DATA MINING

PERTEMUAN KE-2

I.2

SEJARAH DAN PENERAPAN

Versi : 01

Sejarah Data Mining

- Sebelum 1600: Empirical science
  - Disebut sains kalau bentuknya kasat mata
- 1600-1950: Theoretical science
  - Disebut sains kalau bisa dibuktikan secara matematis atau eksperimen
- 1950s-1990: Computational science
  - Seluruh disiplin ilmu bergerak ke komputasi
  - Lahirnya banyak model komputasi
- 1990-sekarang: Data science
  - Kultur manusia menghasilkan data besar
  - Kemampuan komputer untuk mengolah data besar
  - Datangnya data mining sebagai arus utama sains

Jim Gray and Alex Szalay, The World Wide Telescope: An Archetype for Online Science, Comm. ACM, 45(11): 50-54, Nov. 2002

Forbes / Tech

JAN 20, 2016 @ 02:39 PM 15,446 VIEWS The Little Black

Report: Why "Data Scientist" Is The Best Job To Pu 2016

Gregory Ferenstein, CONT
FULL BIO
Opinions expressed by Forbes Contributors ar

(Ferenstein Wire) - Data scientis
jobs in America, according to a r
company review site, Glassdoor.
voluntary reviews and self-repor
company's massive dataset; eacl
a composite score of median rep
openings, and career opportunit

According to the report, the med
Scientist is an impressive $116,0

CNBC

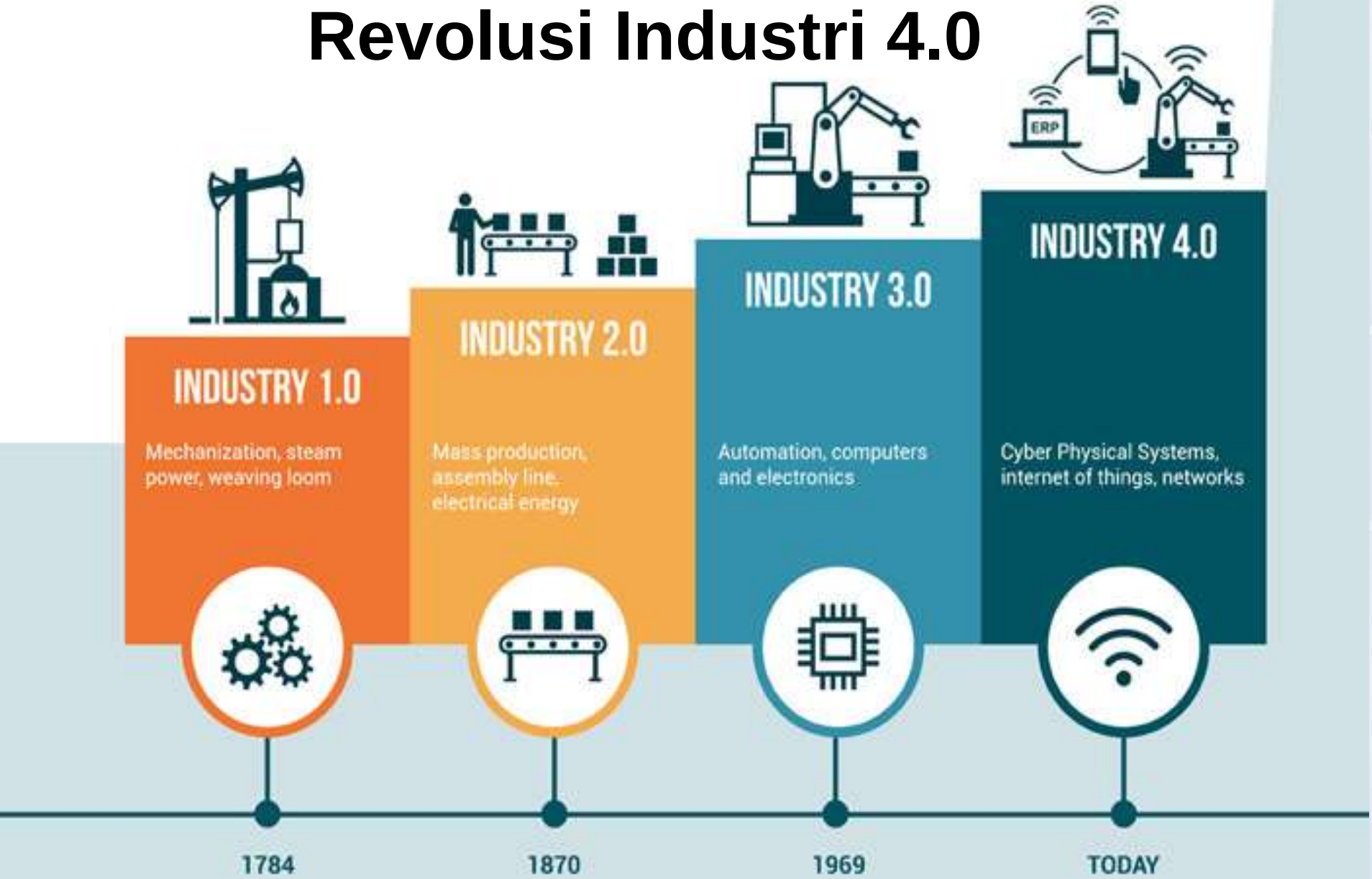
JOBS

ECONOMY | WORLD ECONOMY | US ECONOMY | THE FED | CENTRAL BANKS | JOBS |

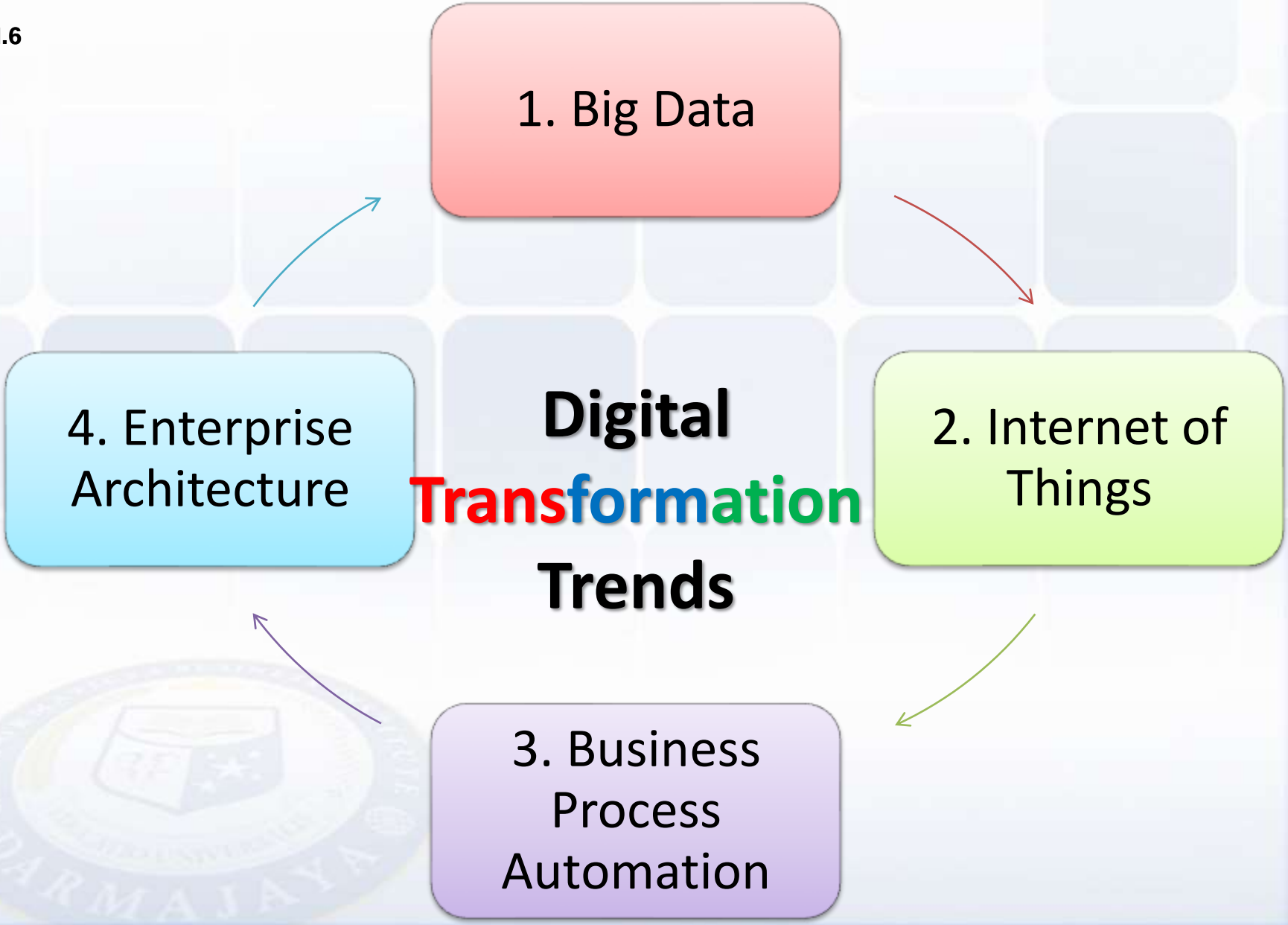
Data science jobs top Glassdoor survey for best work-life balance

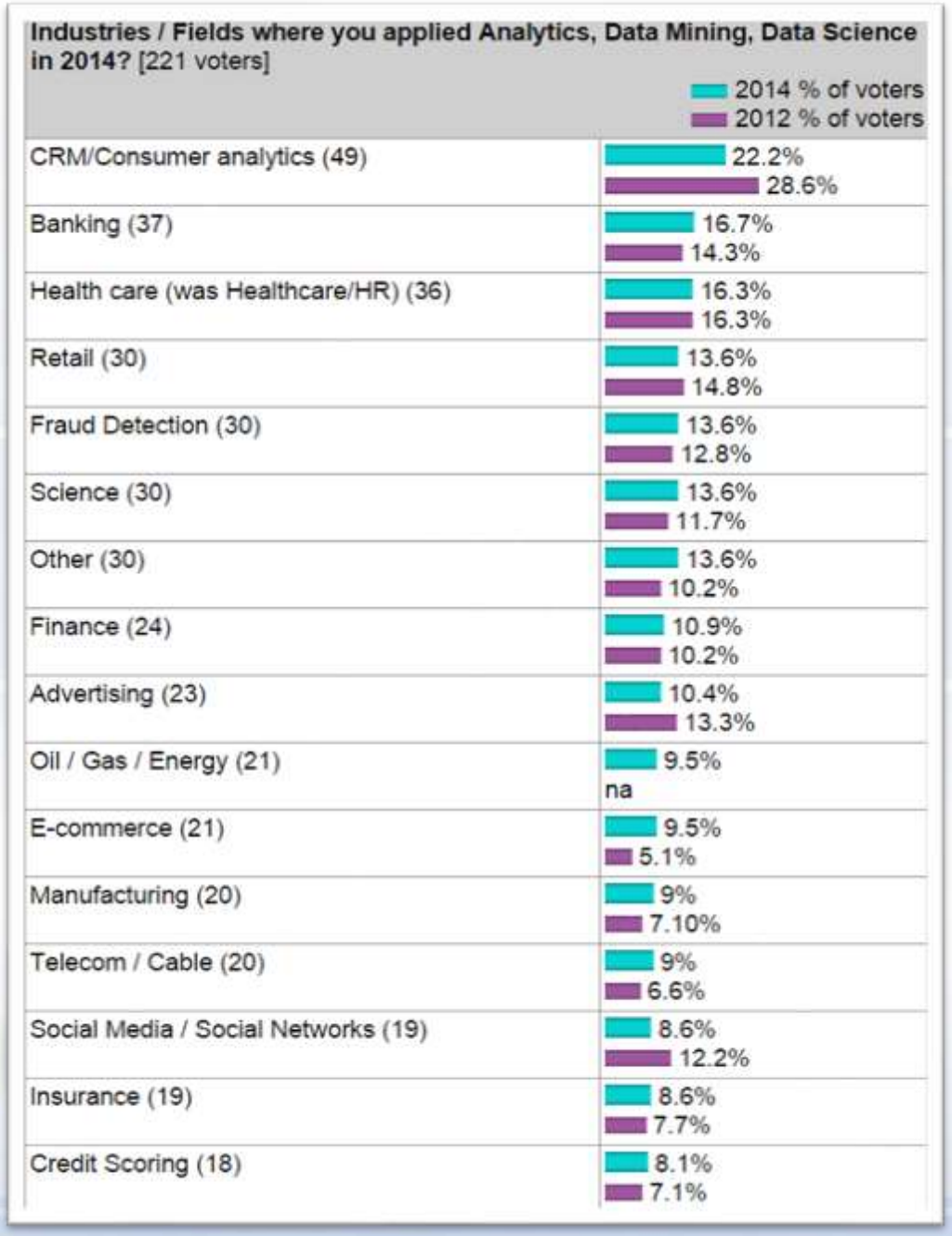
Versi : 01

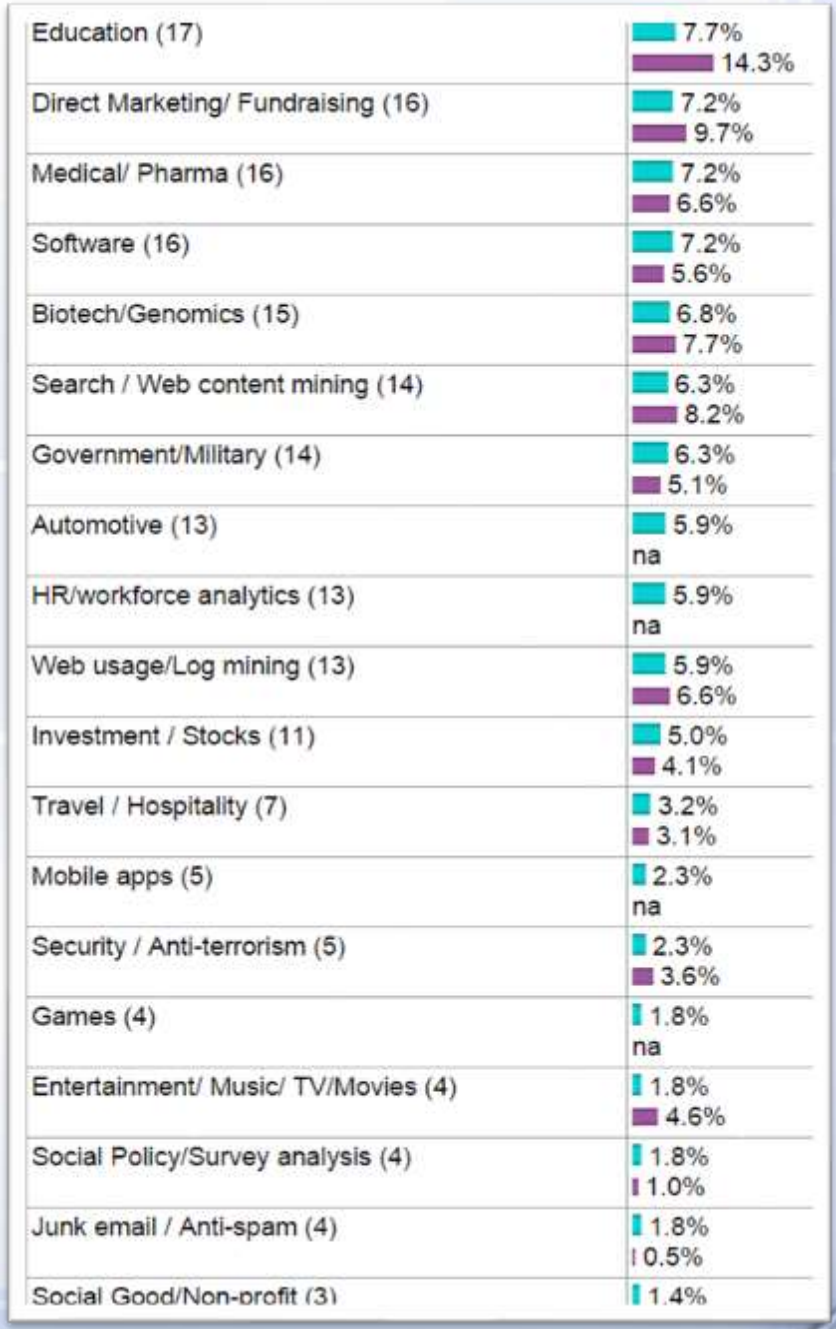
Revolusi Industri 4.0



Versi : 01


1. Big Data
2. Internet of Things
3. Business Process Automation
4. Enterprise Architecture
Digital Transformation Trends

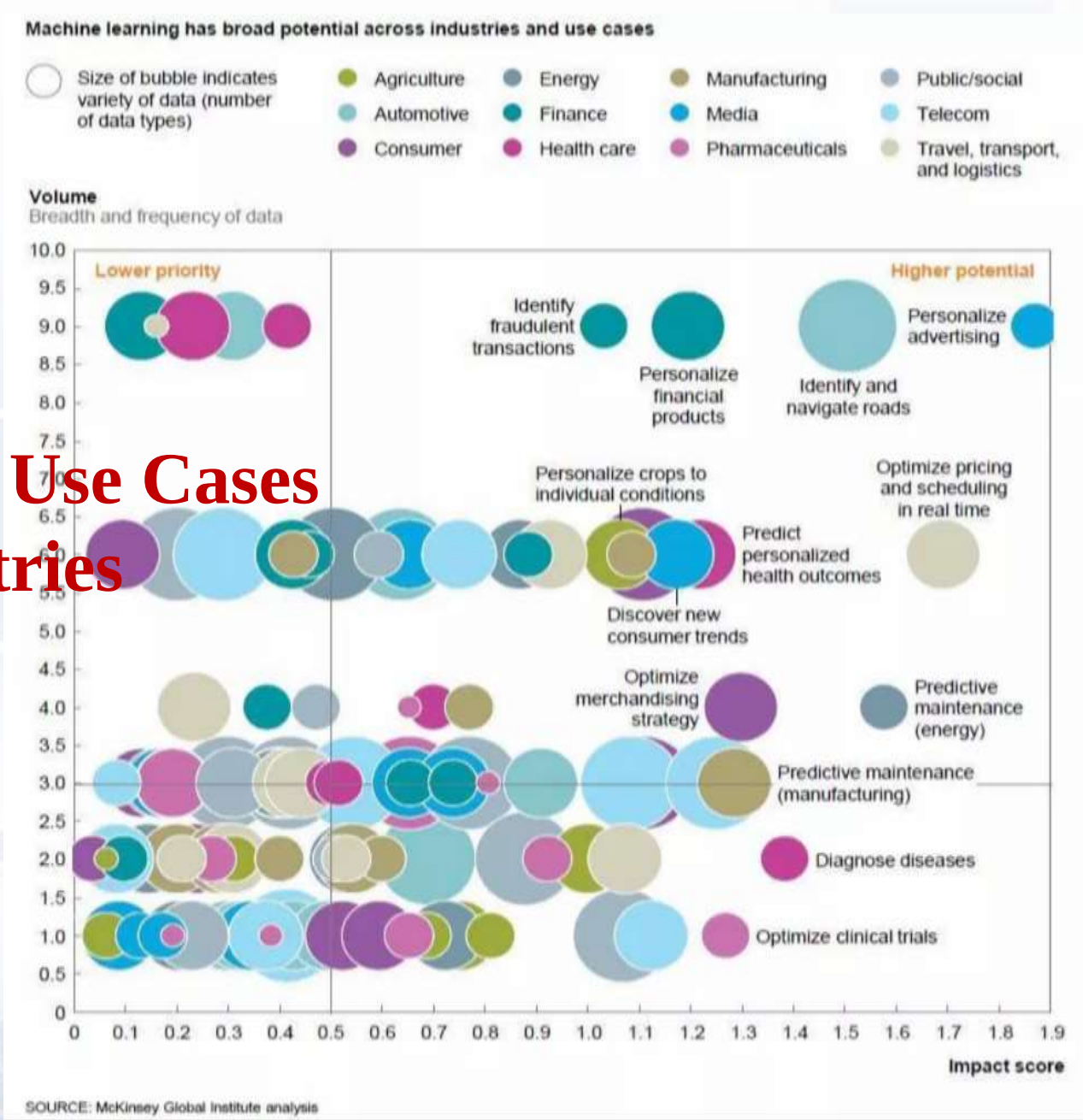

Industries / Fields where you applied Analytics, Data Mining, Data Science in 2014? [221 voters]
2014 % of voters
2012 % of voters
CRM/Consumer analytics (49) 22.2% 28.6%
Banking (37) 16.7% 14.3%
Health care (was Healthcare/HR) (36) 16.3% 16.3%
Retail (30) 13.6% 14.8%
Fraud Detection (30) 13.6% 12.8%
Science (30) 13.6% 11.7%
Other (30) 13.6% 10.2%
Finance (24) 10.9% 10.2%
Advertising (23) 10.4% 13.3%
Oil / Gas / Energy (21) 9.5% na
E-commerce (21) 9.5% 5.1%
Manufacturing (20) 9% 7.10%
Telecom / Cable (20) 9% 6.6%
Social Media / Social Networks (19) 8.6% 12.2%
Insurance (19) 8.6% 7.7%
Credit Scoring (18) 8.1% 7.1%


Education (17) 7.7% 14.3%
Direct Marketing/ Fundraising (16) 7.2% 9.7%
Medical/ Pharma (16) 7.2% 6.6%
Software (16) 7.2% 5.6%
Biotech/Genomics (15) 6.8% 7.7%
Search / Web content mining (14) 6.3% 8.2%
Government/Military (14) 6.3% 5.1%
Automotive (13) 5.9% na
HR/workforce analytics (13) 5.9% na
Web usage/Log mining (13) 5.9% 6.6%
Investment / Stocks (11) 5.0% 4.1%
Travel / Hospitality (7) 3.2% 3.1%
Mobile apps (5) 2.3% na
Security / Anti-terrorism (5) 2.3% 3.6%
Games (4) 1.8% na
Entertainment/ Music/ TV/Movies (4) 1.8% 4.6%
Social Policy/Survey analysis (4) 1.8% 1.0%
Junk email / Anti-spam (4) 1.8% 0.5%
Social Good/Non-profit (3) 1.4%

Versi : 01

I.8

Data Mining Use Cases across Industries



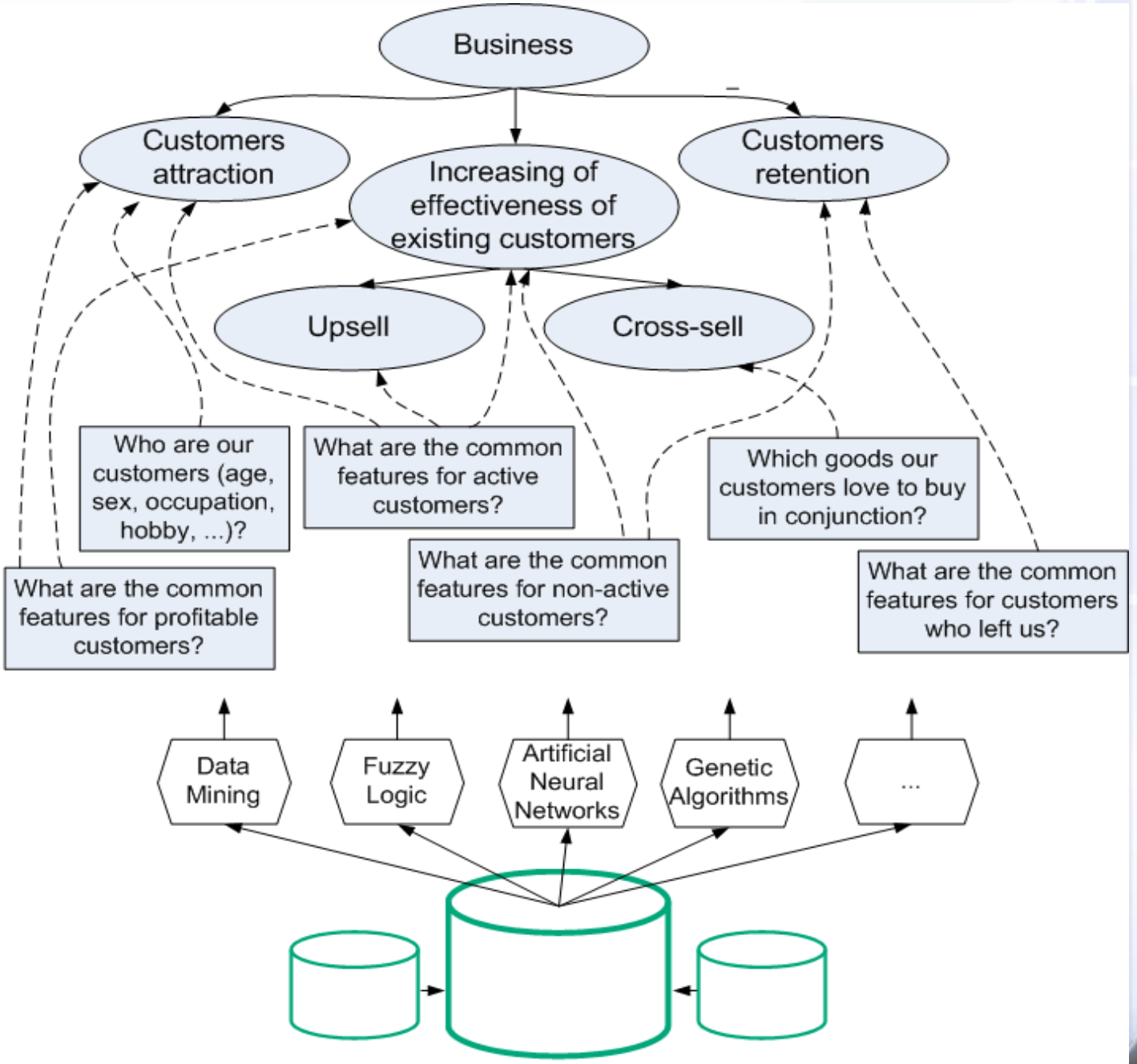
Versi : 01

Business

Knowledge

Methods

Technology



Versi : 01

Business Goals Law (Data Mining Law 1)

Tujuan bisnis adalah asal mula setiap solusi dari Data Mining

- Definisikan bidang Data Mining: Data Mining berkaitan dengan pemecahan masalah bisnis dan pencapaian tujuan bisnis
- Data Mining pada dasarnya bukanlah sebuah teknologi; yaitu proses, yang memiliki satu atau lebih tujuan bisnis sebagai intinya
- Without a business objective, there is no data mining
- The maxim: "Data Mining is a Business Process"

CRISP-DM

- CRISP-DM adalah singkatan dari Cross-Industry Standard Process for Data Mining. Ini adalah metodologi yang umum digunakan dalam dunia data mining dan analisis data untuk mengelola proyek-proyek analisis data secara sistematis. Metodologi ini membantu tim analisis data dalam mengidentifikasi, merencanakan, dan mengimplementasikan proyek analisis data dengan langkah-langkah yang terstruktur. CRISP-DM terdiri dari enam tahap utama, yaitu

Business Knowledge Law (Data Mining Law 2)

Pengetahuan bisnis sangat penting dalam setiap langkah proses penambangan data

Pengetahuan bisnis sangat penting dalam setiap langkah proses penambangan data

- A naive reading of CRISP-DM would see business knowledge used at the start of the process in defining goals, and at the end of the process in guiding deployment of results
- Hal ini berarti kehilangan properti utama dari proses data mining, bahwa pengetahuan bisnis memiliki peran sentral dalam setiap langkah

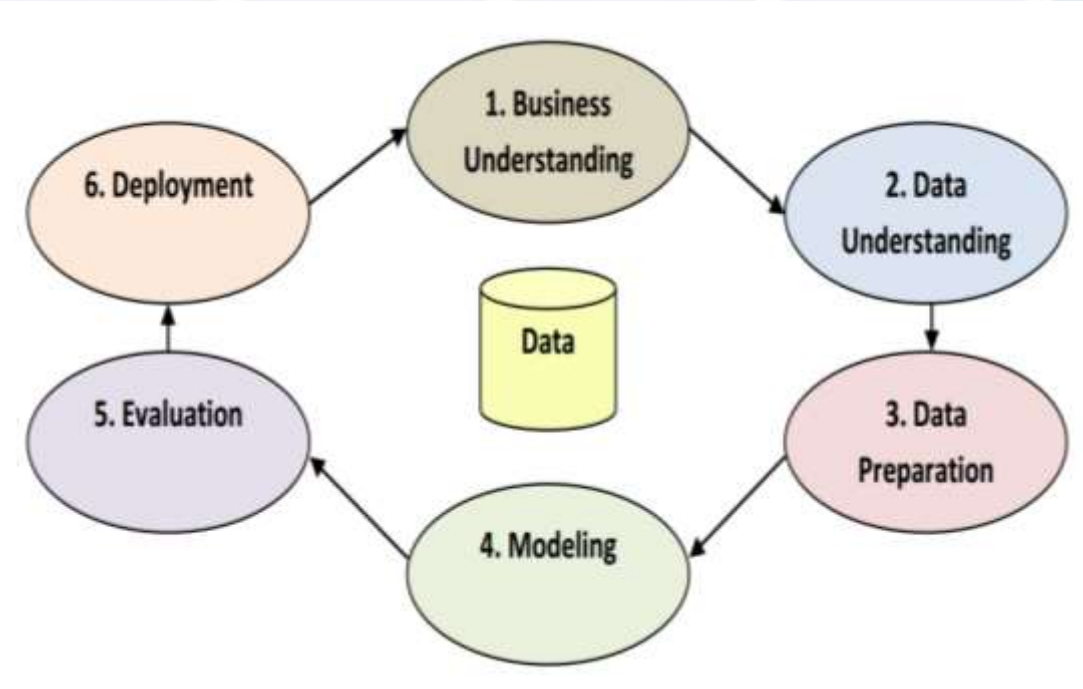
CRISP-DM terdiri dari enam tahap utama, yaitu

1. Business Understanding (Pemahaman Bisnis): Tahap ini melibatkan pemahaman sepenuhnya tentang masalah bisnis yang ingin dipecahkan melalui analisis data. Ini melibatkan identifikasi tujuan bisnis, kebutuhan pengguna, dan pemahaman tentang konteks bisnis.
2. Data Understanding (Pemahaman Data): Di tahap ini, tim mengumpulkan data yang diperlukan dan memahami karakteristiknya. Ini mencakup pengeksplorasian data, pemahaman atas masalah kualitas data, serta pemahaman terhadap hubungan antar data.

3. Data Preparation (Persiapan Data): Data yang telah dikumpulkan kemudian dipersiapkan untuk analisis. Ini termasuk pembersihan data, penggabungan data, transformasi, dan pemilihan atribut yang relevan.
4. Modeling (Modeling): Di tahap ini, model-model analisis data dikembangkan dan dievaluasi. Ini mencakup pemilihan algoritma, pelatihan model, dan pengujian model untuk memahami sejauh mana model tersebut berhasil memecahkan masalah.

5. Evaluation (Evaluasi): Evaluasi model-model yang dikembangkan dilakukan untuk memahami sejauh mana mereka efektif dalam mencapai tujuan bisnis yang ditetapkan pada tahap awal. Evaluasi ini dapat mengarah pada perbaikan model atau perubahan dalam pendekatan analisis data.
6. Deployment (Implementasi): Model-model yang telah diuji dan dievaluasi kemudian diimplementasikan dalam lingkungan produksi. Ini melibatkan integrasi model dalam sistem bisnis dan memastikan bahwa mereka dapat digunakan dalam situasi nyata.

CRISP-DM

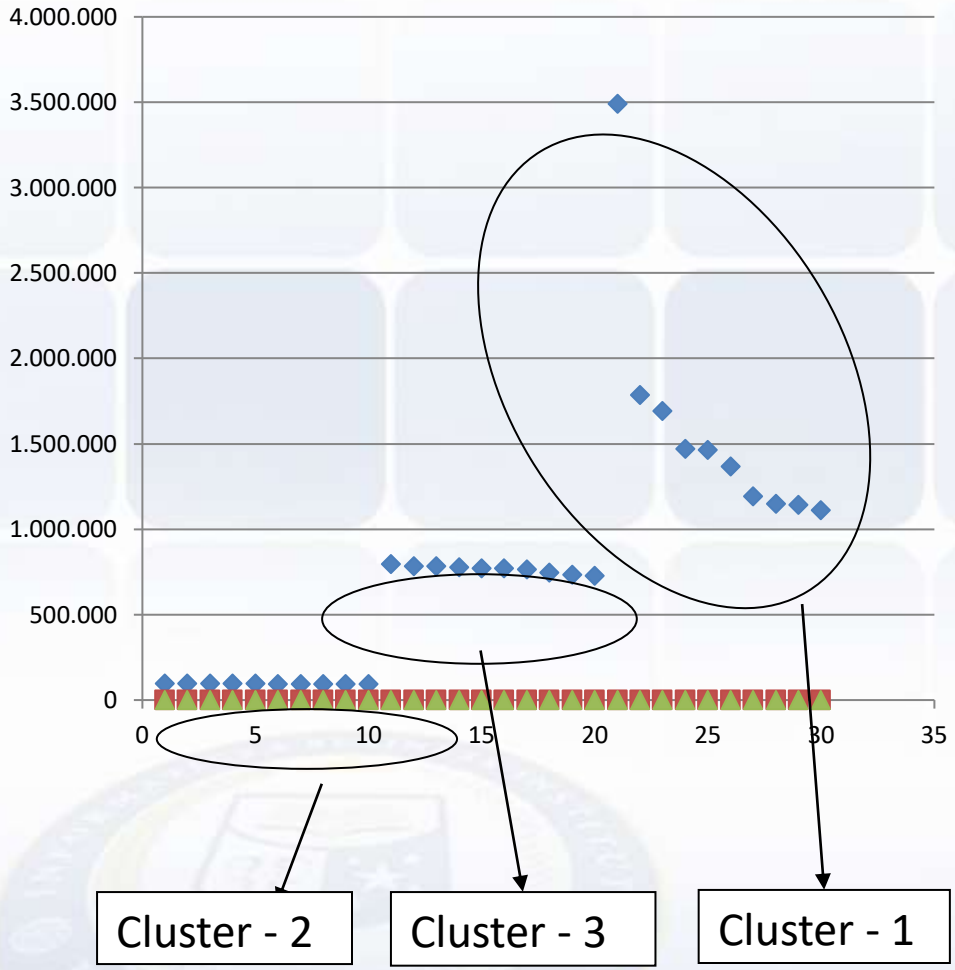

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment
Data

Private and Commercial Sector

- Marketing: product recommendation, market basket analysis, product targeting, customer retention
- Finance: investment support, portfolio management, price forecasting
- Banking and Insurance: credit and policy approval, money laundry detection
- Security: fraud detection, access control, intrusion detection, virus detection
- Manufacturing: process modeling, quality control, resource allocation
- Web and Internet: smart search engines, web marketing
- Software Engineering: effort estimation, fault prediction
- Telecommunication: network monitoring, customer churn prediction, user behavior analysis

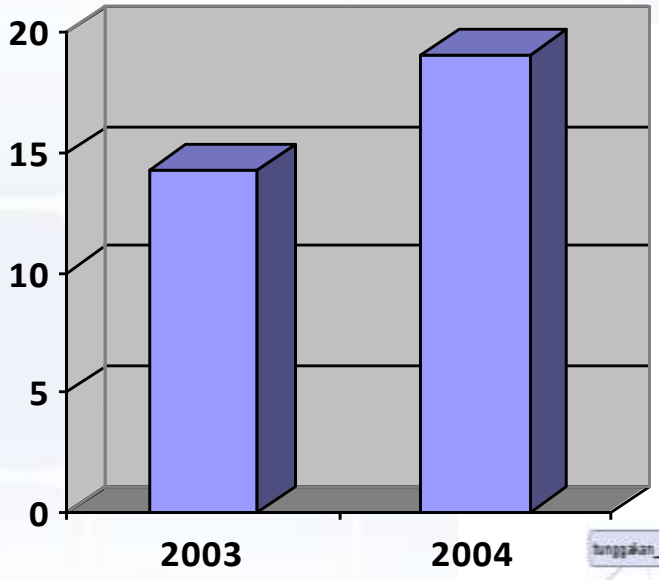
I.18

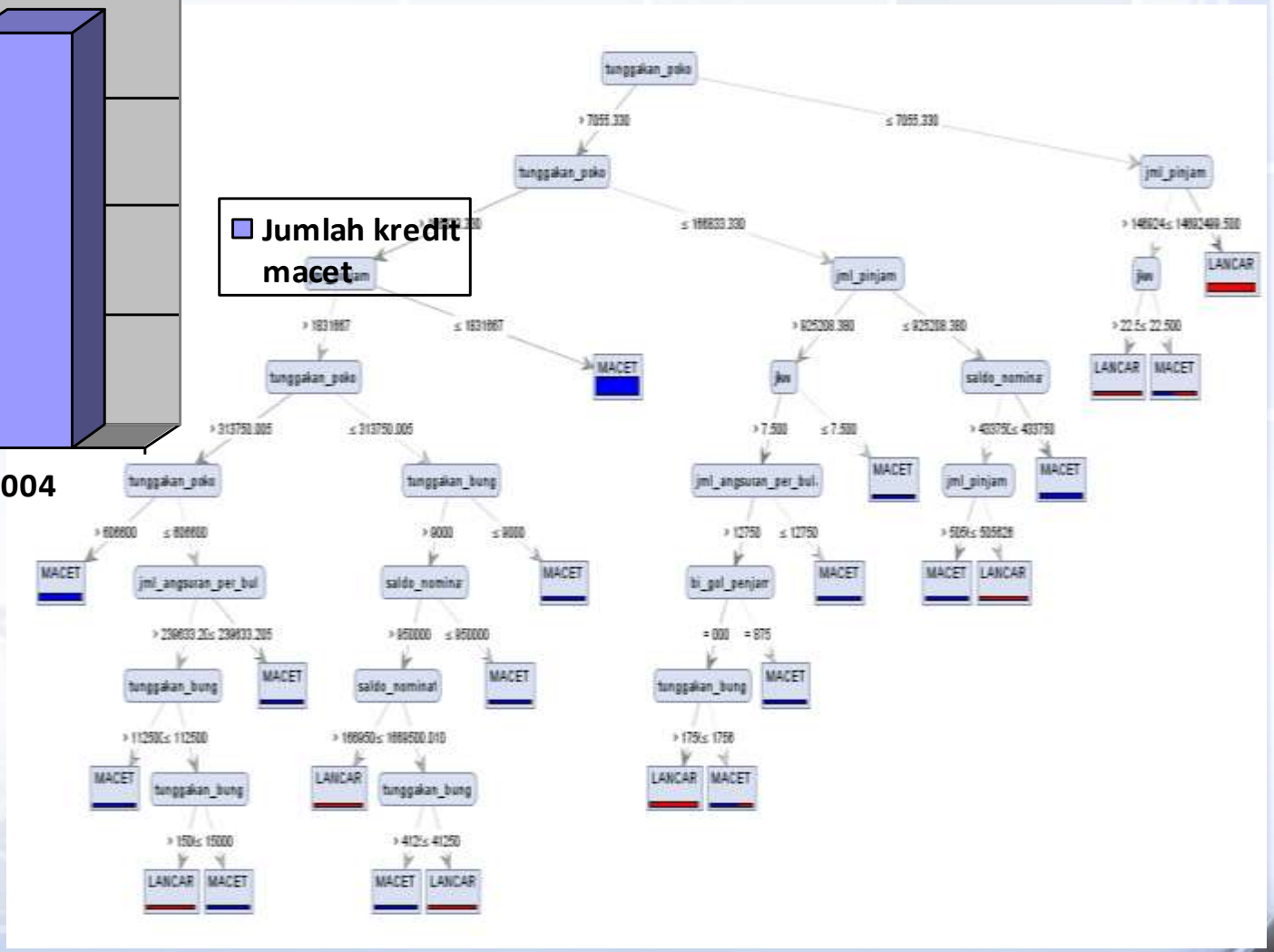
Use Case: Product Recommendation



Versi : 01

I.19 Use Case: Penentuan Kelayakan Kredit





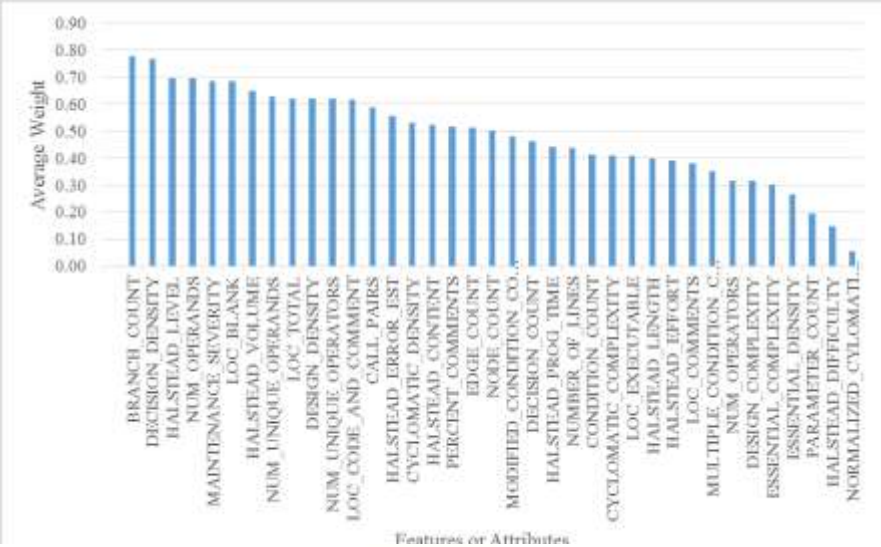
Versi : 01

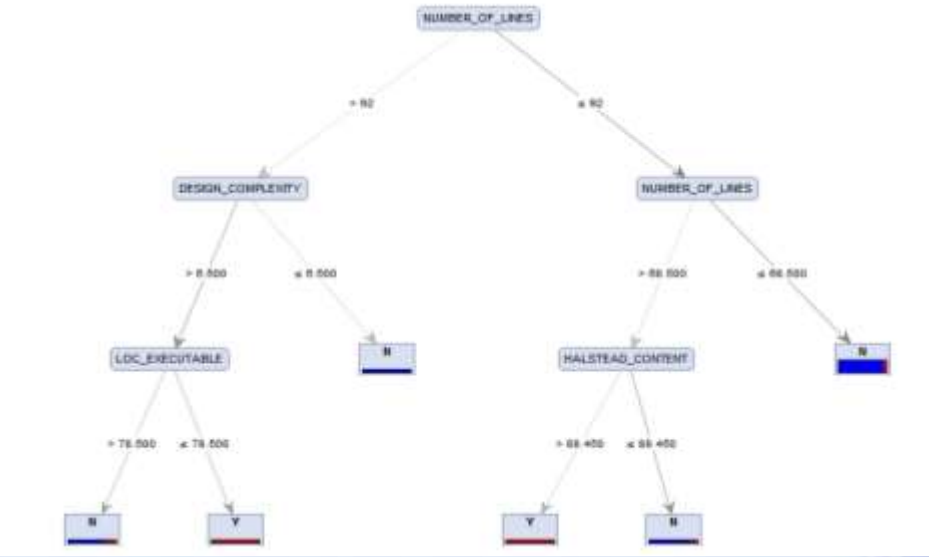
I.20 Use Case: Software Fault Prediction

Biaya untuk menemukan dan memperbaiki kerusakan itu mahal

- $14,102 per defect in post-release phase *(Boehm & Basili 2008)*
- $60 billion per year (NIST 2002)

- Industrial methods of manual software reviews activities can find only 60% of defects (Shull et al. 2002)
- The probability of detection of software fault prediction models is higher (71%) than software reviews (60%)





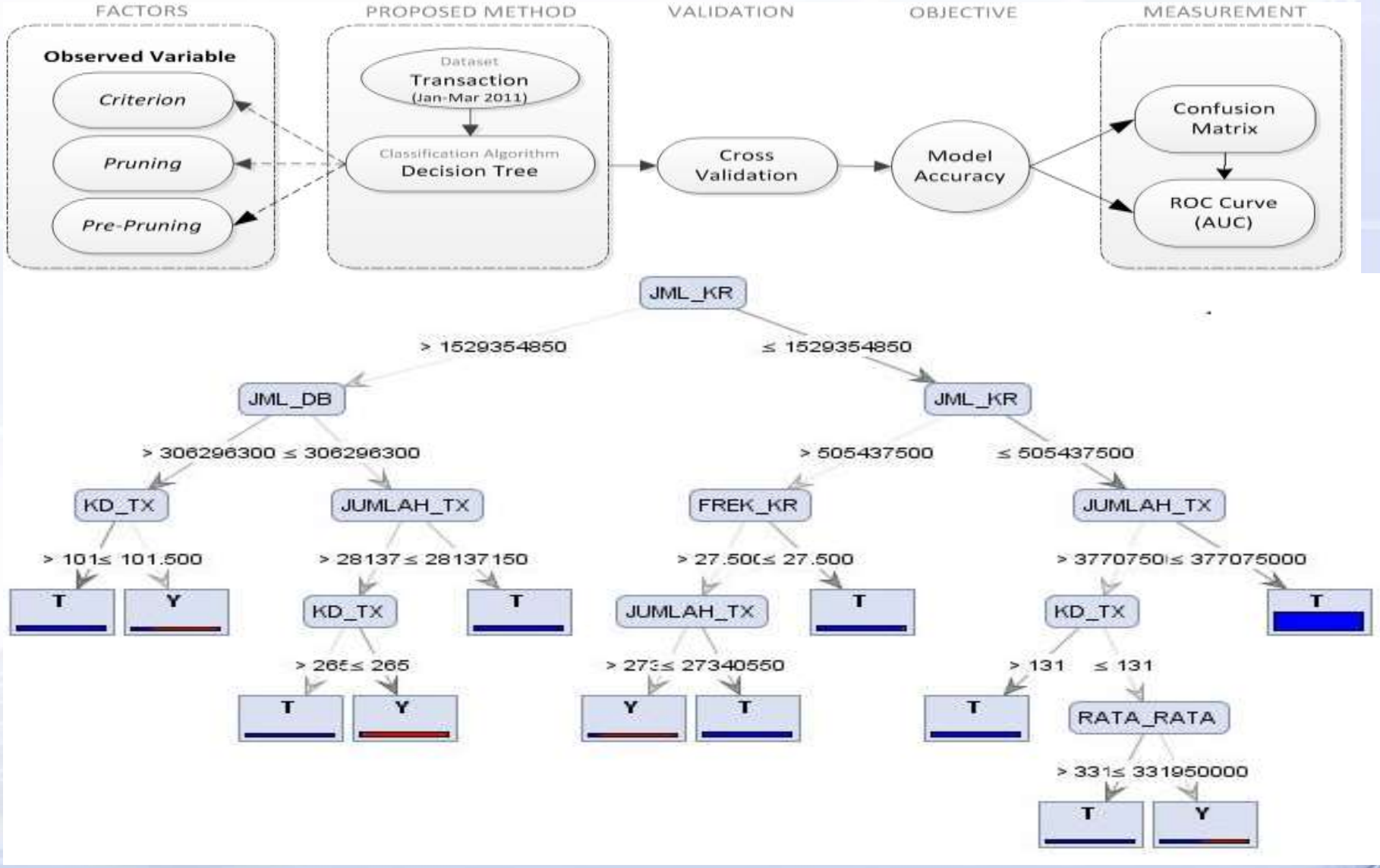
Versi : 01

Public and Government Sector

- Finance: exchange rate forecasting, sentiment analysis
- Taxation: adaptive monitoring, fraud detection
- Medicine and Healt Care: hypothesis discovery, disease prediction and classification, medical diagnosis
- Education: student allocation, resource forecasting
- Insurance: worker's compensation analysis
- Security: bomb, iceberg detection
- Transportation: simulation and analysis, load estimation
- Law: legal patent analysis, law and rule analysis
- Politic: election prediction

I.22

Use Case: Deteksi Pencucian Uang



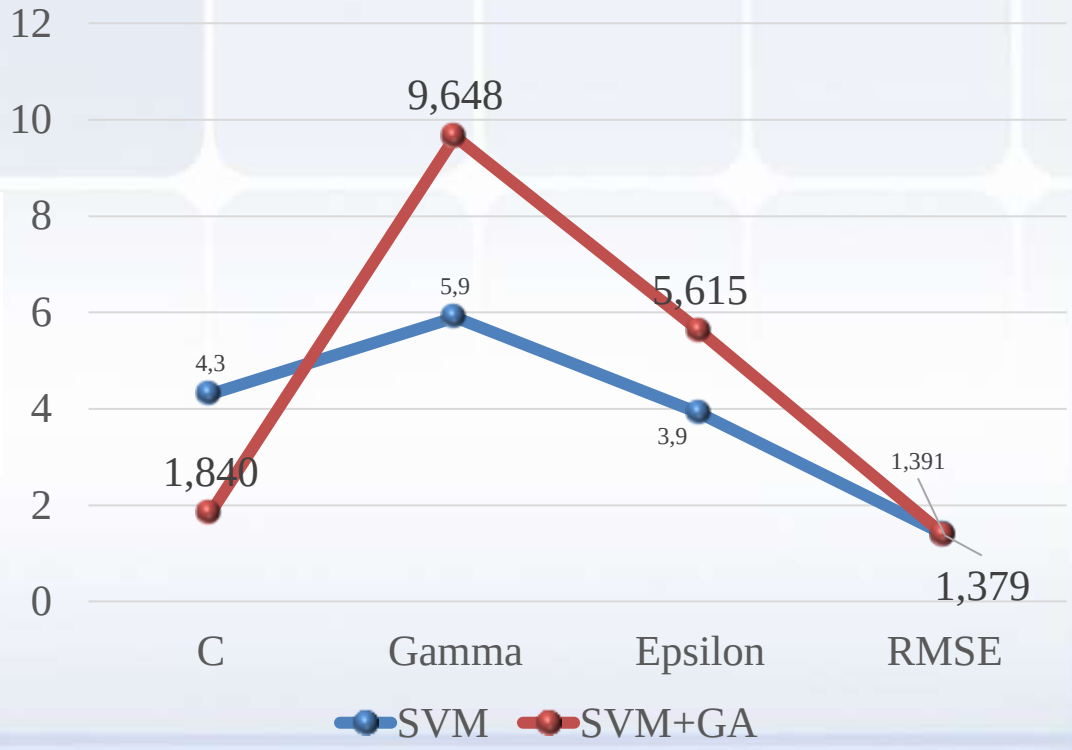
Versi : 01

I.23

Use Case: Prediksi Kebakaran Hutan

FFMC	DMC	DC	ISI	temp	RH	wind	rain	ln(area+1)
93.5	139.4	594.2	20.3	17.6	52	5.8	0	0
92.4	124.1	680.7	8.5	17.2	58	1.3	0	0
90.9	126.5	686.5	7	15.6	66	3.1	0	0
85.8	48.3	313.4	3.9	18	42	2.7	0	0.307485
91	129.5	692.6	7	21.7	38	2.2	0	0.357674
90.9	126.5	686.5	7	21.9	39	1.8	0	0.385262
95.5	99.9	513.3	13.2	23.3	31	4.5	0	0.438255

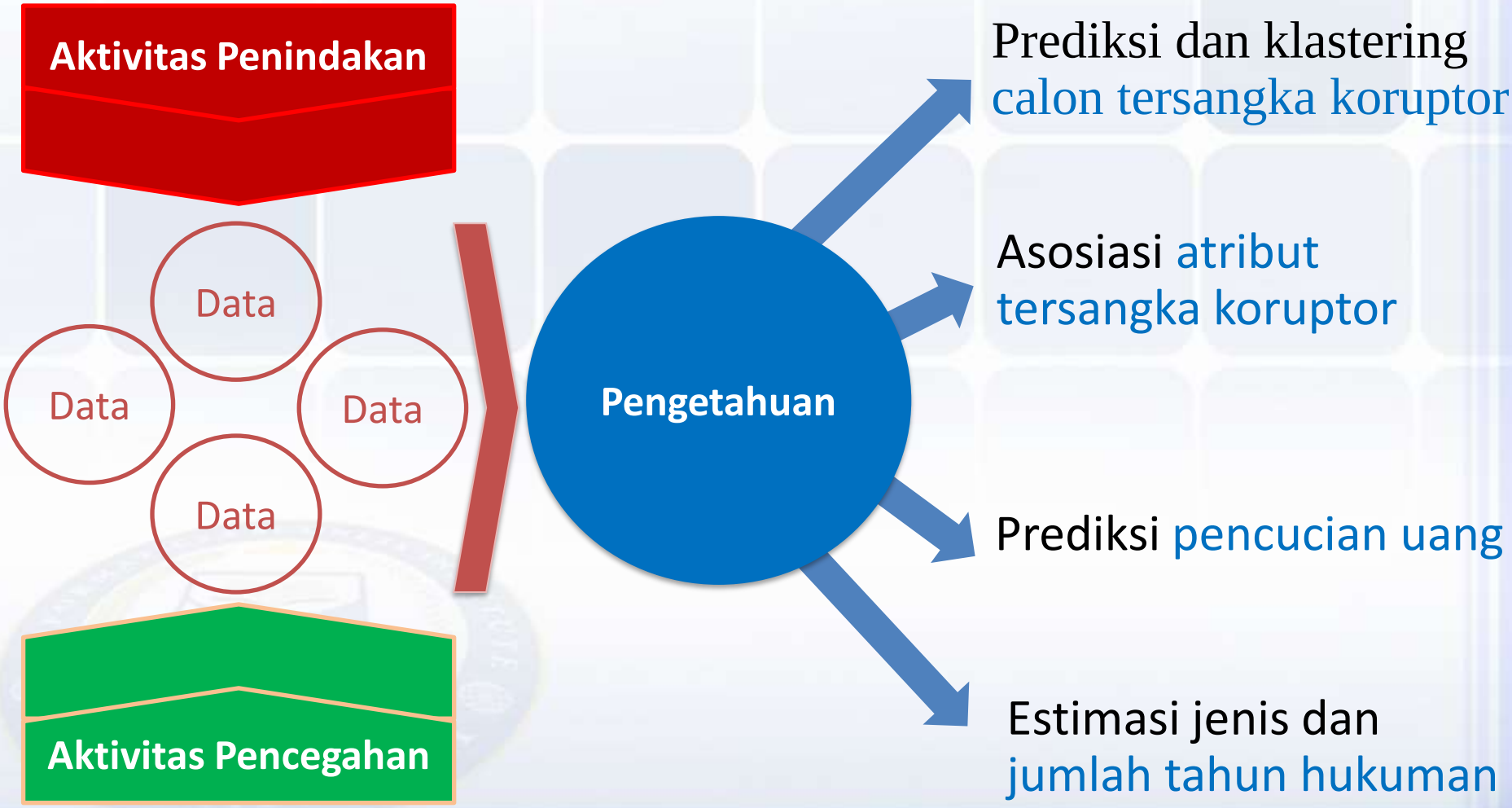
	SVM	SVM+GA
C	4.3	1,840
Gamma (γ)	5.9	9,648
Epsilon (ε)	3.9	5,615
RMSE	1.391	1.379



Versi : 01

I.24

Use Case: Profiling dan Prediksi Koruptor



Versi : 01

Contoh Penerapan Data Mining

- Penentuan kelayakan kredit pemilihan rumah di bank
- Penentuan pasokan listrik PLN untuk wilayah Jakarta
- Prediksi profile tersangka koruptor dari data pengadilan
- Perkiraan harga saham dan tingkat inflasi
- Analisis pola belanja pelanggan
- Memisahkan minyak mentah dan gas alam
- Penentuan pola pelanggan yang loyal pada perusahaan operator telepon
- Deteksi pencucian uang dari transaksi perbankan
- Deteksi serangan (*intrusion*) pada suatu jaringan

Data Mining Society

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
  - Knowledge Discovery in Databases *(G. Piatetsky-Shapiro and W. Frawley, 1991)*
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining *(U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)*
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), WSDM (2008), etc.
- ACM Transactions on KDD (2007)

Review dan Latihan

☺ END ☺