, 2006	1286.570	1290.930	1286.450	193810000(
3	1289.120	Apr 13, 2006	1288.120	1292.090	1283.370	189194000(
4	1285.330	Apr 17, 2006	1289.120	1292.450	1280.740	179465000(
5	1307.280	Apr 18, 2006	1285.330	1309.020	1285.330	259544000(
6	1309.930	Apr 19, 2006	1307.650	1310.390	1302.790	244731000(
7	1311.460	Apr 20, 2006	1309.930	1318.160	1306.380	251292000(
8	1311.280	Apr 21, 2006	1311.460	1317.670	1306.590	239263000(
9	1308.110	Apr 24, 2006	1311.280	1311.280	1303.790	211733000(
10	1301.740	Apr 25, 2006	1308.110	1310.790	1299.170	236638000(
11	1305.410	Apr 26, 2006	1301.740	1310.970	1301.740	250269000(
12	1309.720	Apr 27, 2006	1305.410	1315	1295.570	277201000(
13	1310.610	Apr 28, 2006	1309.720	1316.040	1306.160	241992000(

Dataset harga saham dalam bentuk time series (rentet waktu)

Pembelajaran dengan Metode **Forecasting (*Neural Network*)**

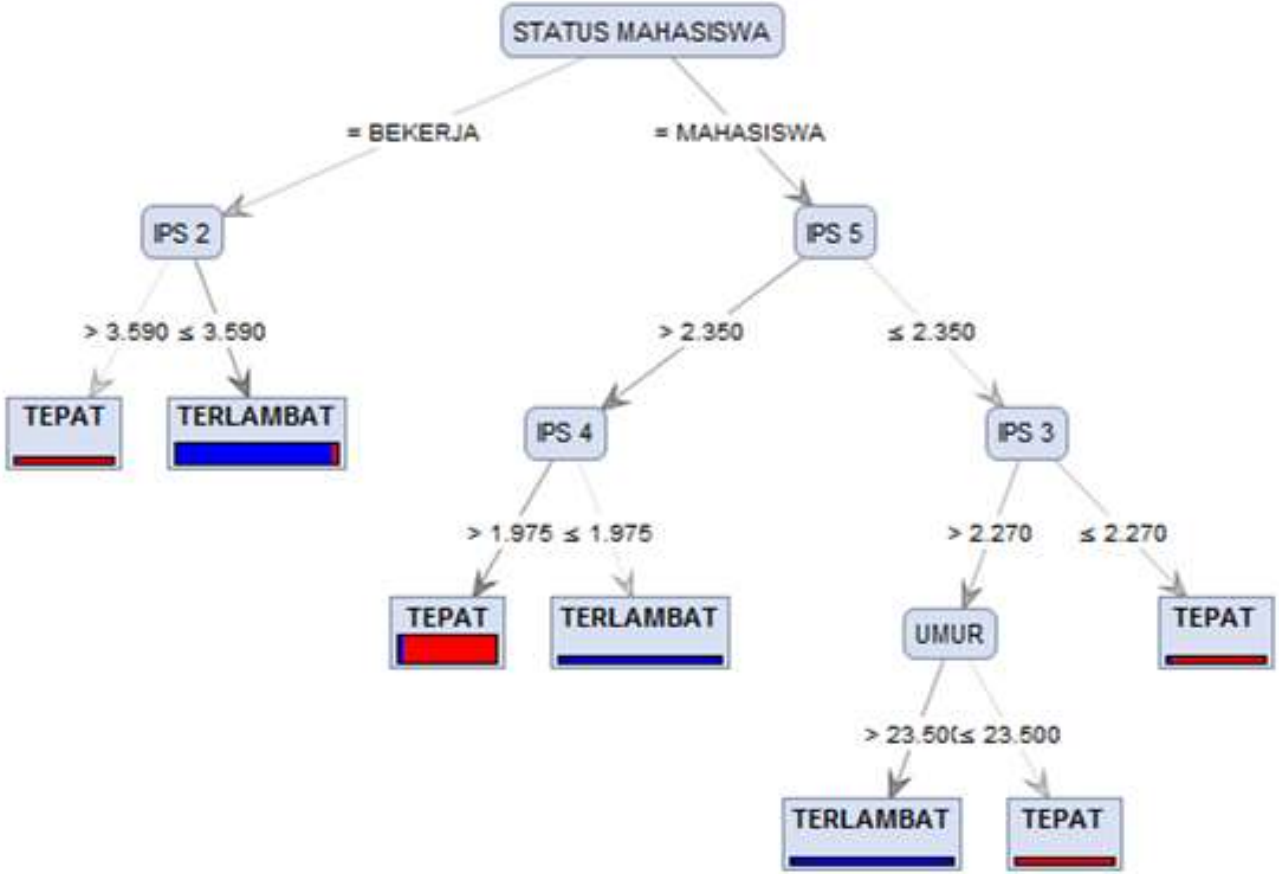
3. Klasifikasi Kelulusan Mahasiswa

Label

NIM	Gender	Nilai UN	Asal Sekolah	IPS1	IPS2	IPS3	IPS 4	...	Lulus Tepat Waktu
10001	L	28	SMAN 2	3.3	3.6	2.89	2.9		Ya
10002	P	27	SMA DK	4.0	3.2	3.8	3.7		Tidak
10003	P	24	SMAN 1	2.7	3.4	4.0	3.5		Tidak
10004	L	26.4	SMAN 3	3.2	2.7	3.6	3.4		Ya
...									
...									
11000	L	23.4	SMAN 5	3.3	2.8	3.1	3.2		Ya

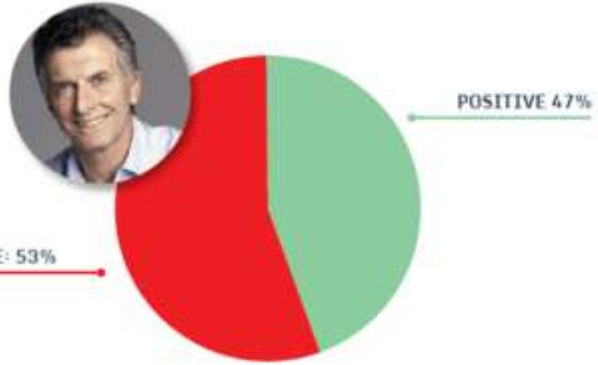
Pembelajaran dengan Metode Klasifikasi (*C4.5*)

Pengetahuan Berupa Pohon Keputusan

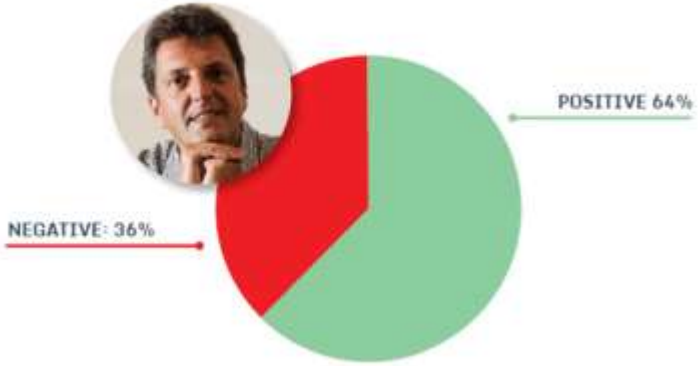


Klasifikasi Sentimen Analisis

Mauricio Macri



Sergio Massa



Daniel Scioli

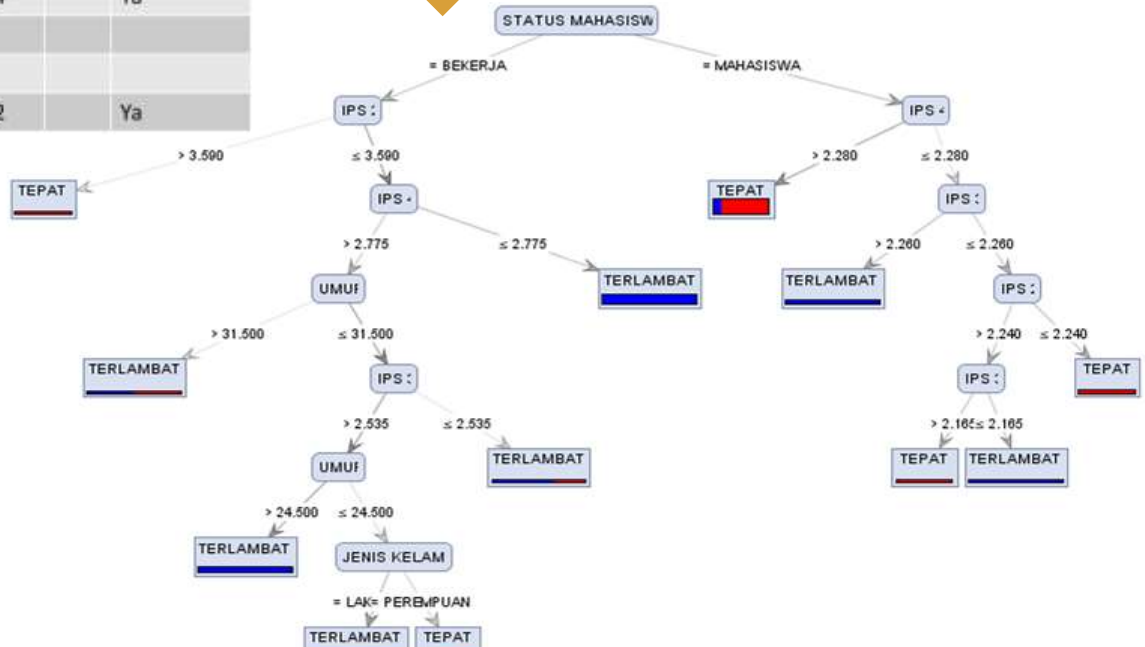
Contoh Data di Kampus

- Puluhan ribu data mahasiswa di kampus yang diambil dari sistem informasi akademik
- Apakah pernah kita ubah menjadi pengetahuan yang lebih bermanfaat? TIDAK!
- Seperti apa pengetahuan itu? Rumus, Pola, Aturan

NIM	Gender	Nilai UN	Asal Sekolah	IPS1	IPS2	IPS3	IPS 4	...	Lulus Tepat Waktu
10001	L	28	SMAN 2	3.3	3.6	2.89	2.9		Ya
10002	P	27	SMA DK	4.0	3.2	3.8	3.7		Tidak
10003	P	24	SMAN 1	2.7	3.4	4.0	3.5		Tidak
10004	L	26.4	SMAN 3	3.2	2.7	3.6	3.4		Ya
...									
...									
11000	L	23.4	SMAN 5	3.3	2.8	3.1	3.2		Ya

Prediksi Kelulusan Mahasiswa

NIM	Gender	Nilai UN	Asal Sekolah	IPS1	IPS2	IPS3	IPS 4	...	Lulus Tepat Waktu
10001	L	28	SMAN 2	3.3	3.6	2.89	2.9		Ya
10002	P	27	SMA DK	4.0	3.2	3.8	3.7		Tidak
10003	P	24	SMAN 1	2.7	3.4	4.0	3.5		Tidak
10004	L	26.4	SMAN 3	3.2	2.7	3.6	3.4		Ya
...									
...									
11000	L	23.4	SMAN 5	3.3	2.8	3.1	3.2		Ya



Contoh Data di Komisi Pemilihan Umum

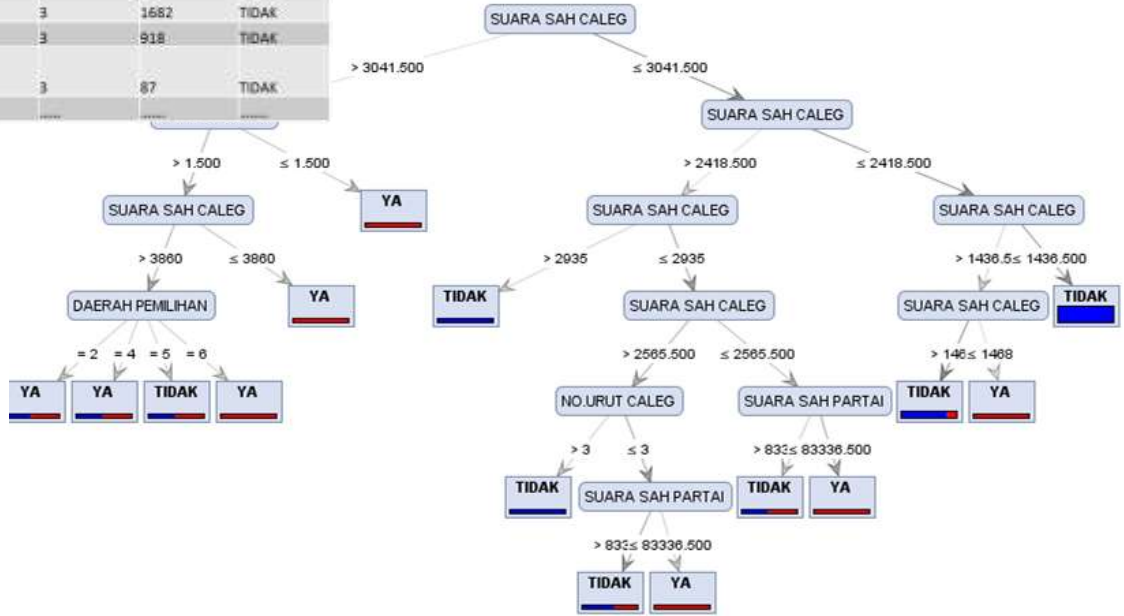
Puluhan ribu data calon anggota legislatif di KPU

Apakah pernah kita ubah menjadi pengetahuan yang lebih bermanfaat? TIDAK!

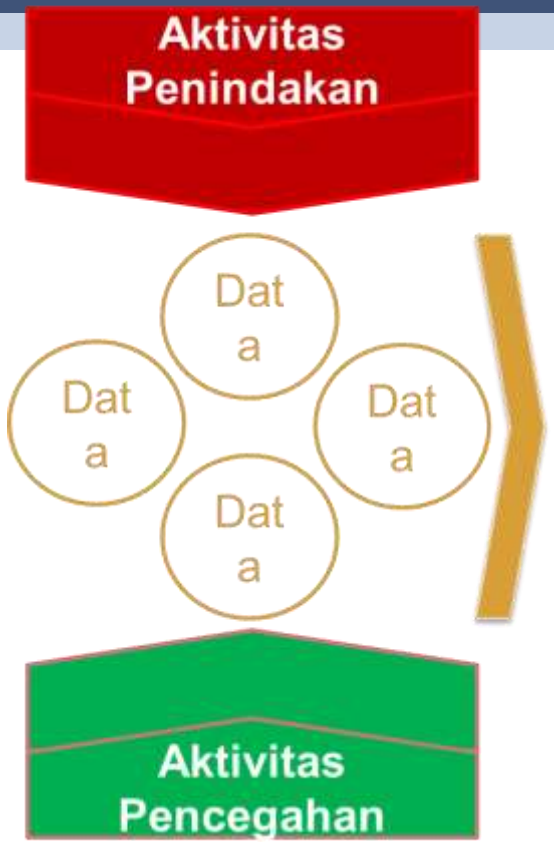
NAMA PARTAI POLITIK	NAMA CALON LEGESLATIF	JENIS KELAMIN	KECAMATAN	SUARA SAH PARTAI	DAERAH PEMILIHAN	SUARA SAH CALEG	TERPILIH ATAU TIDAK
HANURA	TOTO SUKISNO,BSc	L	LEBAKSIU	18578	1	594	TIDAK
HANURA	EDI PURYANTO,SH	L	SLAWI	18578	1	943	TIDAK
PKB	ELI RETNOWATI,SH	P	SLAWI	18578	1	1730	TIDAK
PKB	SAHYUDIN	L	DUKUHWARU	18578	1	2508	YA
GOLKAR	H.FAJAR SIGIT KUSUMAJAYA,SH	L	SLAWI	18578	2	923	TIDAK
GOLKAR	SUMIRAH	P	TARUB	18578	2	308	TIDAK
GOLKAR	DARYOTO	L	TARUB	18578	2	54	TIDAK
PKS	KHAPIP APRONI,S.Pdi	L	BOJONG	18578	3	1682	TIDAK
PKS	ENDANG SUCI RAHAYU	P	JATINEGARA	18578	3	918	TIDAK
PDI-P	KH.CHAFIDZ ISA MUFTI ,LC	L	SLAWI	18578	3	87	TIDAK
.....	.....	.....	.....	.....	.....	.....	.....

Prediksi Calon Legislatif DKI Jakarta

NAMA PARTAI POLITIK	NAMA CALON LEGESLATIF	JENIS KELAMIN	KECAMATAN	SUARA SAH PARTAI	DAERAH PEMILIHAN	SUARA SAH CALEG	TERPILIH ATAU TIDAK
HANURA	TOTO SUKISNO,BSc	L	LEBAKSIU	18578	1	594	TIDAK
HANURA	EDI PURYANTO,SH	L	SLAWI	18578	1	943	TIDAK
PKB	ELI RETNOWATI,SH	P	SLAWI	18578	1	1730	TIDAK
PKB	SAHYUDIN	L	DUKUHWARU	18578	1	2508	YA
GOLKAR	H.FAJAR SIGIT KUSUMAJAYA,SH	L	SLAWI	18578	2	923	TIDAK
GOLKAR	SUMIRAH	P	TARUB	18578	2	308	TIDAK
GOLKAR	DARYOTO	L	TARUB	18578	2	54	TIDAK
PKS	KHAPIP APRONI,S.Pdi	L	BOJONG	18578	3	1682	TIDAK
PKS	ENDANG SUCI RAHAYU	P	JATINEGARA	18578	3	918	TIDAK
PDI-P	KH.CHAFIDZ ISA MUFTI ,LC	L	SLAWI	18578	3	87	TIDAK
.....	.....	.....	.....	.....	.....	.....	.....



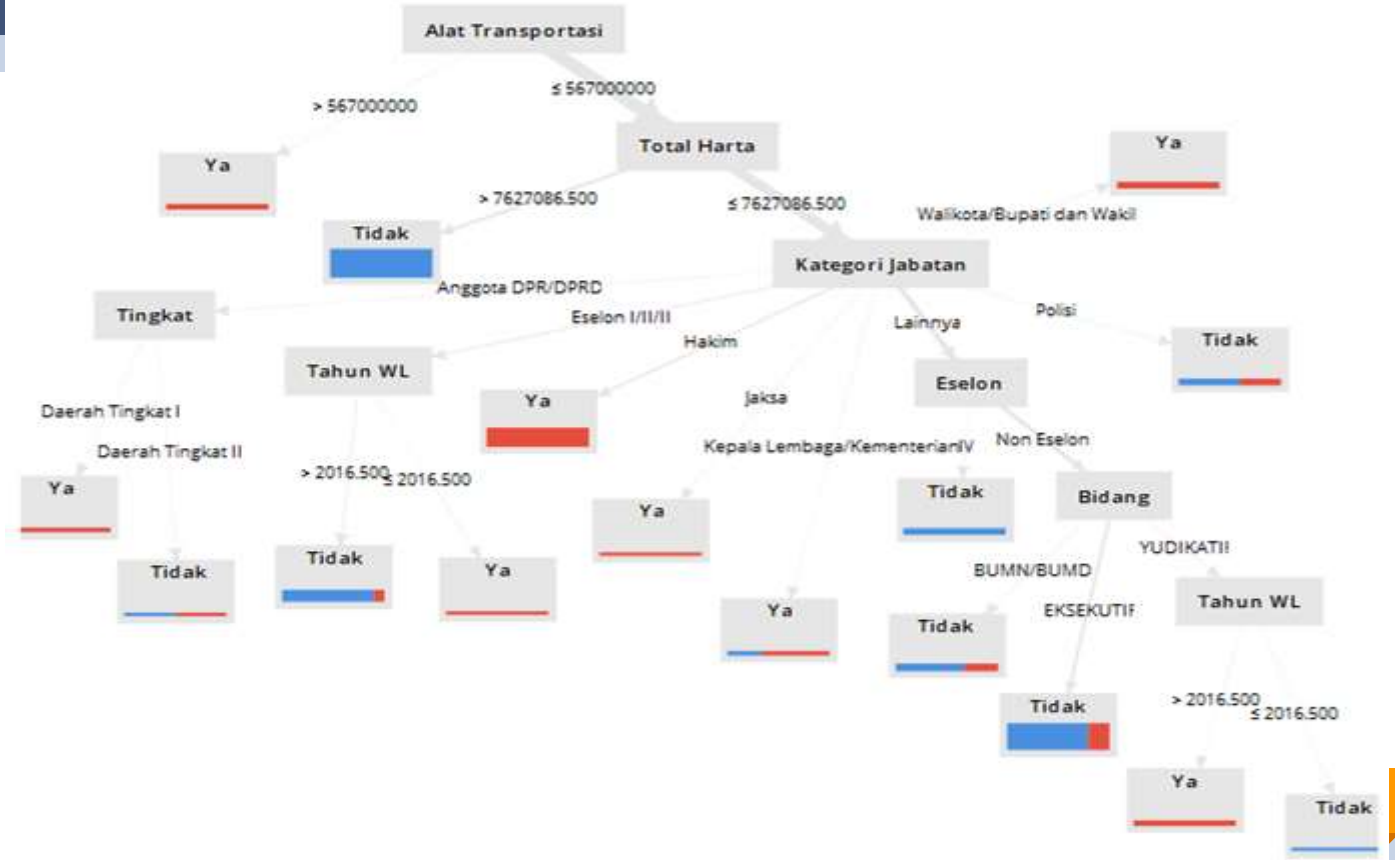
Profiling dan Prediksi Koruptor



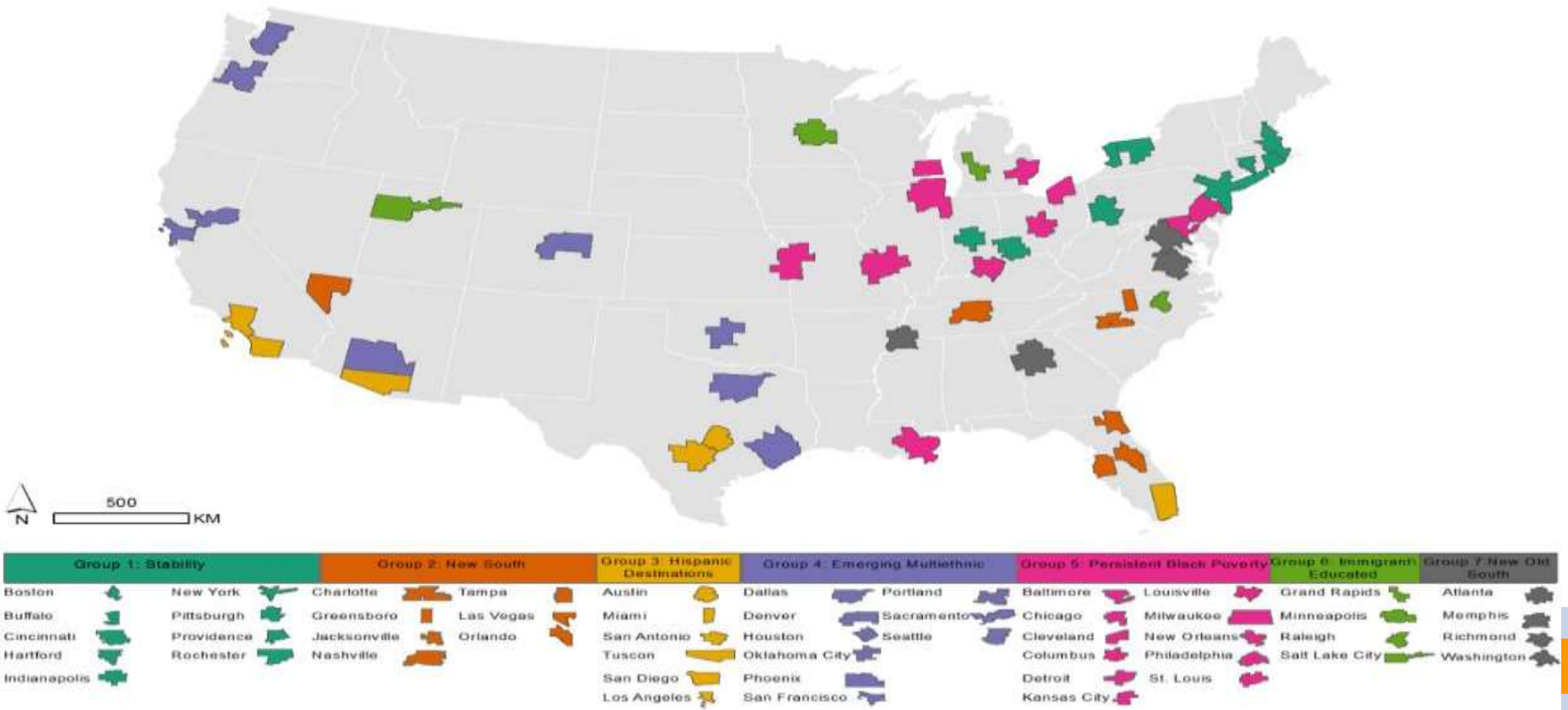
Pengetahuan

- Prediksi dan klastering calon tersangka koruptor
- Asosiasi atribut tersangka koruptor
- Prediksi pencucian uang
- Estimasi jenis dan jumlah tahun hukuman

Pola Profil Tersangka Koruptor



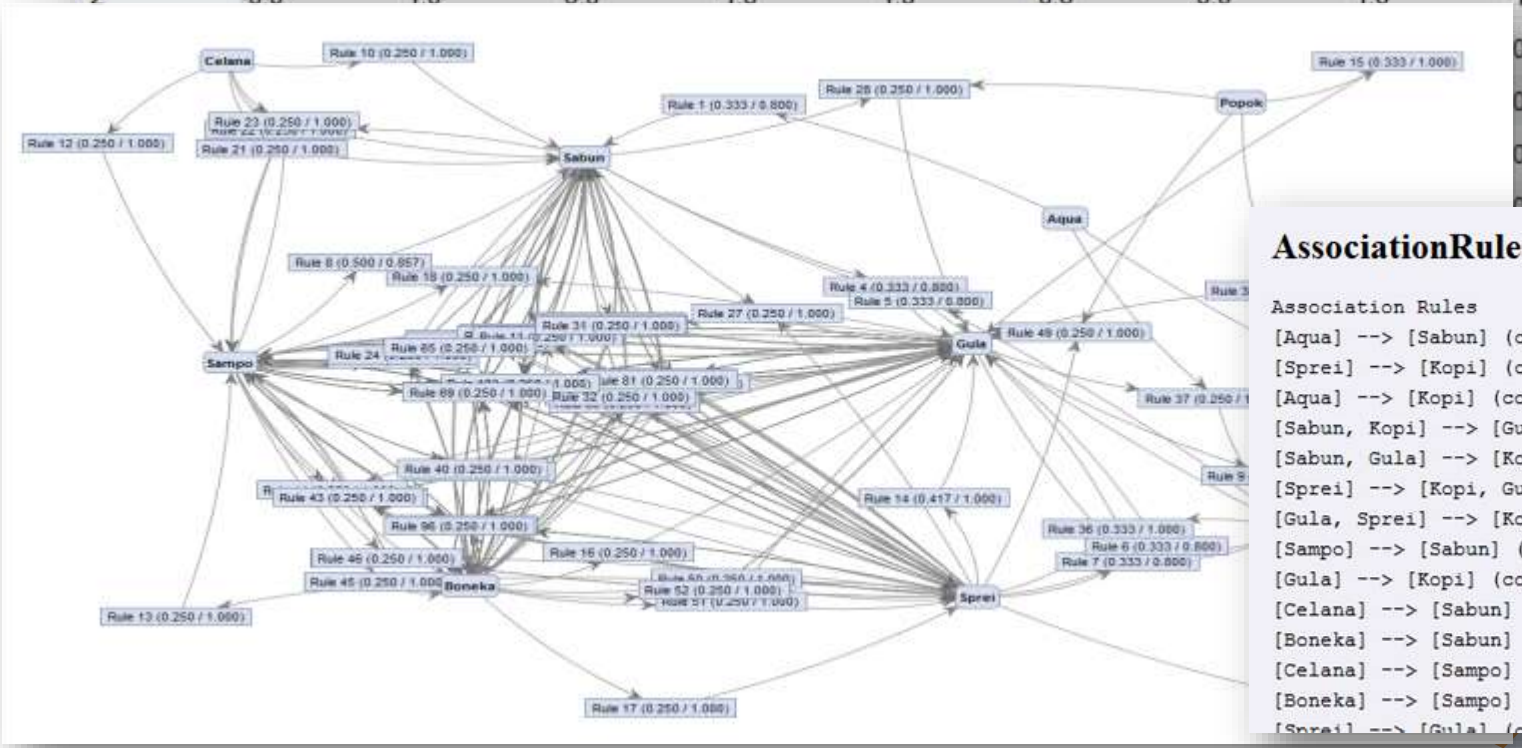
Klasterisasi Tingkat Kemiskinan



Pola Aturan Asosiasi dari Data Transaksi

ExampleSet (12 examples, 0 special attributes, 10 regular attributes)

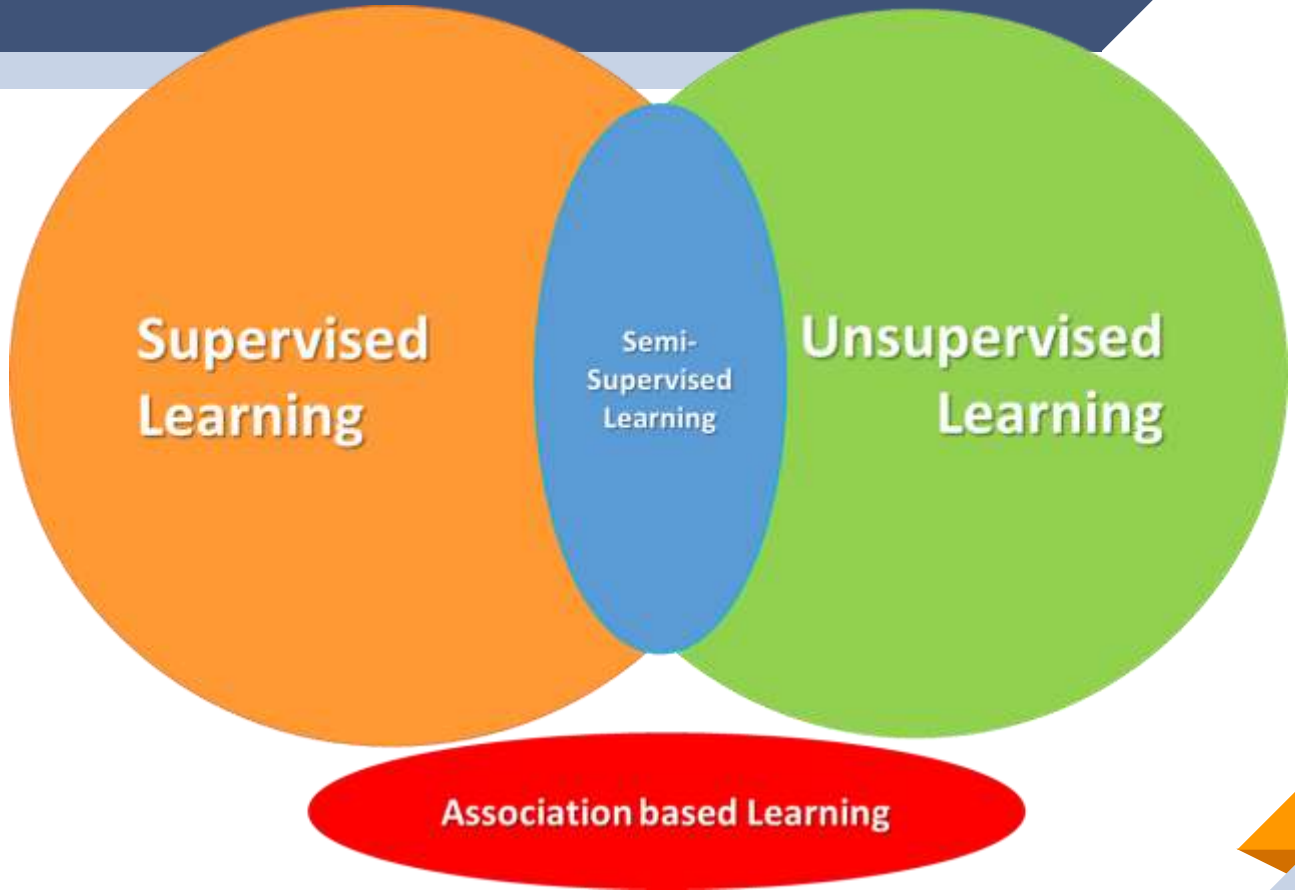
Row No.	Gula	Kopi	Aqua	Popok	Sprei	Sabun	Sampo	Kemeja	Celana	Boneka
1	1.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0
2	0.0	1.0	0.0	1.0	1.0	0.0	0.0	1.0	1.0	1.0
									0.0	1.0
									0.0	0.0
									0.0	0.0



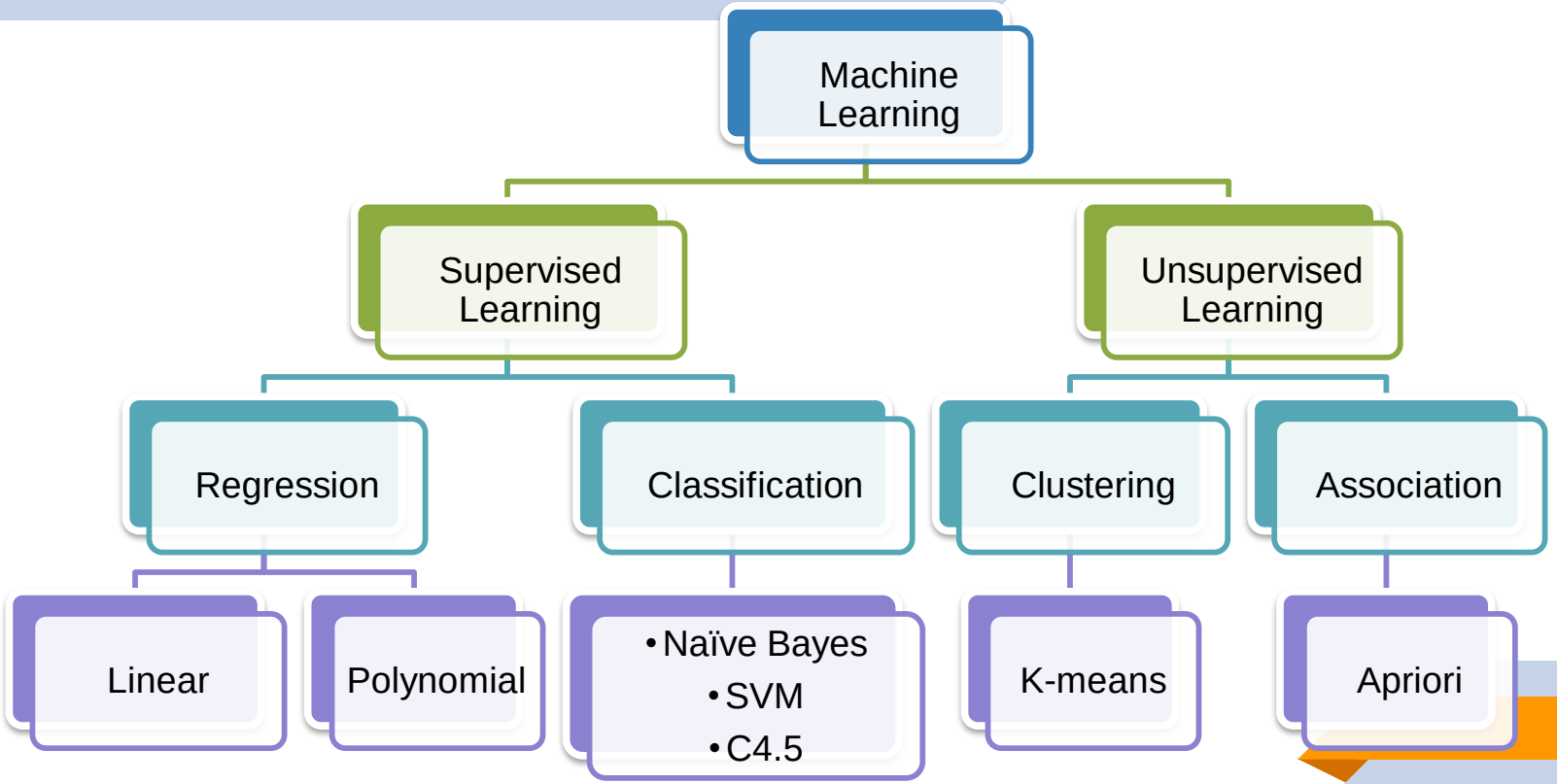
AssociationRules

```
Association Rules
[Aqua] --> [Sabun] (confidence: 0.800)
[Sprei] --> [Kopi] (confidence: 0.800)
[Aqua] --> [Kopi] (confidence: 0.800)
[Sabun, Kopi] --> [Gula] (confidence: 0.800)
[Sabun, Gula] --> [Kopi] (confidence: 0.800)
[Sprei] --> [Kopi, Gula] (confidence: 0.800)
[Gula, Sprei] --> [Kopi] (confidence: 0.800)
[Sampo] --> [Sabun] (confidence: 0.857)
[Gula] --> [Kopi] (confidence: 0.857)
[Celana] --> [Sabun] (confidence: 1.000)
[Boneka] --> [Sabun] (confidence: 1.000)
[Celana] --> [Sampo] (confidence: 1.000)
[Boneka] --> [Sampo] (confidence: 1.000)
[Sprei] --> [Gula] (confidence: 1.000)
```

Metode Learning Algoritma Data Mining



Metode Data Mining



1. Supervised Learning

- Pembelajaran dengan guru, data set memiliki target/label/class
- Sebagian besar algoritma data mining (estimation, prediction/forecasting, classification) adalah supervised learning
- Algoritma melakukan proses belajar berdasarkan nilai dari variabel target yang terasosiasi dengan nilai dari variable prediktor

Dataset dengan Class

Attribute/Feature/Dimension

Class/Label/Target

	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	*Iris setosa*
2	4.9	3.0	1.4	0.2	*Iris setosa*
3	4.7	3.2	1.3	0.2	*Iris setosa*
4	4.6	3.1	1.5	0.2	*Iris setosa*
5	5.0	3.6	1.4	0.2	*Iris setosa*
...					
51	7.0	3.2	4.7	1.4	*Iris versicolor*
52	6.4	3.2	4.5	1.5	*Iris versicolor*
53	6.9	3.1	4.9	1.5	*Iris versicolor*
54	5.5	2.3	4.0	1.3	*Iris versicolor*
55	6.5	2.8	4.6	1.5	*Iris versicolor*
...					
101	6.3	3.3	6.0	2.5	*Iris virginica*
102	5.8	2.7	5.1	1.9	*Iris virginica*
103	7.1	3.0	5.9	2.1	*Iris virginica*

Nominal

Numerik

2. Unsupervised Learning

- Algoritma data mining mencari pola dari semua variable (atribut)
- Variable (atribut) yang menjadi target/label/class tidak ditentukan (tidak ada)
- Algoritma clustering adalah algoritma unsupervised learning

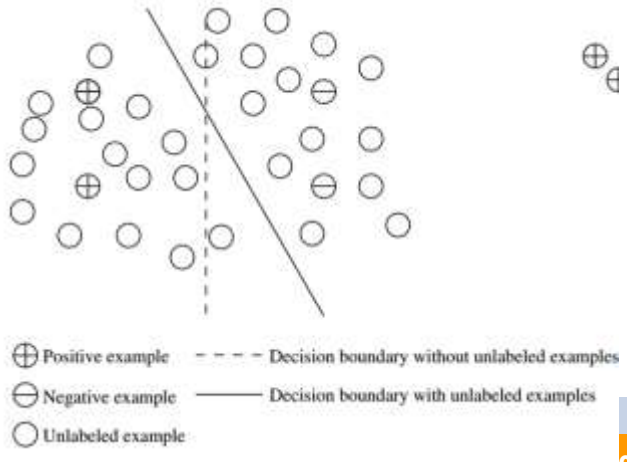
Dataset tanpa Class

Attribute/Feature/Dimension

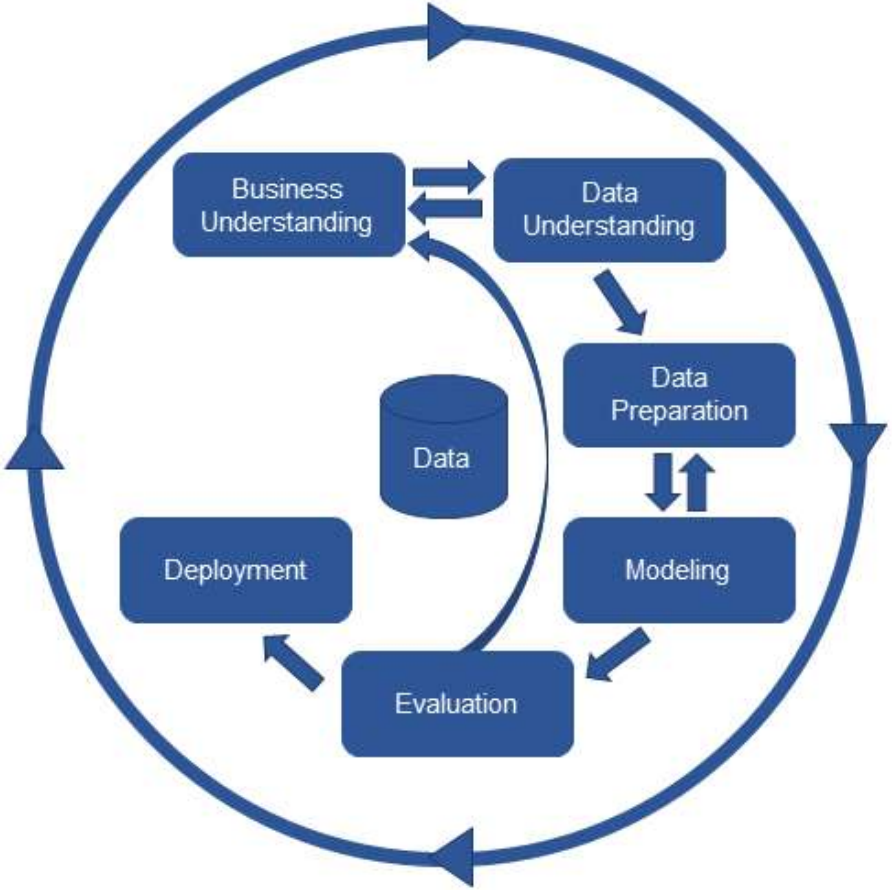
	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
...				
51	7.0	3.2	4.7	1.4
52	6.4	3.2	4.5	1.5
53	6.9	3.1	4.9	1.5
54	5.5	2.3	4.0	1.3
55	6.5	2.8	4.6	1.5
...				
101	6.3	3.3	6.0	2.5
102	5.8	2.7	5.1	1.9
103	7.1	3.0	5.9	2.1

3. Semi-Supervised Learning

- **Semi-supervised learning** adalah metode data mining yang menggunakan data dengan label dan tidak berlabel sekaligus dalam proses pembelajarannya
- Data yang memiliki kelas digunakan untuk membentuk model (pengetahuan), data tanpa label digunakan untuk membuat batasan antara kelas



TAHAPAN CRISDM UNTUK INDUSTRI



1. Business Understanding:
Menentukan Masalah Bisnis

Kasus: Kegagalan Kredit

Problem:
Bagaimana menurunkan NPL suatu bank
Pertanyaan:
Bagaimana memperbaiki perhitungan *Credit score*
Measurable outcomes:
% Penurunan kredit gagal bayar

1. Business Understanding:
Menentukan Kebutuhan Data

Data apa yang diperlukan?
Dari mana bisa diperoleh?

Struktur Data: Bagaimana deskripsi data (atribut) yang diperlukan

Jumlah Data: Berapa banyak (record) data yang diperlukan

Sumber Data: Darimana data bisa diperoleh? Apakah sudah tersedia?

- Internal: Sistem Informasi/ ERP, Excel, dokumen
- Eksternal: Web API, Web Scraping
- Dataset via public data
- Dataset via open data

1. Business Understanding:
Merencanakan Manajemen Proyek

Bagaimana rencana pelaksanaan projeknya?

Cost Benefit Analysis: Apakah menguntungkan untuk melakukannya?

Situation Assessment: Analisa keadaan organisasi

Project Plan: Scope (WBS), Time, Schedule, Tim Pengembang

2. Data Understanding:
Mengenali / Mendalami data yang dimiliki

01 **Mengumpulkan Data**
Mengumpulkan Data yang Diperlukan

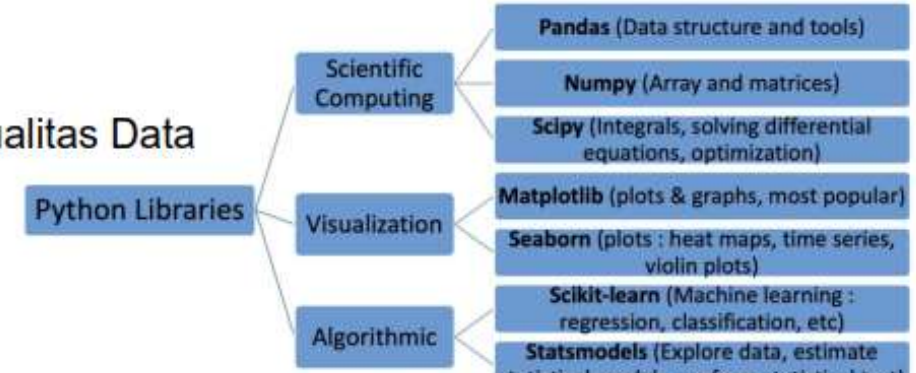
Jumlah Data (Baris dan Kolom)
Deskripsi data

02 **Menelaah data**
Menganalisa data secara eksploratif

Karakteristik atribut/ fitur
Keterkaitan antar data

03 **Memvalidasi Data**
Menilai kesesuaian kualitas data dengan masalah yang akan dipecahkan

Kualitas Data



2. Data Understanding:
Mengumpulkan Data

Mengumpulkan Data yang Diperlukan

Jumlah Data: Berapa banyak yang dapat diperoleh

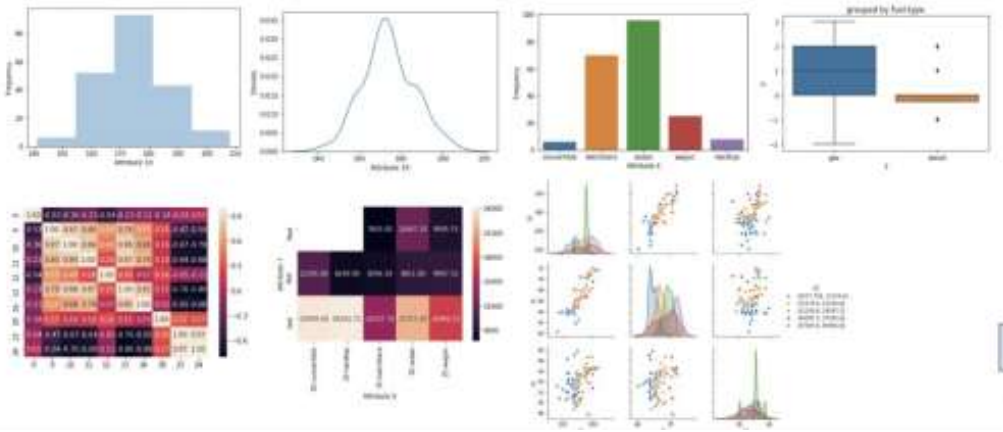
Deskripsi Data: Penjelasan arti atribut/ fitur

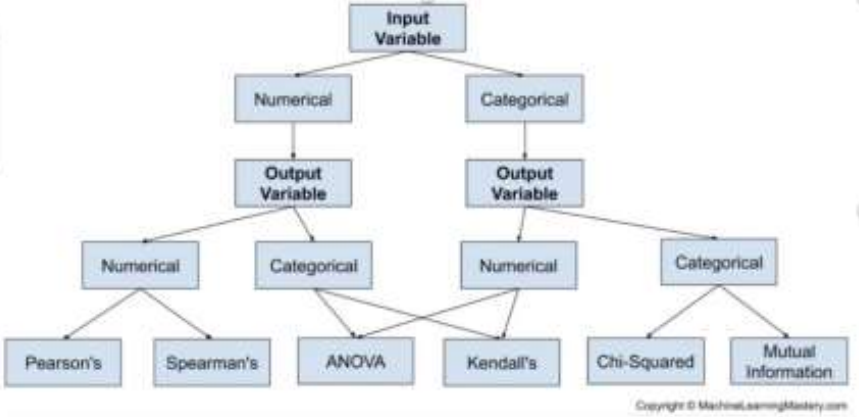
2. Data Understanding:
Menelaah Data

Menganalisa data secara eksploratif (EDA)

Karakteristik Atribut: Deskripsi data (atribut) yang diperoleh

Keterkaitan antar Data: Analisis statistik korelasi, Anova, Chi-Squared,...





2. Data Understanding:
Validasi Data

Menilai kesesuaian kualitas data de-
ngan masalah yang akan dipecahkan

Laporan Kualitas Data:

- Ukuran Data (Atribut/ fitur dan Jumlah record
- Deskripsi statistical atribut
- Relasi antar atribut (dan label)
- Visualisasi data

3. Data Preparation:
Memperbaiki Kualitas Data untuk Pemodelan

01	**Memilih dan memilah data** Memilih data yang akan dipergunakan	Rekord terpakai Atribut terpakai
02	**Membersihan Data** Meminimalkan noise (tidak lengkap, salah)	Data lengkap Data yang diperbaiki Data Pecilan
03	**Mengkonstruksi data** Menambahkan fitur dan transformasi data	Fitur tambahan (Feature Engineering) Transformasi data (standardisasi, transformasi)
04	**Integrasi Data** Menggabungkan data	Gabungan data

4. Modeling:
Mengembangkan Model

01

Membangun Skenario Pemodelan

Membuat strategi pencarian model terbaik

Pemilihan Algoritma Machine Learning (ML)
Pembagian Data
Penentuan Langkah Eksperimen

02

Membangun model

Mengembangkan model dengan Teknik ML

Eksekusi Algoritma
Pengaturan Parameter
Pengukuran Performance Metrics

4. Modeling:
Membangun Skenario Pemodelan

Membuat strategi pencarian model terbaik

Pemilihan Algoritma Machine Learning (ML)
Pembagian Data
Penentuan Langkah Eksperimen

4. Modeling:
Membangun Skenario Pemodelan

Membuat strategi pencarian model terbaik

A. Memilih Algoritma: Disesuaikan dengan Tugas Analytics yang dipilih

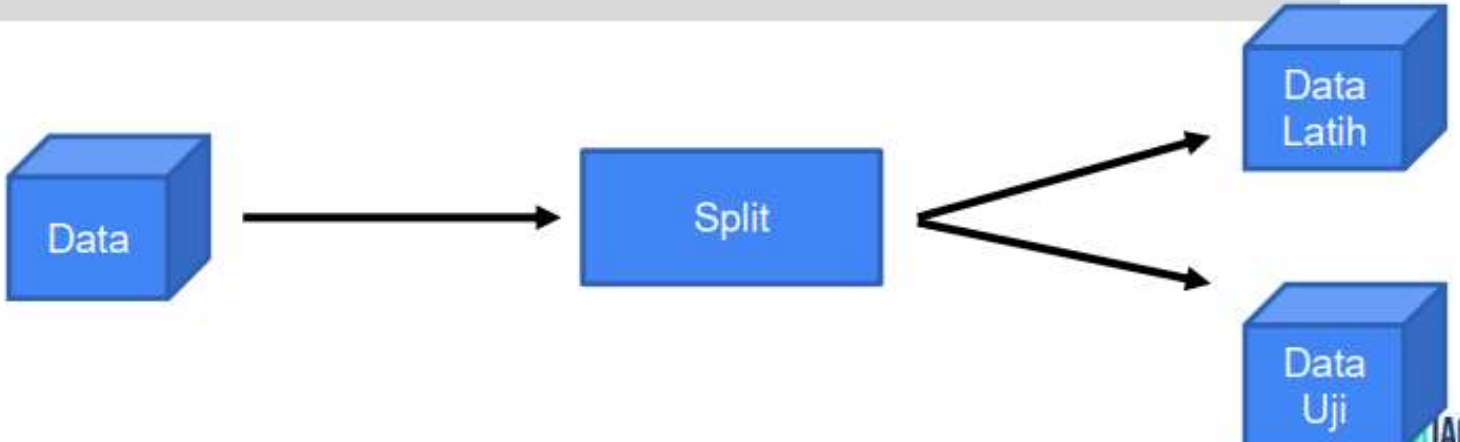
1. k-Nearest Neighbor (k-NN)
2. Naïve Bayes
3. Regression Techniques
4. Support Vector Machines (SVMs)
5. Decision Trees
6. Random Forests
7. Deep Learning Algorithms
8. . . .

4. Modeling:
Membangun Skenario Pemodelan

Membuat strategi pencarian model terbaik

B. Membagi data: Sesuai dengan ketersediaan data

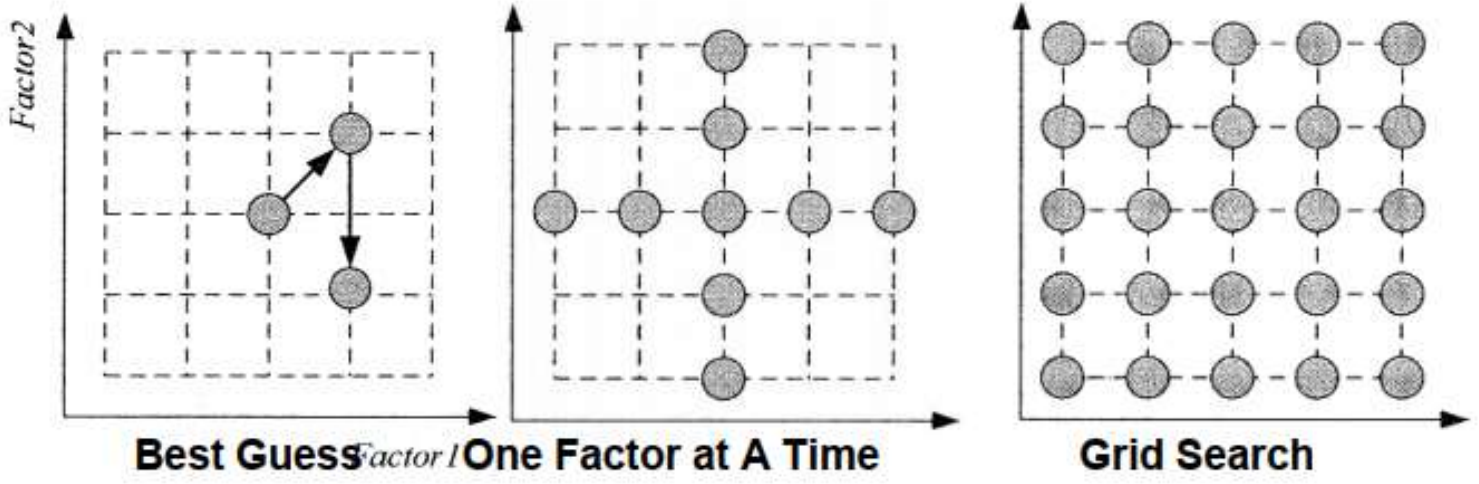
1. Data Latih: Untuk mengembangkan model
2. Data Uji: Untuk Mengukur performansi model



4. Modeling:
Membangun Skenario Pemodelan

Membuat strategi pencarian model terbaik

C. Menentukan Langkah Eksperimen: Untuk mendapatkan model terbaik secara efisien dan efektif



4. Modeling:
Membangun Model

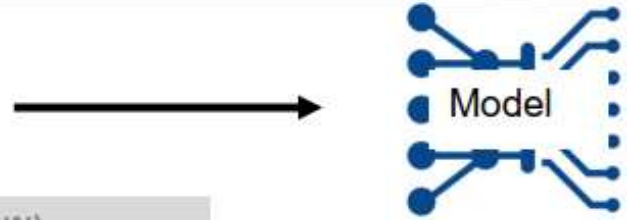
Mengembangkan model dengan Teknik ML

Pemilihan Algoritma Machine Learning (ML)
Pembagian Data
Penentuan Langkah Eksperimen

4. Modeling:
Membangun Model

Mengembangkan model dengan Teknik ML

A. Proses Pelatihan : Untuk mendapatkan model

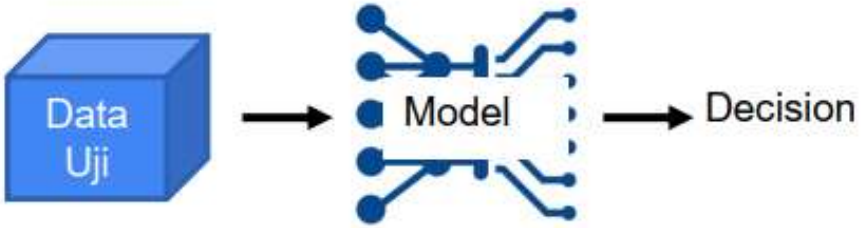


1. k-Nearest Neighbor (k-NN)
2. Naïve Bayes
3. Regression Techniques
4. Support Vector Machines (SVMs)
5. Decision Trees
6. Random Forests
7. Deep Learning Algorithms
8. . . .

4. Modeling:
Membangun Model

Mengembangkan model dengan Teknik ML

B. Proses Pengujian : Untuk mengukur Performansi



TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

$$\text{Precision} = \frac{TP}{TP + FP}$$

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

$$\text{F1-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

4. Modeling:
Model Evaluation

Mengevaluasi Performansi Model Yang Dihasilkan

01 **Mengevaluasi Model**
Mengukur performansi model

Performansi Capaian vs Target
Memilih Model terbaik

02 **Mengevaluasi Proses**
Menilai apakah proses sudah maksimal

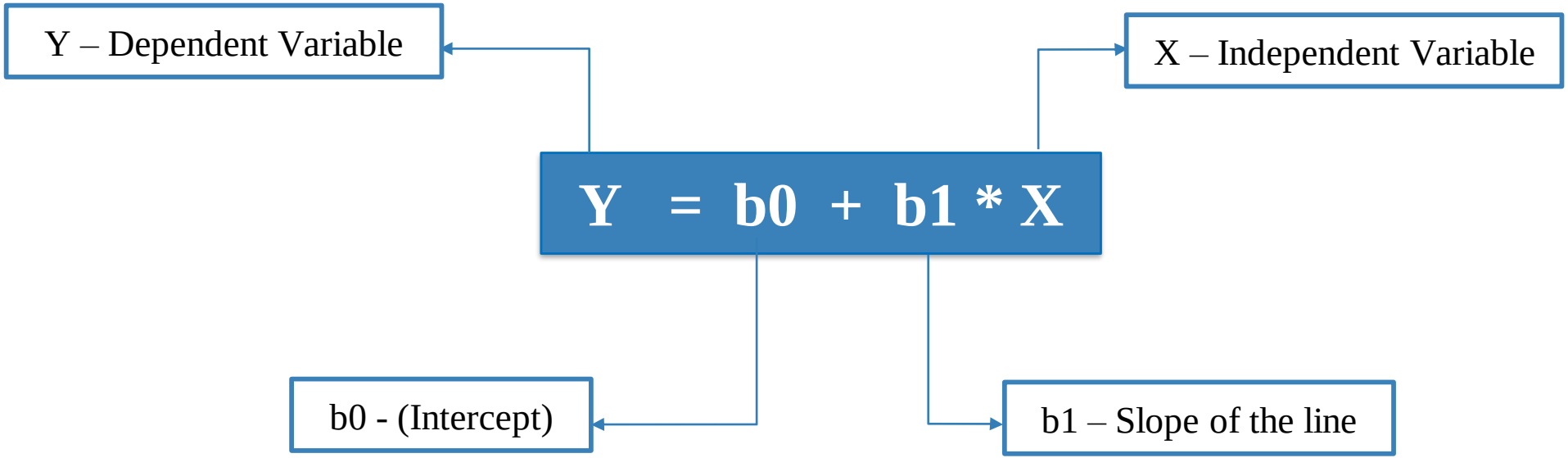
Review Proses untuk mencari batasan atau kekurangan model

Linear Regression

- Simple linear regression is useful for finding relationship between two continuous variables.
- One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship.

Linear Regression

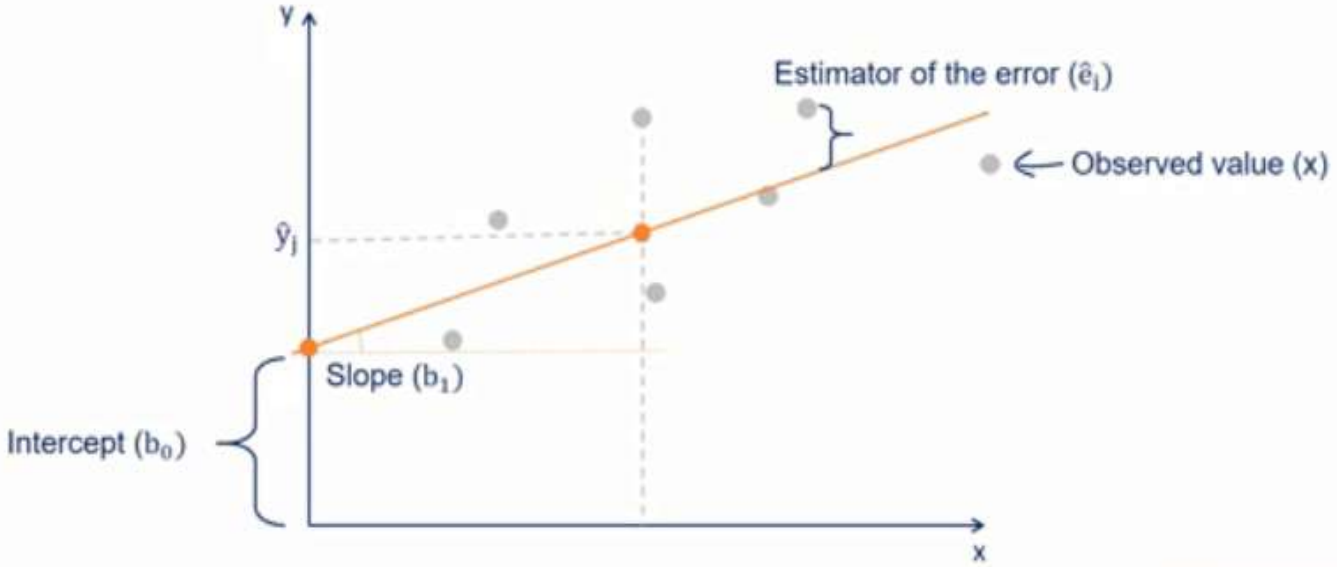
Given that,



Linear Regression

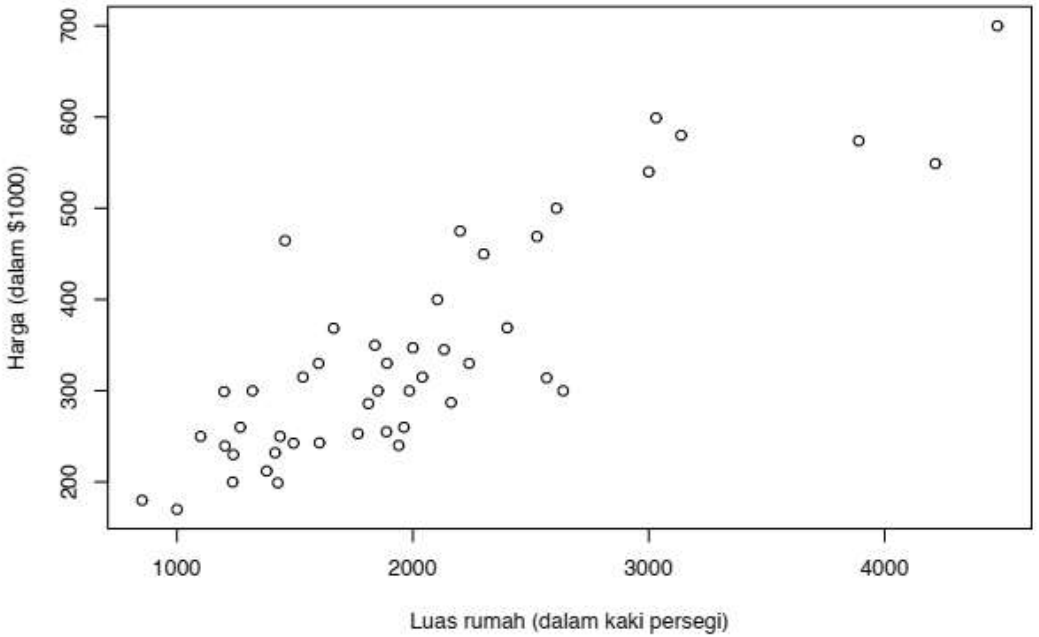
Given that,

$$\hat{y}_i = b_0 + b_1 x_i$$

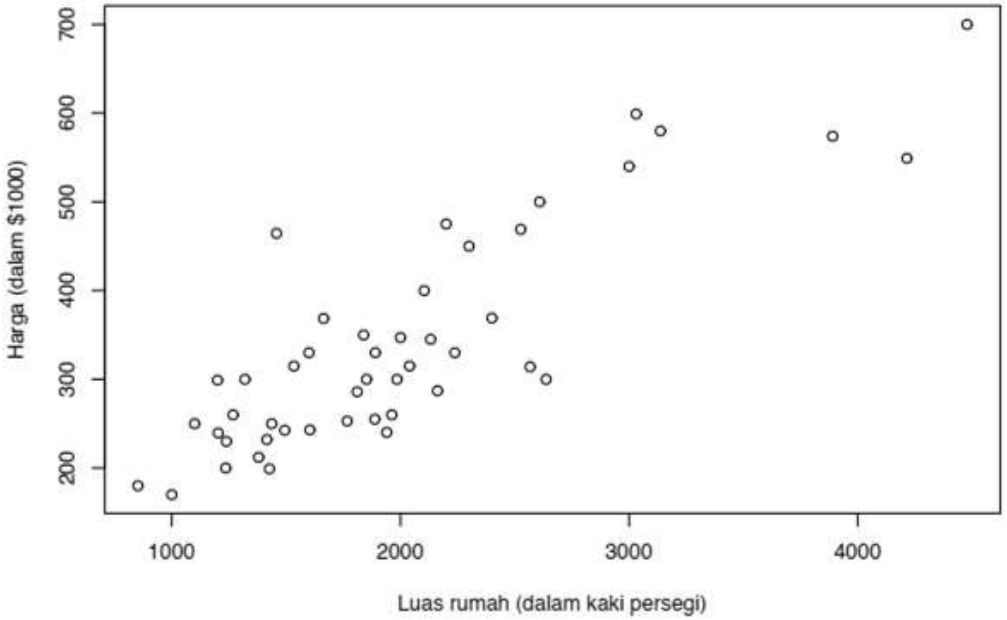


Linear Regression

Luas rumah (x)	Harga (y)
2104	400
1600	330
2400	369
1416	232
3000	540
....	….
….	….

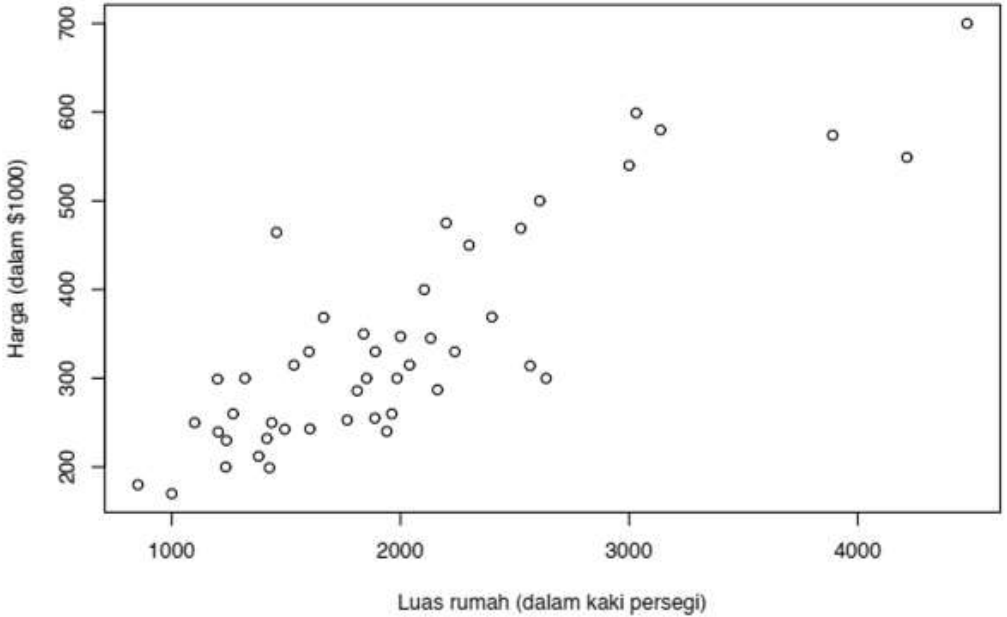


Linear Regression



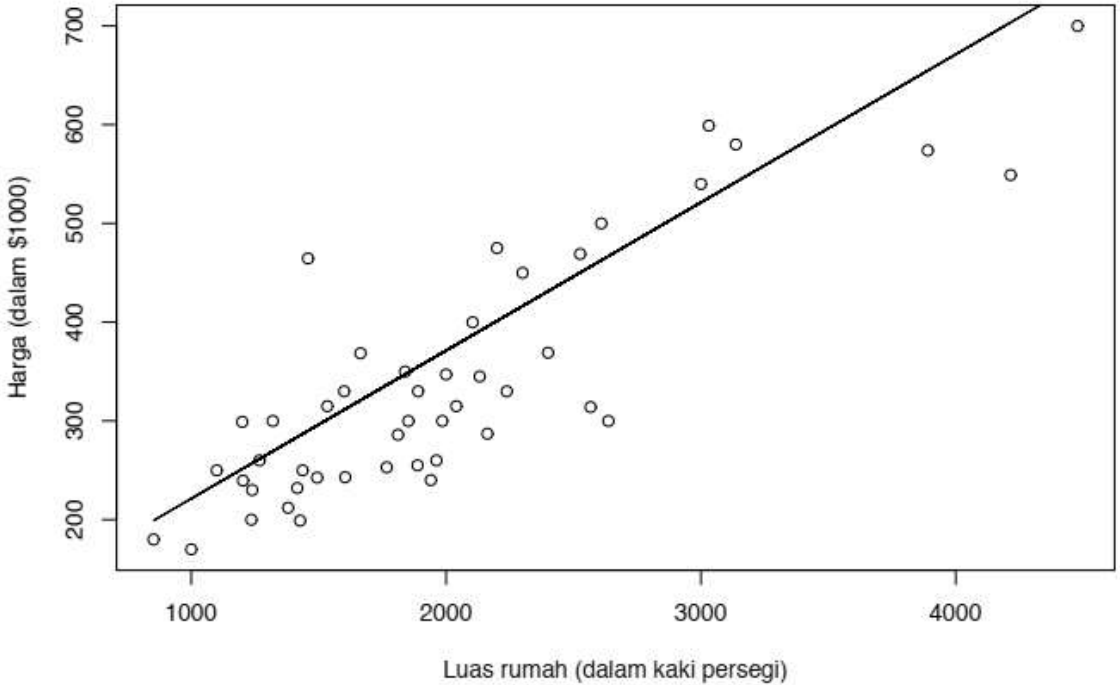
Suppose that we want to predict house price (Harga) based on the training set plotted above.

Linear Regression



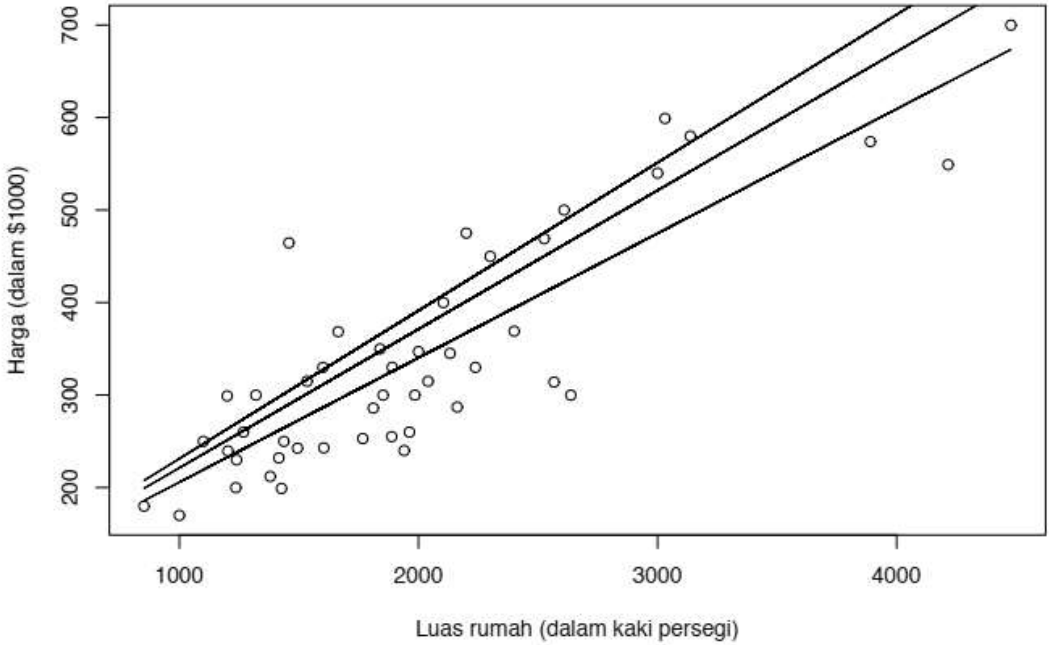
If we are about to *draw* a good hypothesis model *h*, what would you draw to *best* model the data?

Linear Regression



Yes, a straight line seems to fit the data well

Linear Regression

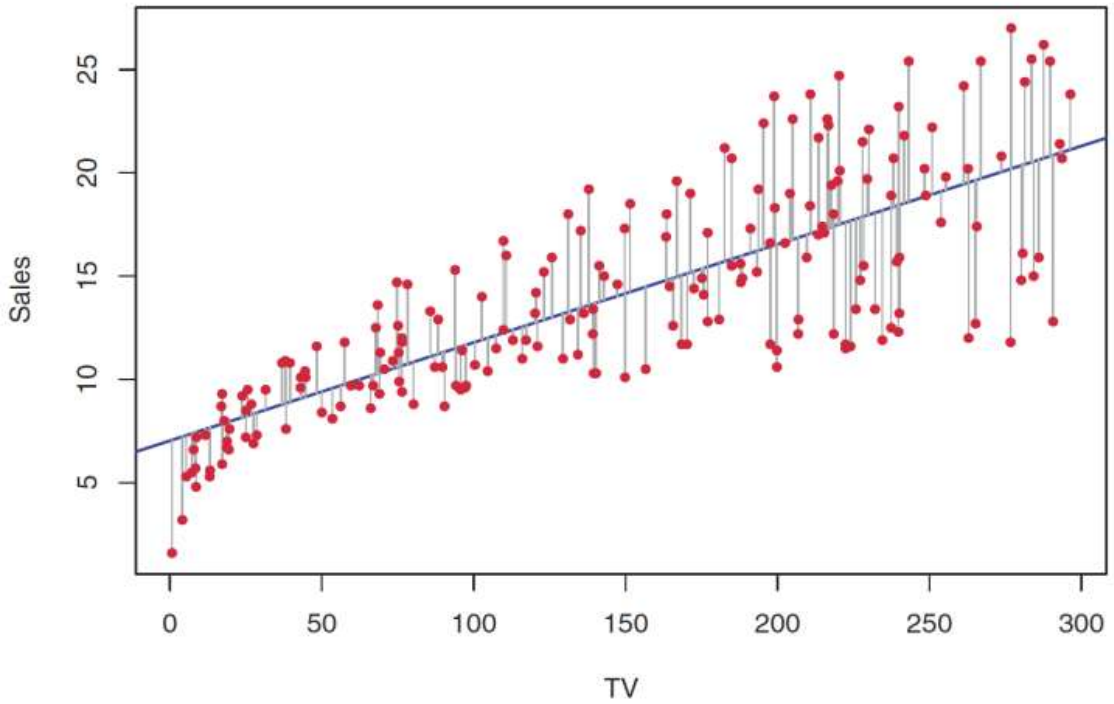


Yes, a straight line seems to fit the data well. But which line? There are infinitely many possible lines (hypotheses, recall the earlier slides)

Linear Regression

Which line ?

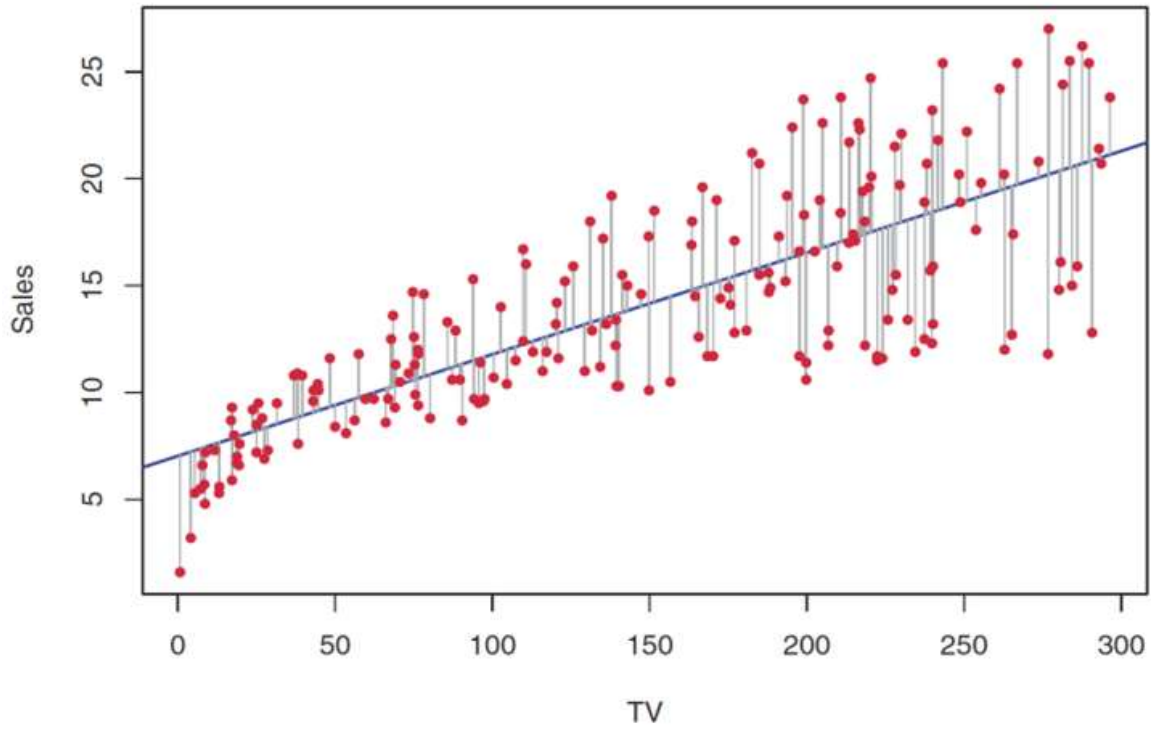
Linear Regression



We can think a distance between the line and a data point as an **error**.

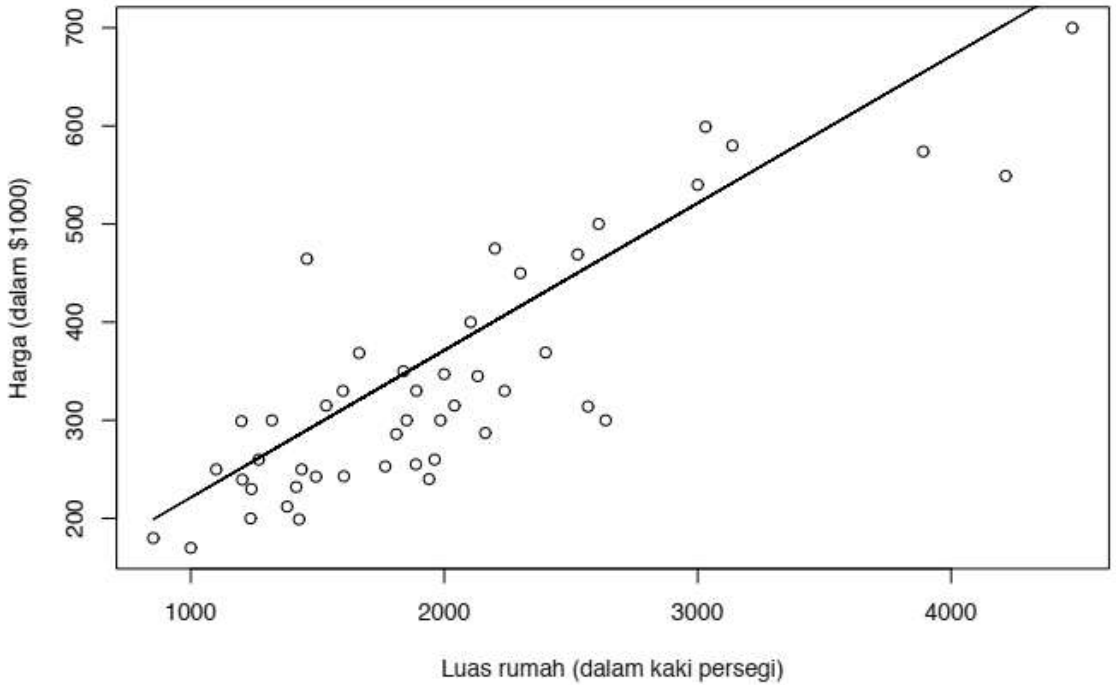
$$\text{Error} = \sum_{i=1}^{n} (actual\_output - predicted\_output) ** 2$$

Linear Regression

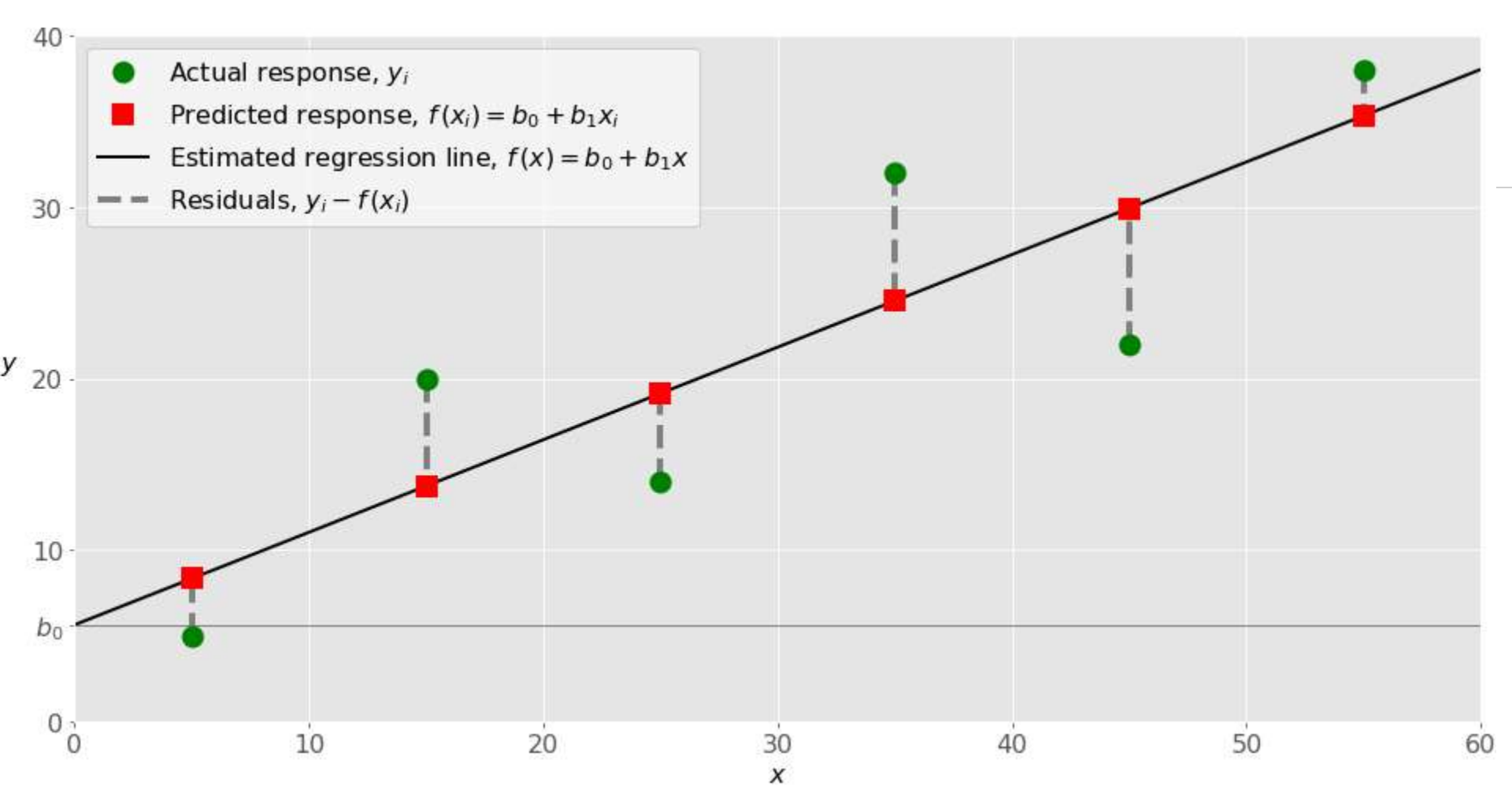


Thus, the *best* line to model our hypothesis is the one that has the smallest accumulated error.

Linear Regression



Mathematically, the above straight line can be written by Harga = *b0* + b1Luas


Actual response, y_i
Predicted response, $f(x_i) = b_0 + b_1 x_i$
Estimated regression line, $f(x) = b_0 + b_1 x$
Residuals, $y_i - f(x_i)$
40
30
20
10
b_0
0
y
0
10
20
30
40
50
60
x

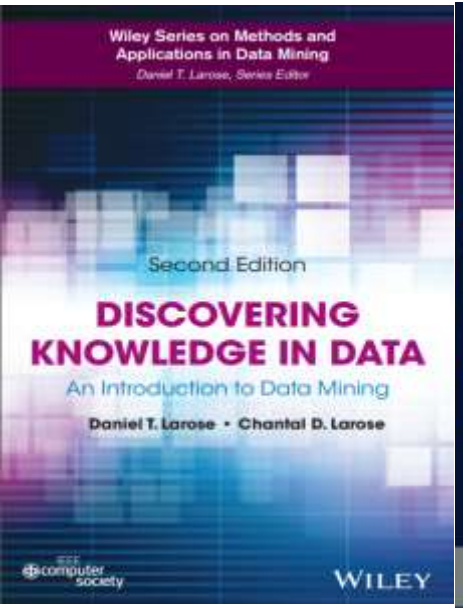
Linear Regression

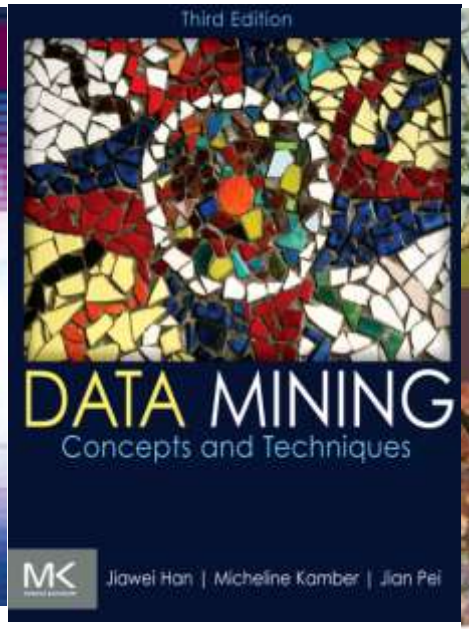
Given that,

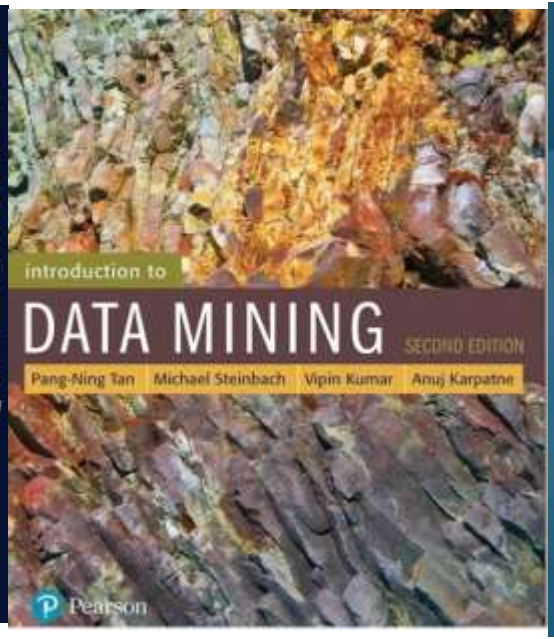
Harga = *b0* + *b1* Luas;

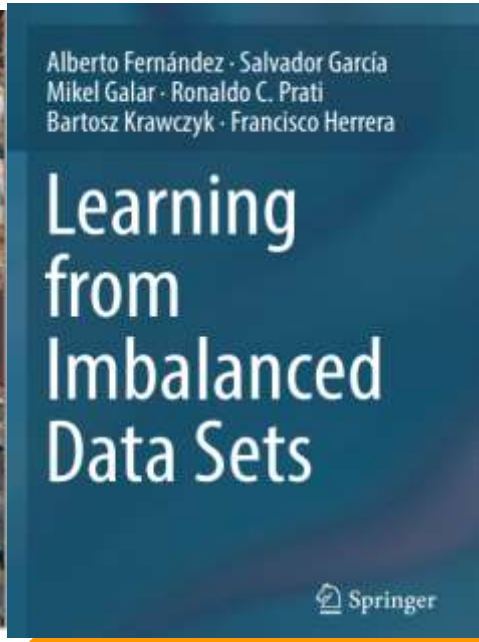
linear regression is a procedure to find the best line (hypothesis or model) by *searching* parameters ***m*** and ***b*** that give the *smallest* total error.

REFERENSI


Wiley Series on Methods and Applications in Data Mining
Daniel T. Larose, Series Editor
Second Edition
DISCOVERING KNOWLEDGE IN DATA
An Introduction to Data Mining
Daniel T. Larose • Chantal D. Larose
computer society
WILEY


Third Edition
DATA MINING
Concepts and Techniques
MK
Jiawei Han | Micheline Kamber | Jian Pei


introduction to
DATA MINING
SECOND EDITION
Pang-Ning Tan
Michael Steinbach
Vipin Kumar
Anuj Karpatne
Pearson


Alberto Fernández · Salvador García
Mikel Galar · Ronaldo C. Prati
Bartosz Krawczyk · Francisco Herrera
Learning from Imbalanced Data Sets
Springer

THANKS!

SERTIFIKAT

No : PLT.01.0037/ISB/LOK/IV/2022

Diberikan Kepada :

Hairani, S.Kom., M.Eng.

Sebagai Narasumber

KULIAH UMUM

"IoT dan Data Mining"

Diselenggarakan Oleh :

PROGRAM STUDI S1 TEKNOLOGI INFORMASI INSTITUT SHANTI BHUANA

Pada Tanggal : 21 April 2022

KEPALA PRODI TEKNOLOGI INFORMASI

Azriel Christian Nurcahyo, S.Kom.,M.Kom

Kampus Merdeka INDONESIA JAYA

PEMBINA HMTI

Santi Thomas, S.Kom.,M.M.S.I