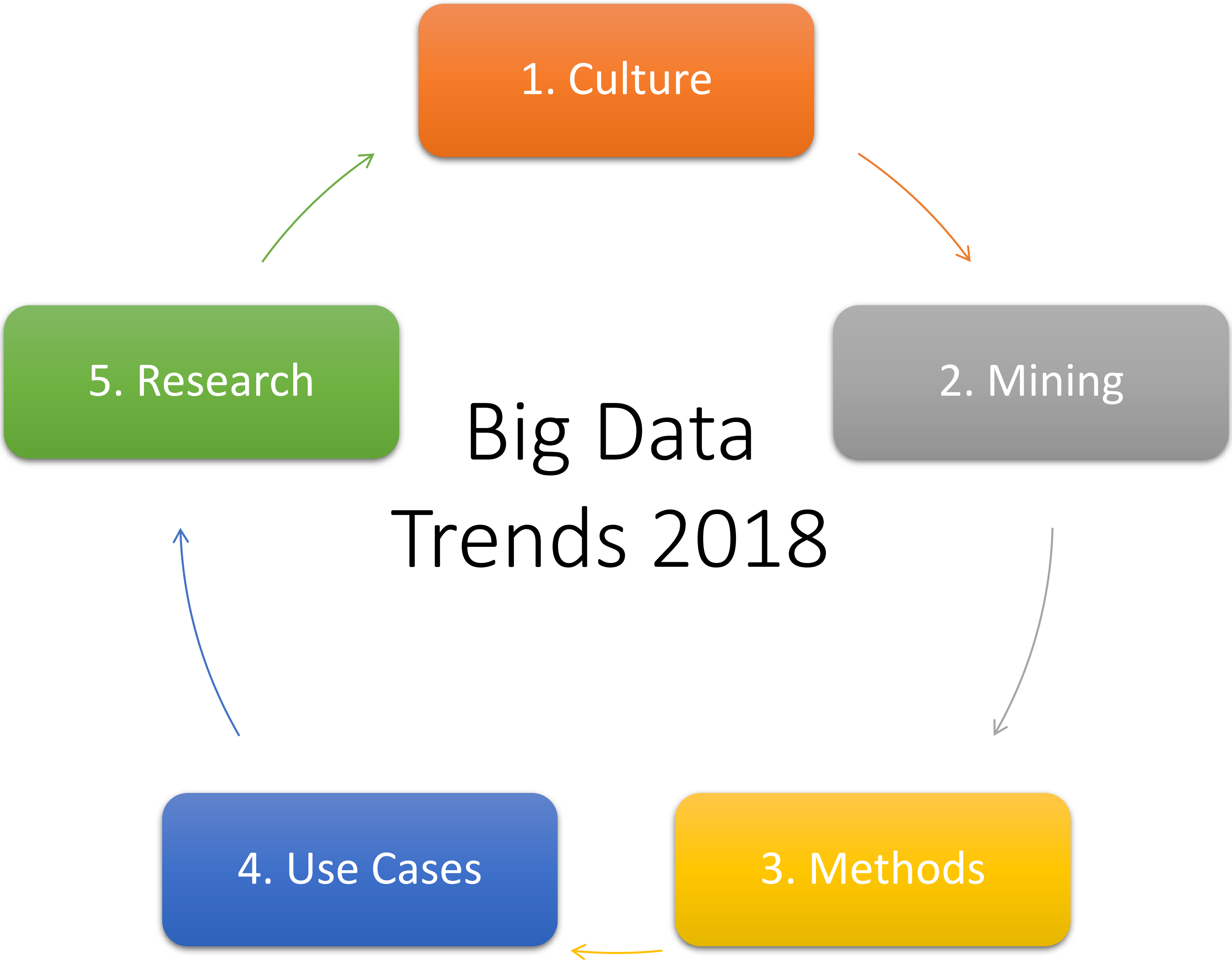
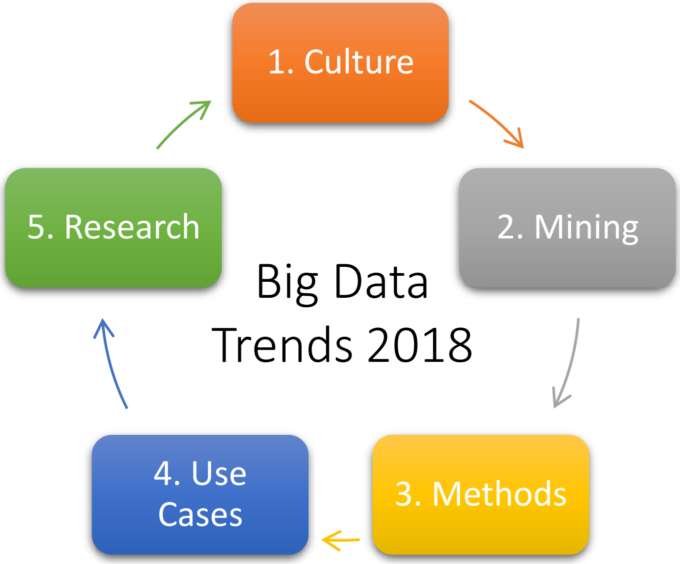
Big Data Trends 2018

Romi Satria Wahono
romi@romisatriawahono.net
http://romisatriawahono.net
Telegram/WA: +6281586220090

Romi Satria Wahono

- SMA Taruna Nusantara Magelang (1993)
- B.Eng, M.Eng and Ph.D in Software Engineering
  Saitama University Japan (1994-2004)
  Universiti Teknikal Malaysia Melaka (2014)
- Research Interests in Software Engineering and Machine Learning
- LIPI Researcher (2004-2007)
- Founder and CEO:
  - PT IlmuKomputerCom Braindevs Sistema
  - PT Brainmatics Cipta Informatika
- Professional Member of IEEE, ACM and PMI
- IT and Research Award Winners from WSIS (United Nations), Kemdikbud, LIPI, etc
- SCOPUS/ISI Indexed Journal Reviewer: Information and Software Technology, Journal of Systems and Software, Software: Practice and Experience, etc
- Industrial IT Certifications: TOGAF, ITIL, CCNA, etc
- Enterprise Architecture Consultant: KPK, Ristek Dikti, LIPI, DJPK Kemenkeu, Kemsos, INSW, Telkom, PLN, PJB, UNSRI, etc


1. Culture
2. Mining
3. Methods
4. Use Cases
5. Research
Big Data
Trends 2018



1. Big Data Culture

Manusia Memproduksi Data

Manusia memproduksi beragam data yang jumlah dan ukurannya sangat besar

- Astronomi
- Bisnis
- Kedokteran
- Ekonomi
- Olahraga
- Cuaca
- Financial
- …

Pertumbuhan Data

Astronomi

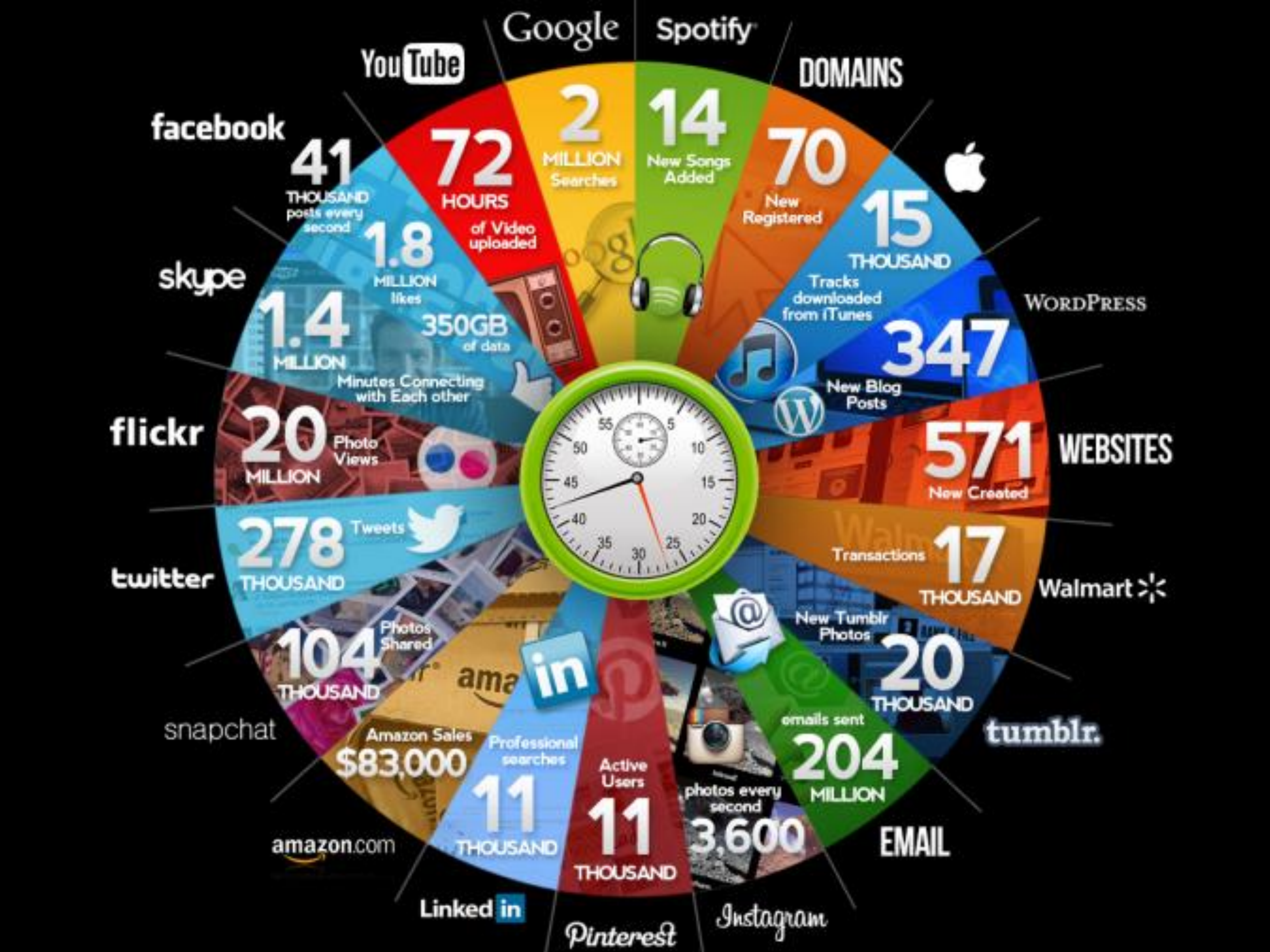
- Sloan Digital Sky Survey
  - New Mexico, 2000
  - 140TB over 10 years
- Large Synoptic Survey Telescope
  - Chile, 2016
  - Will acquire 140TB every five days

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

Perubahan Kultur dan Perilaku


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

Datangnya Tsunami Data

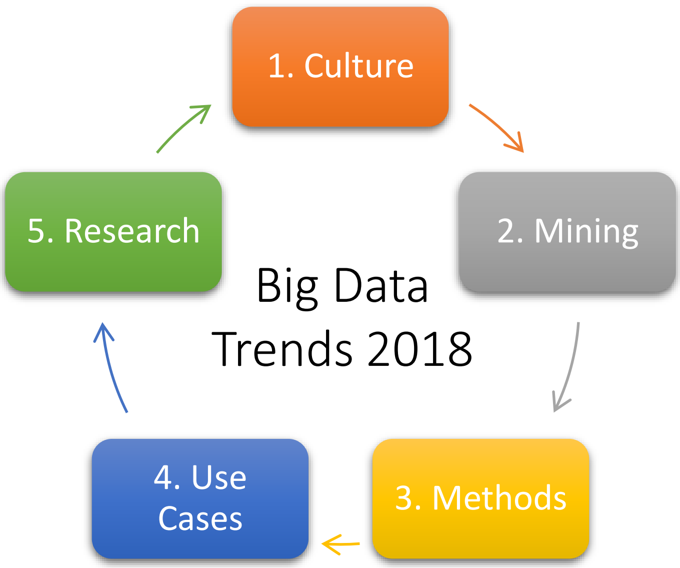
kilobyte (**kB**)	10^3
megabyte (**MB**)	10^6
gigabyte (**GB**)	10^9
terabyte (**TB**)	10^{12}
petabyte (**PB**)	10^{15}
exabyte (**EB**)	10^{18}
zettabyte (**ZB**)	10^{21}
yottabyte (**YB**)	10^{24}

- Mobile Electronics market
  - 4.43B mobile phone users in 2015
  - 7B mobile phone subscriptions in 2015
- Web and 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
  - 235 TBs data collected by the US Library of Congress
  - 15 out of 17 sectors in the US have more data stored per company than the US Library of Congress

Kebanjiran Data tapi Miskin Pengetahuan

We are drowning in data, but starving for knowledge!

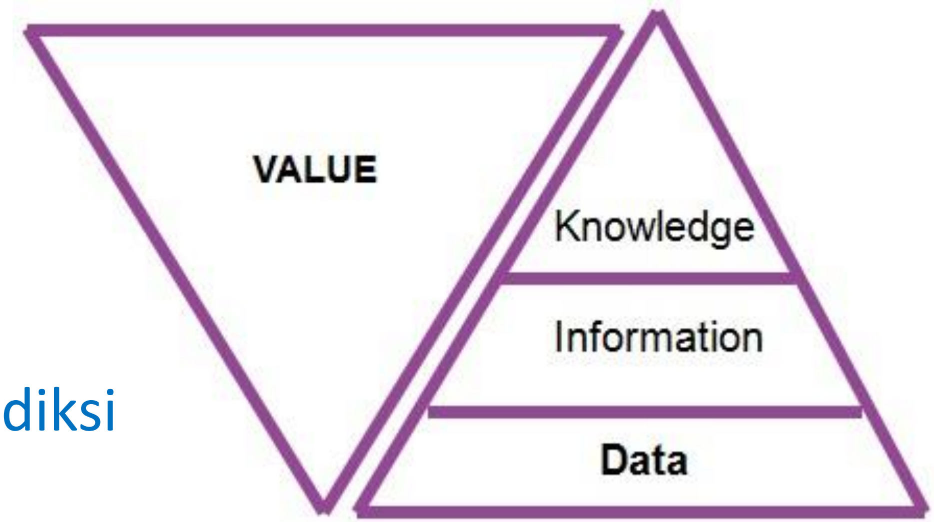
(John Naisbett, Megatrends, 1988)



2. Big Data Mining

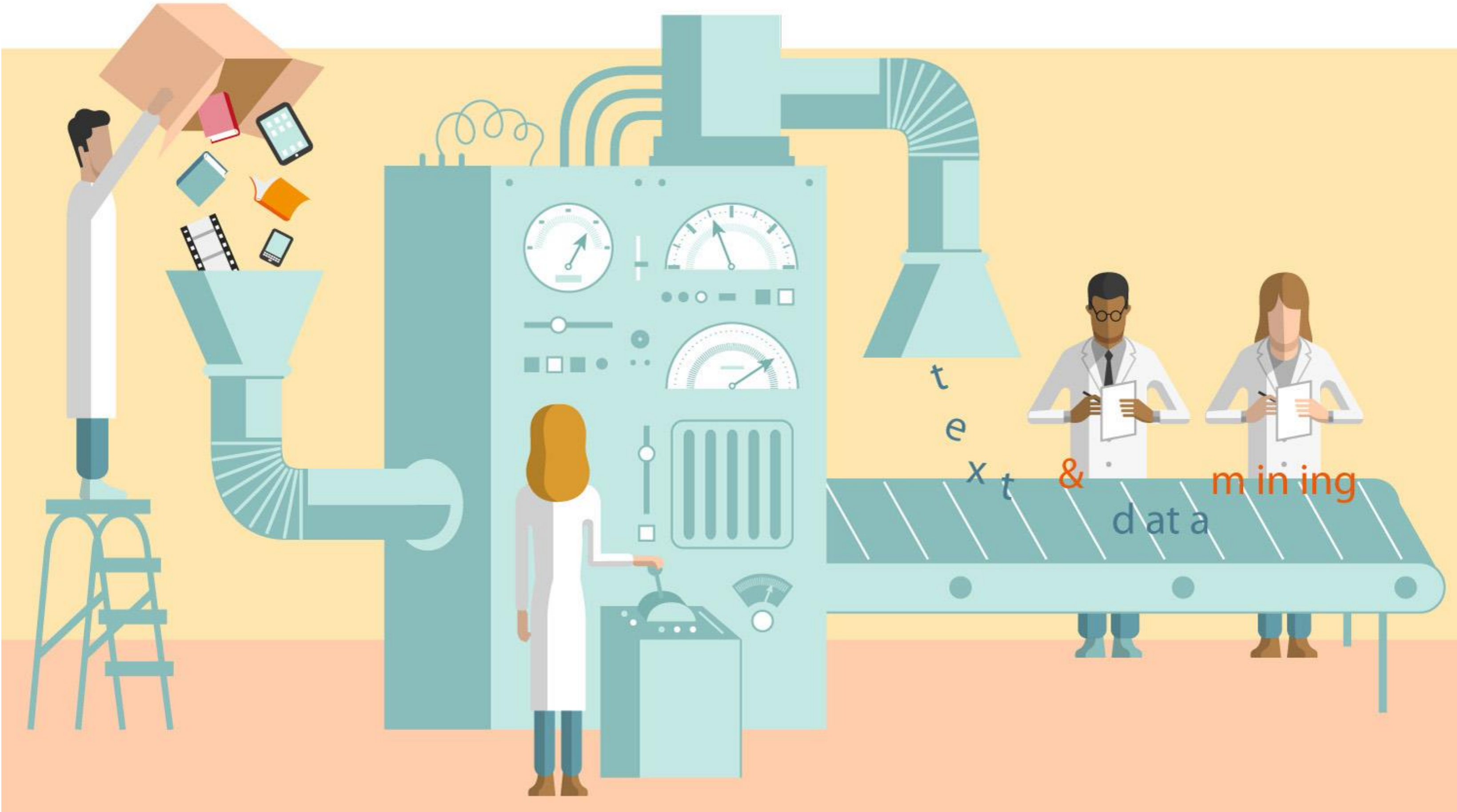
Mengubah Data Menjadi Pengetahuan

- Data harus kita olah menjadi pengetahuan supaya bisa bermanfaat bagi manusia



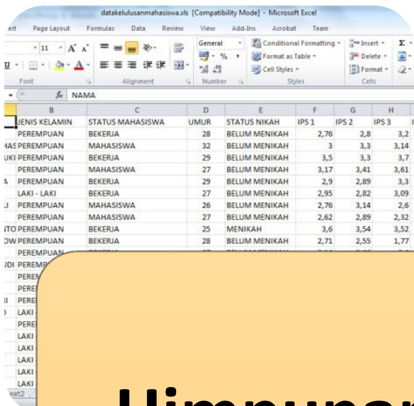
- Dengan pengetahuan tersebut, manusia dapat:
  - Melakukan estimasi dan prediksi apa yang terjadi di depan
  - Melakukan analisis tentang asosiasi, korelasi dan pengelompokan antar data dan atribut
  - Membantu pengambilan keputusan dan pembuatan kebijakan

Memining Data Menjadi Pengetahuan

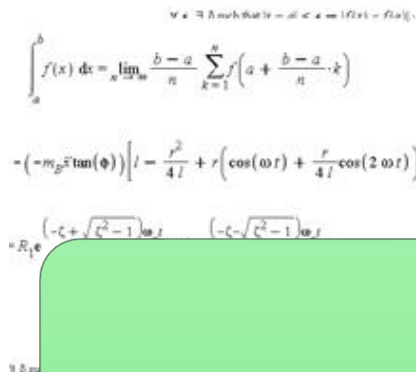


Data Mining: Disiplin ilmu yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data yang besar

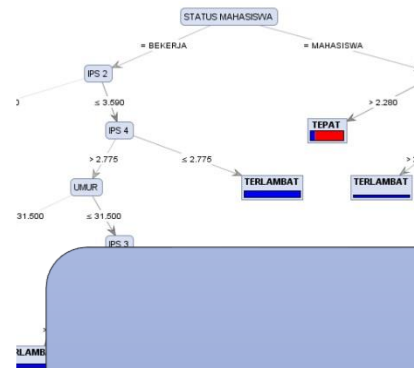
Memining Data Menjadi Pengetahuan



Himpunan Data



Data Mining Methods



Knowledge

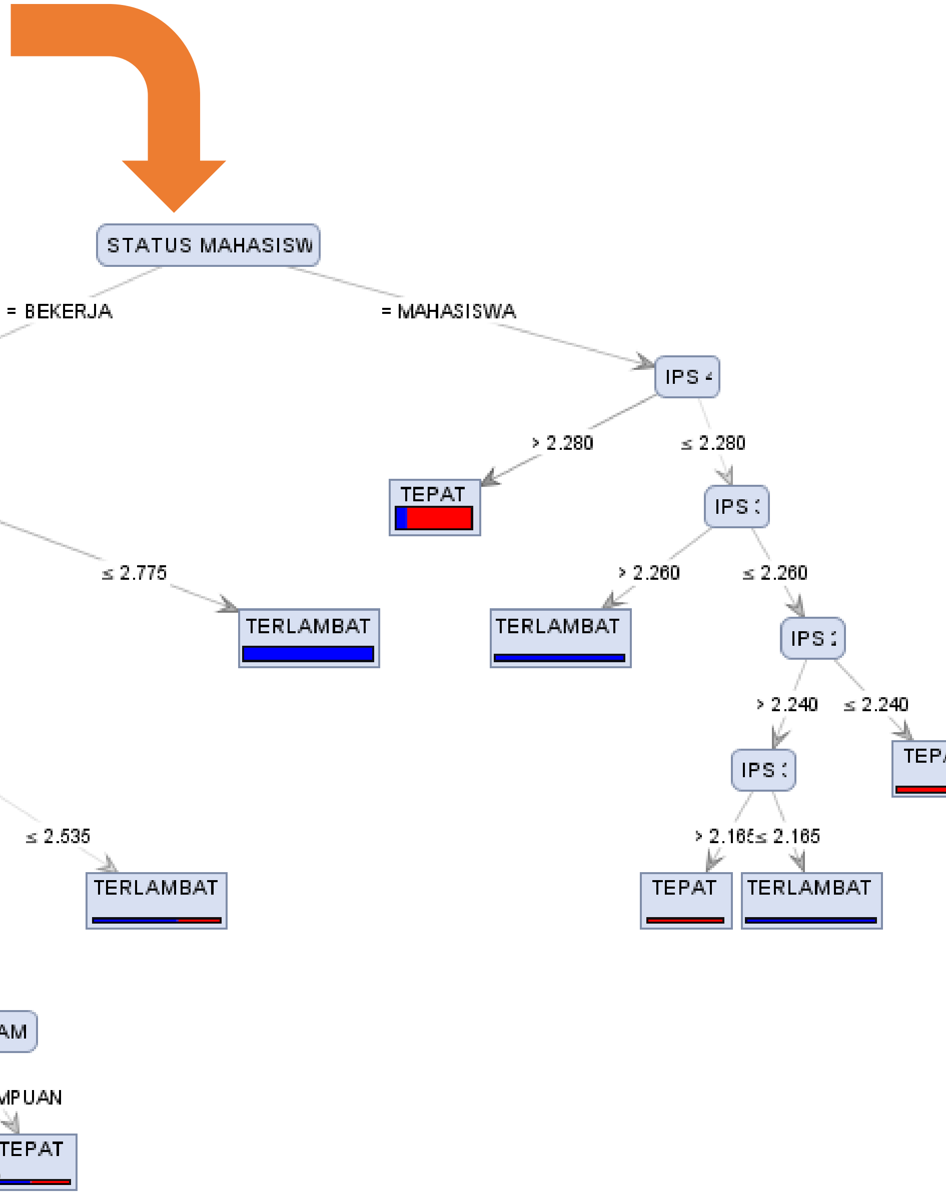
Contoh Data di Kampus

- 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



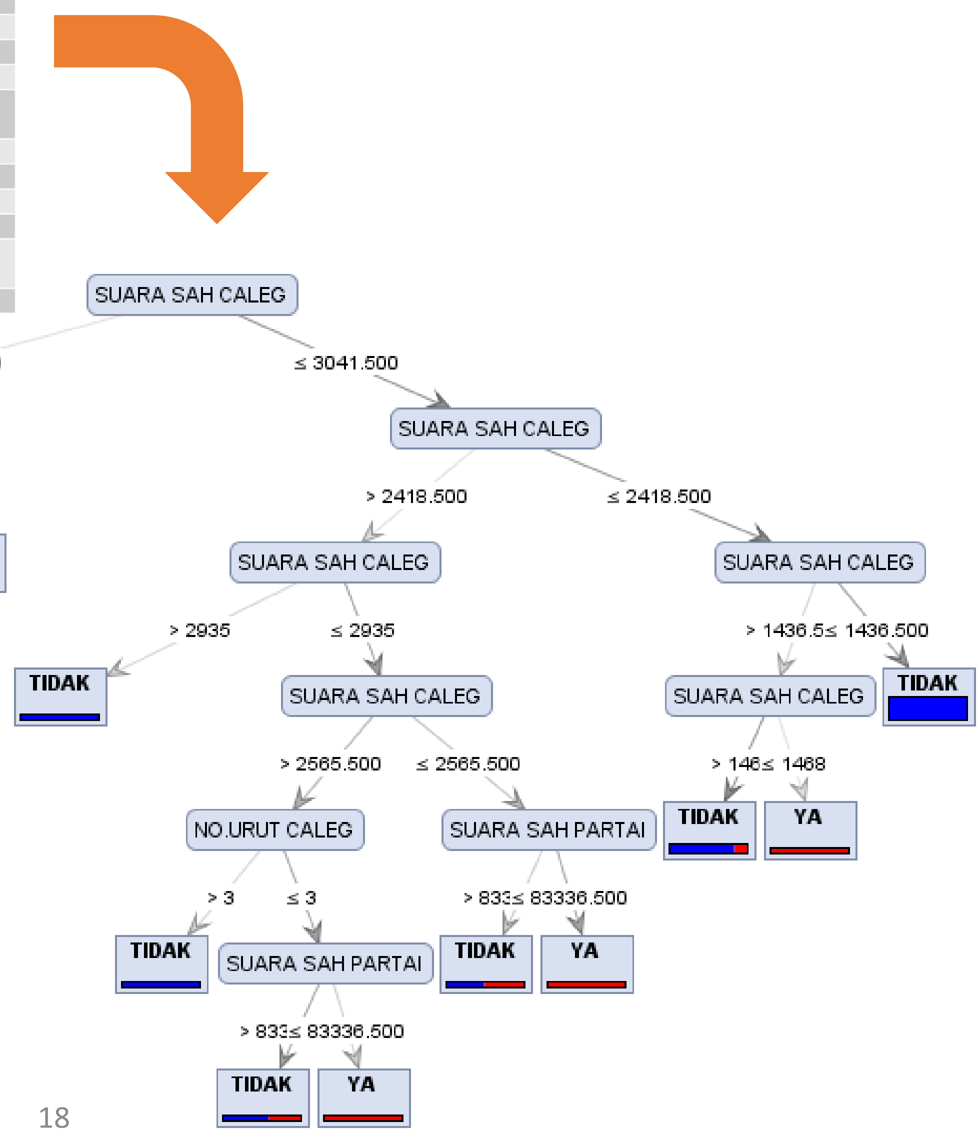
Contoh Data di Komisi Pemilihan Umum

- 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
………	……….	………	……….	………	……	……	……



From Stupid Apps to Smart Apps

Stupid Applications

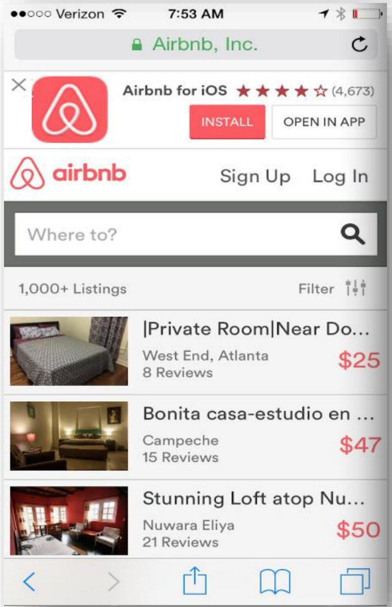
- Sistem Informasi Akademik
- Sistem Pencatatan Pemilu
- Sistem Laporan Kekayaan Pejabat
- Sistem Pencatatan Kredit

Smart Applications

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

Perusahaan Pengolah Pengetahuan

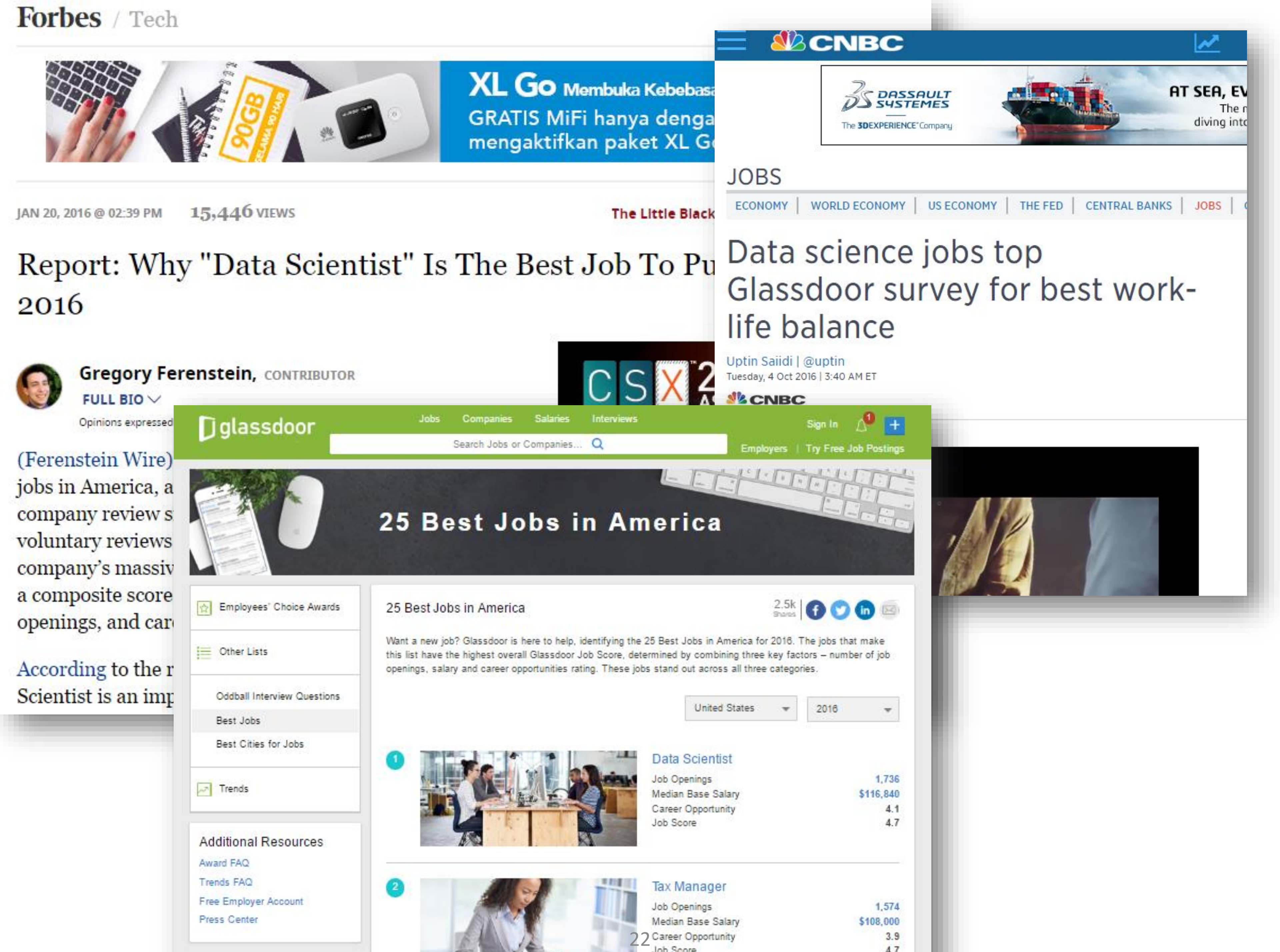
- Uber - the world's largest taxi company, owns no vehicles
- Google - world's largest media/advertising company, creates no content
- Alibaba - the most valuable retailer, has no inventory
- Airbnb - the world's largest accommodation provider, owns no real estate
- Gojek - perusahaan angkutan umum, tanpa memiliki kendaraan
- Groceria – perusahaan penjual sayur dan daging di pasar, tanpa punya toko dan barang dagangan



Evolution of Sciences

- 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

XL Go Membuka Kebebas
GRATIS MiFi hanya denga
mengaktifkan paket XL G

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, CONTRIBUTOR
FULL BIO
Opinions expressed

(Ferenstein Wire)
jobs in America, a
company review s
voluntary reviews
company's massiv
a composite score
openings, and car

According to the r
Scientist is an imp

CNBC

DASSAULT SYSTEMES
The 3DEXPERIENCE Company

AT SEA, EV
The r
diving into

JOBS

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

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

Uptin Saiidi | @uptin
Tuesday, 4 Oct 2016 | 3:40 AM ET

glassdoor

Jobs Companies Salaries Interviews
Search Jobs or Companies...
Sign In
Employers | Try Free Job Postings

25 Best Jobs in America

Employees' Choice Awards
Other Lists
Oddball Interview Questions
Best Jobs
Best Cities for Jobs
Trends

Additional Resources
Award FAQ
Trends FAQ
Free Employer Account
Press Center

25 Best Jobs in America

2.5k Shares

Want a new job? Glassdoor is here to help, identifying the 25 Best Jobs in America for 2016. The jobs that make this list have the highest overall Glassdoor Job Score, determined by combining three key factors – number of job openings, salary and career opportunities rating. These jobs stand out across all three categories.

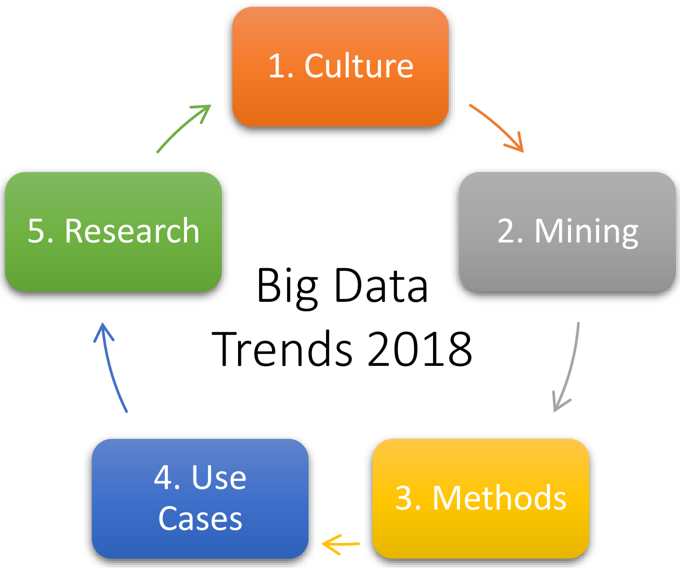
United States
2016

1 Data Scientist

Job Openings	1,736
Median Base Salary	$116,840
Career Opportunity	4.1
Job Score	4.7

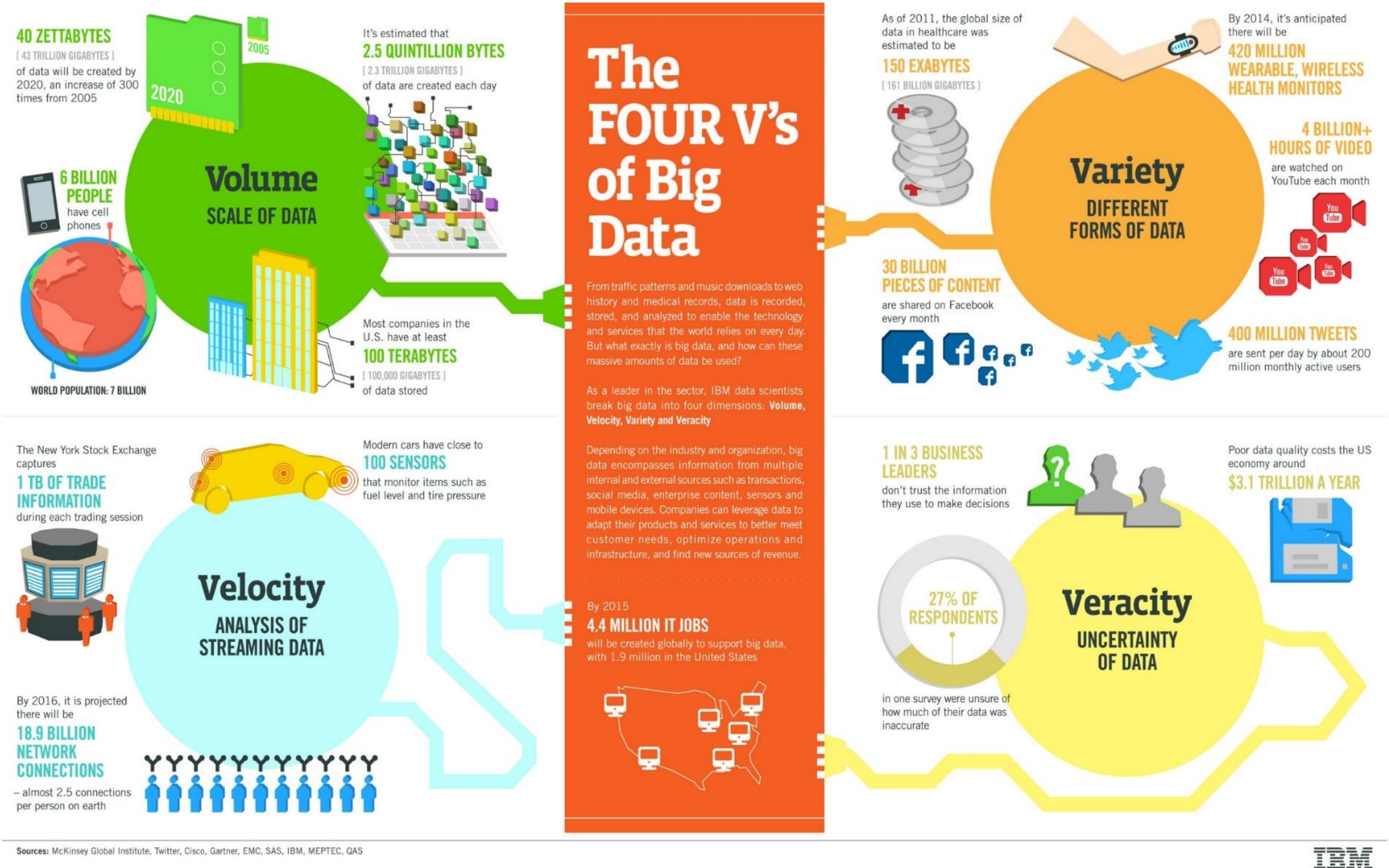
2 Tax Manager

Job Openings	1,574
Median Base Salary	$108,000
Career Opportunity	3.9
Job Score	4.7



3. Big Data Methods and Technologies

Empat Dimensi Masalah Big Data

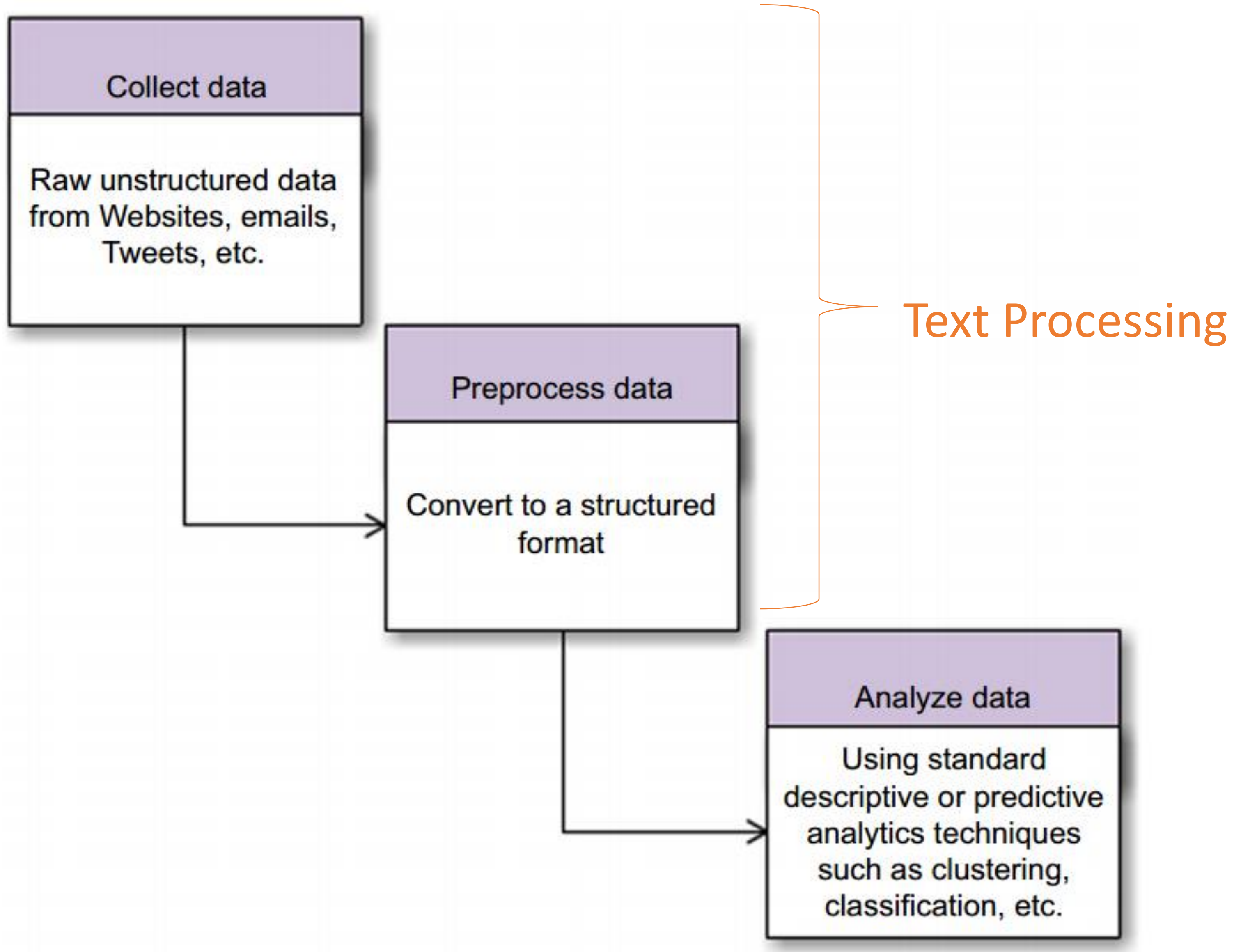


Big Data Methods

1. Text Mining:
   - Mengolah data tidak terstruktur dalam bentuk text, web, social media, dsb
   - Menggunakan metode text processing untuk mengkonversi data tidak terstruktur menjadi terstruktur
     - Kemudian diolah dengan data mining

2. Data Mining:
   - Mengolah data terstruktur dalam bentuk tabel yang memiliki atribut dan kelas
   - Menggunakan metode data mining, yang terbagi menjadi metode estimasi, forecasting, klasifikasi, klastering atau asosiasi
     - Yang dasar berpikirnya menggunakan konsep statistika atau heuristik ala machine learning

Text Mining

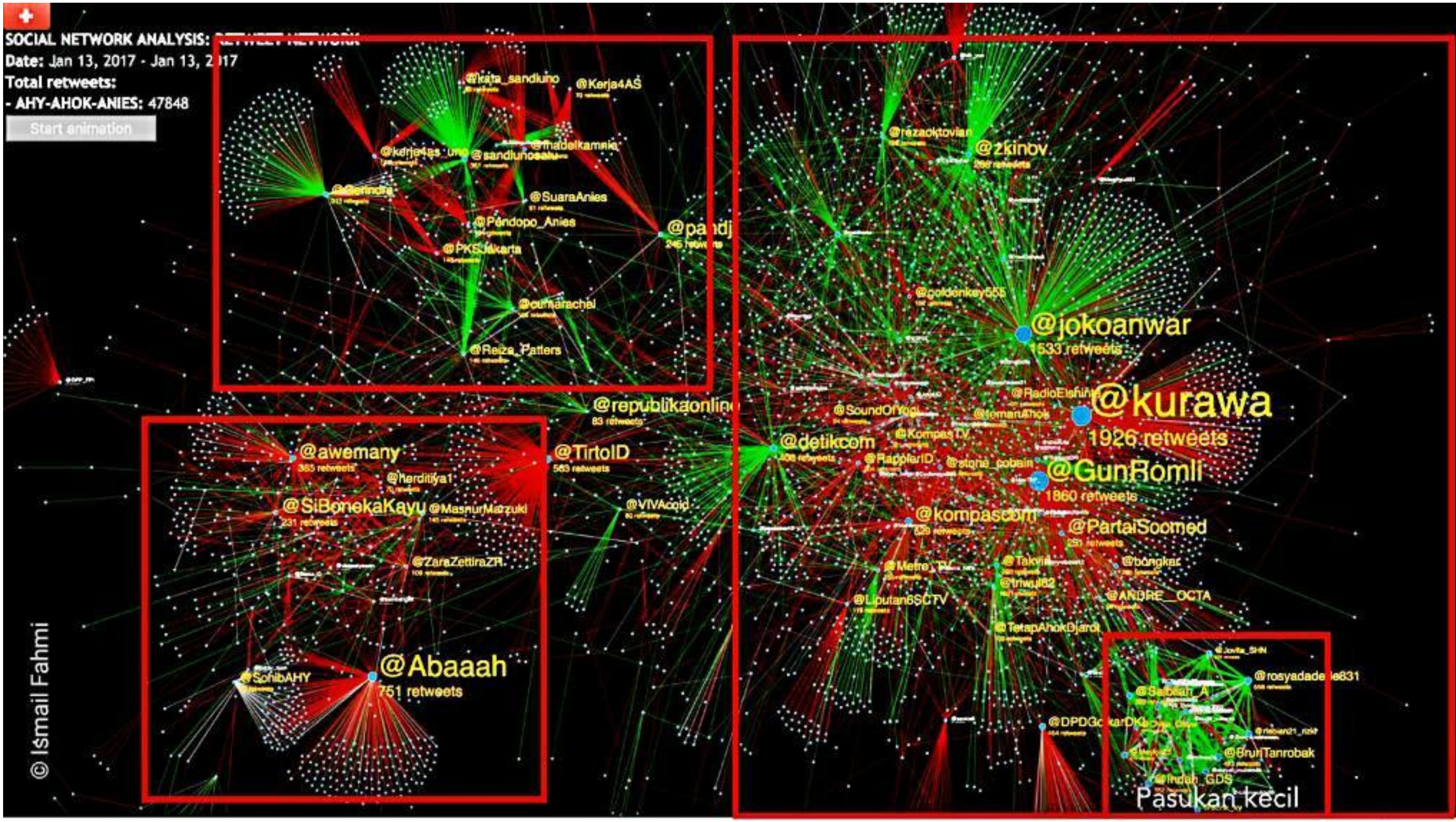


Collect data
Raw unstructured data from Websites, emails, Tweets, etc.
Preprocess data
Convert to a structured format
Analyze data
Using standard descriptive or predictive analytics techniques such as clustering, classification, etc.
Text Processing

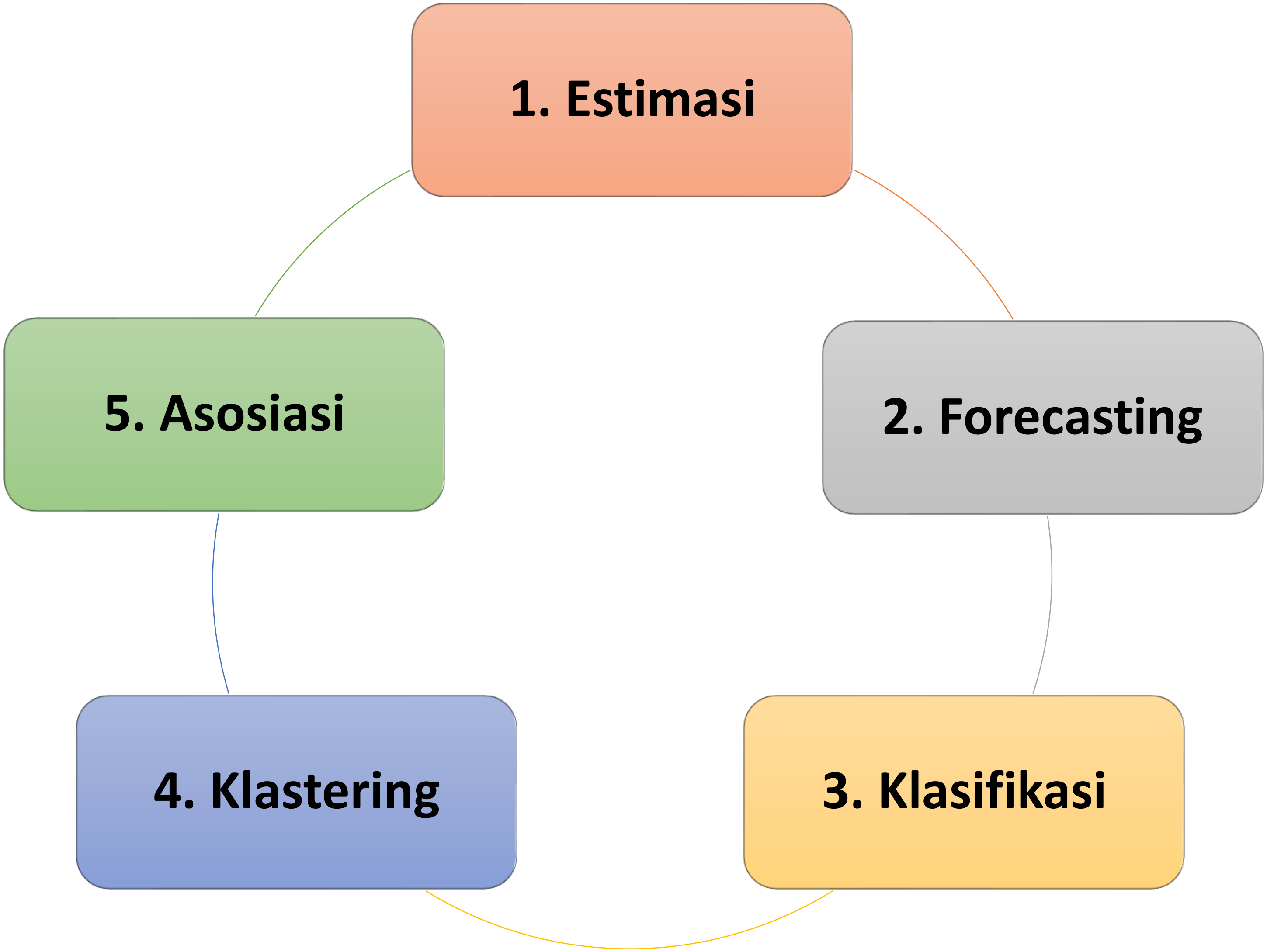
Text Mining

Jejak Pornografi di Indonesia

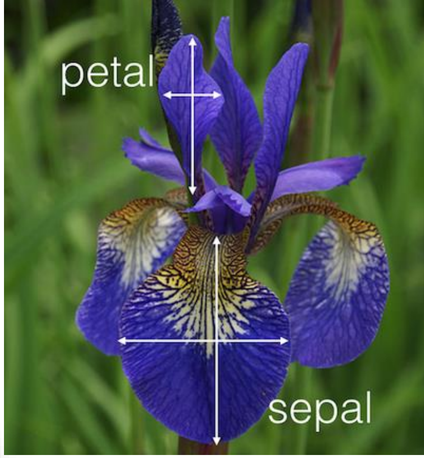
Text Mining: AHY-AHOK-ANIES



Data Mining


1. Estimasi
2. Forecasting
3. Klasifikasi
4. Klastering
5. Asosiasi

Dataset (Himpunan Data)



Attribute/Feature

Class/Label

	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*

Record/
Object/
Sample

Nominal

Numerik

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

Label

Pembelajaran dengan
Metode Estimasi (*Regresi Linier*)

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

Pengetahuan

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

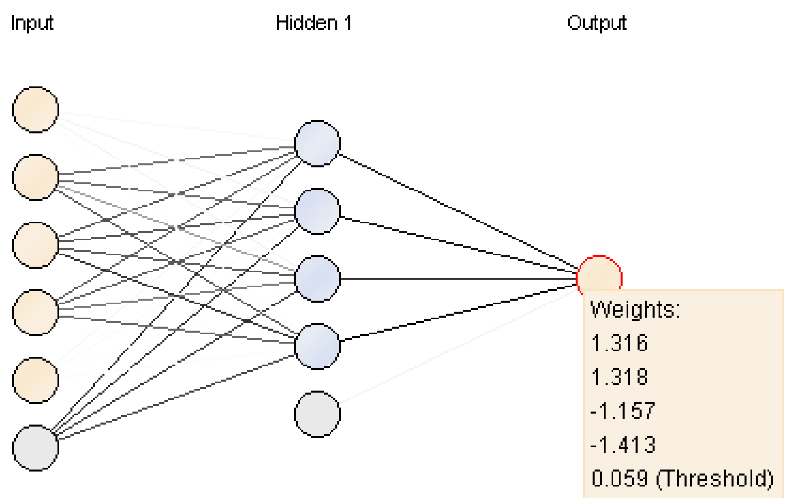
+ 0.6410 CACH - 0.2700 CHMIN + 1.480 CHMAX

2. Forecasting Harga Saham

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

Dataset harga saham dalam bentuk time series (rentet waktu)

Pembelajaran dengan Metode Forecasting (*Neural Network*)



Forecasting Kurs Mata Uang

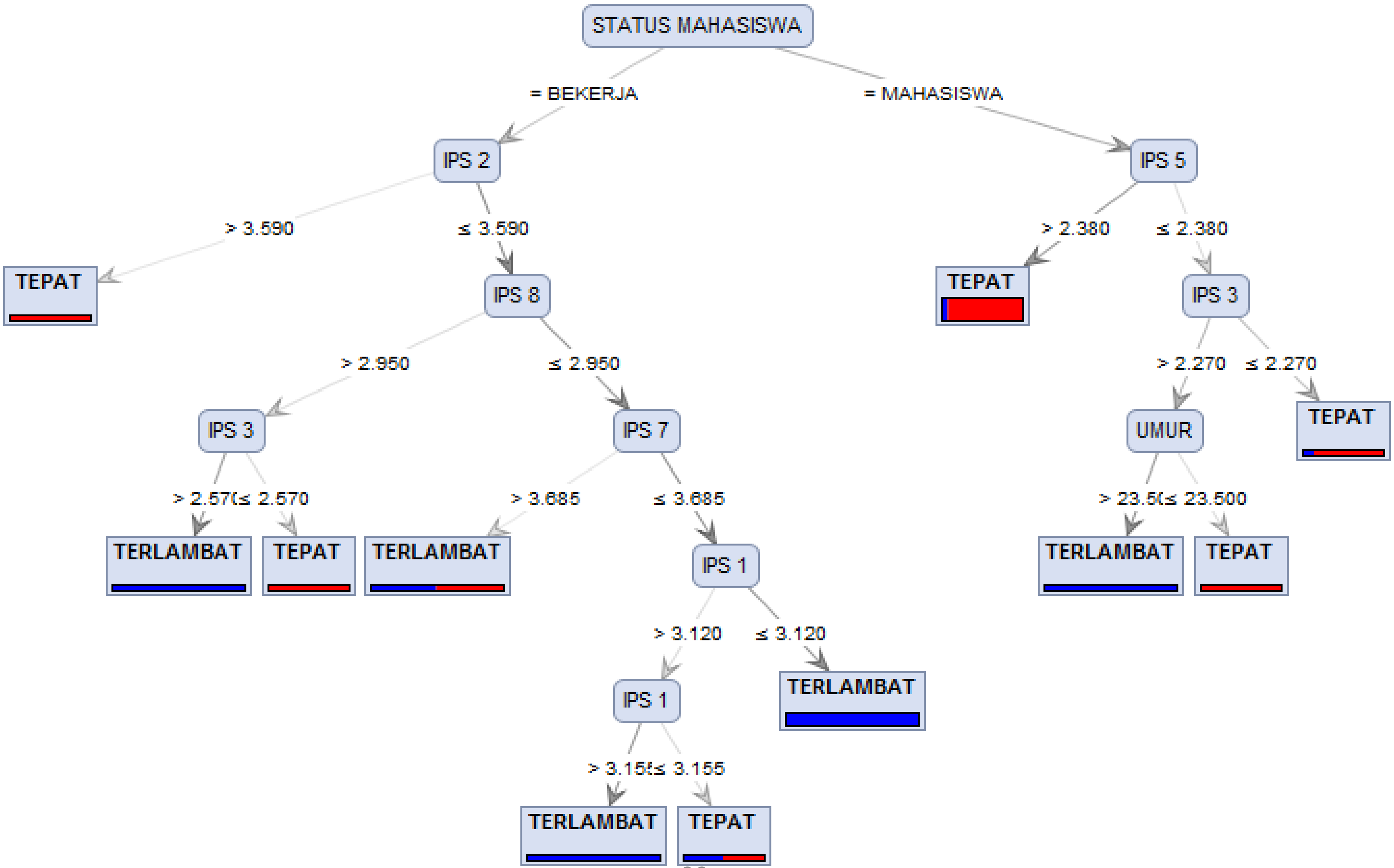
3. Klasifikasi Kelulusan Mahasiswa

Label

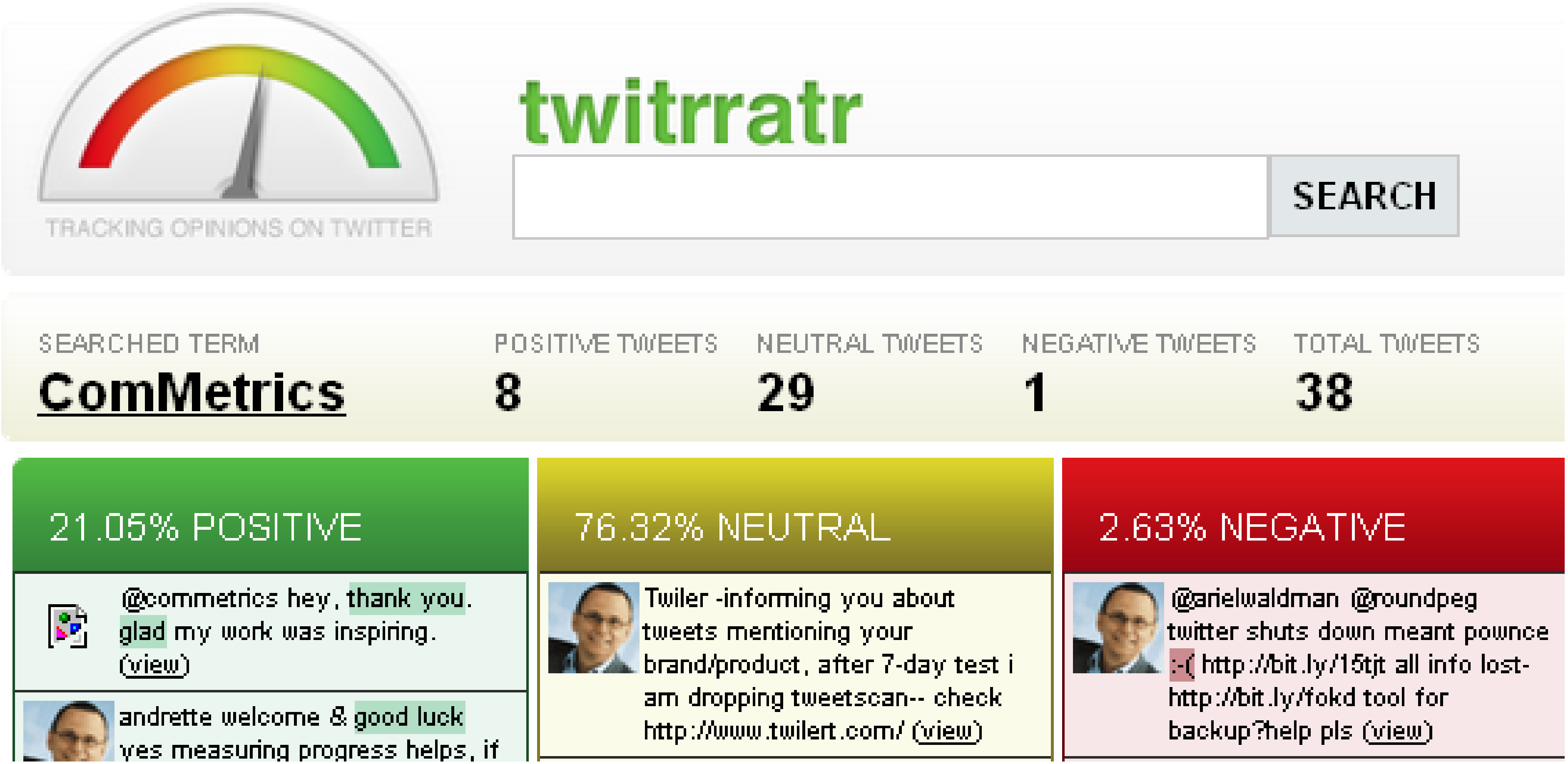
NIM	Gender	Nilai UN	Asal Sekolah	IPS1	IPS2	IPS3	IPS 4	...	Lulus Tepat Waktu
10001	L	28	SMAN 2	3.3	3.6	2.89	2.9		Ya
10002	P	27	SMA DK	4.0	3.2	3.8	3.7		Tidak
10003	P	24	SMAN 1	2.7	3.4	4.0	3.5		Tidak
10004	L	26.4	SMAN 3	3.2	2.7	3.6	3.4		Ya
…									
…									
11000	L	23.4	SMAN 5	3.3	2.8	3.1	3.2		Ya

Pembelajaran dengan
Metode Klasifikasi (*C4.5*)

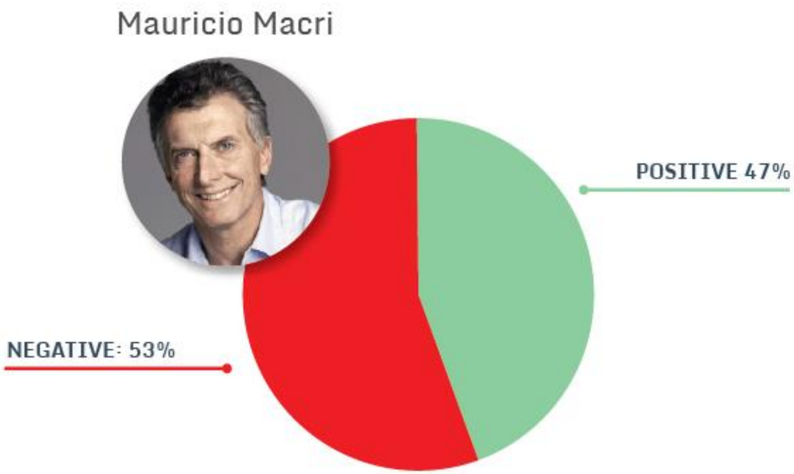
Pengetahuan Berupa Pohon Keputusan

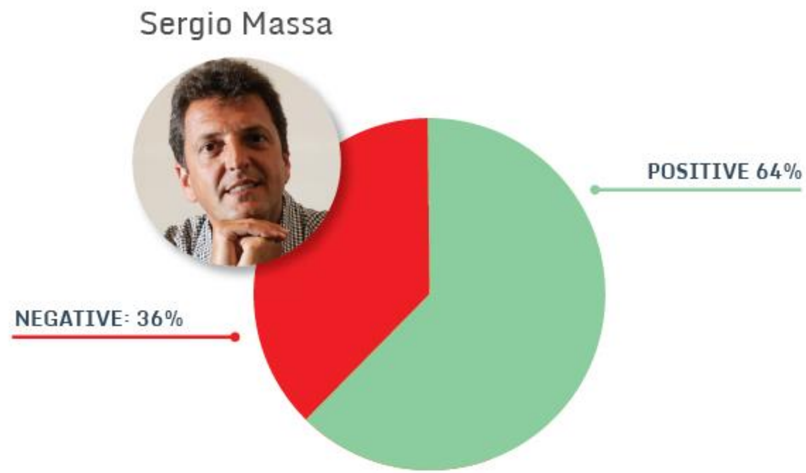


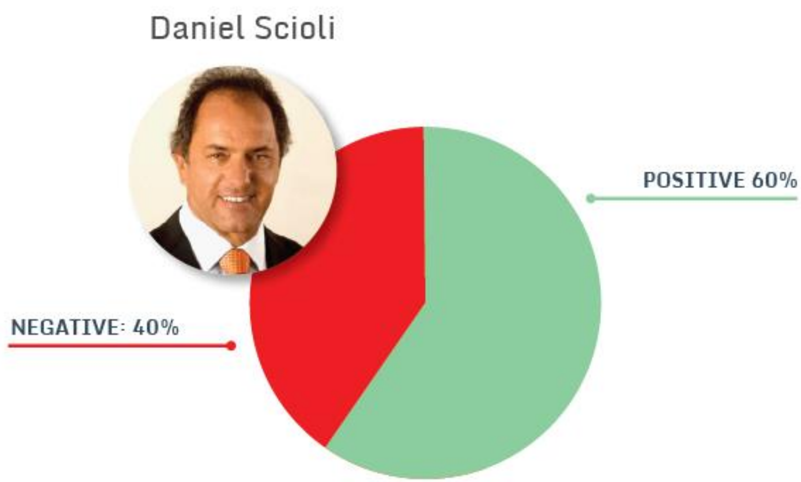
Klasifikasi Sentimen Analisis



Klasifikasi Sentimen Analisis





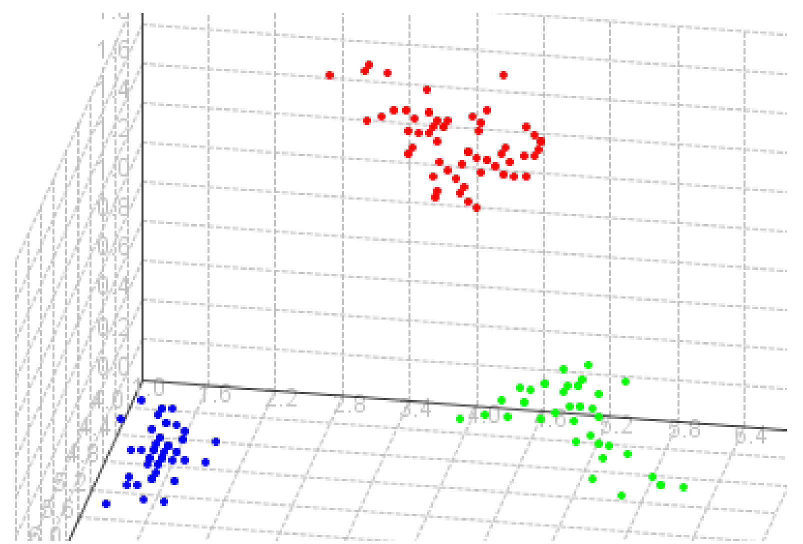


4. Klastering Bunga Iris

Dataset Tanpa Label

Row No.	id	a1	a2	a3	a4
1	id_1	5.100	3.500	1.400	0.200
2	id_2	4.900	3	1.400	0.200
3	id_3	4.700	3.200	1.300	0.200
4	id_4	4.600	3.100	1.500	0.200
5	id_5	5	3.600	1.400	0.200
6	id_6	5.400	3.900	1.700	0.400
7	id_7	4.600	3.400	1.400	0.300
8	id_8	5	3.400	1.500	0.200
9	id_9	4.400	2.900	1.400	0.200
10	id_10	4.900	3.100	1.500	0.100
11	id_11	5.400	3.700	1.500	0.200

Pembelajaran dengan
Metode Klastering (*K-Means*)



Klastering Jenis Pelanggan

Klastering Sentimen Warga

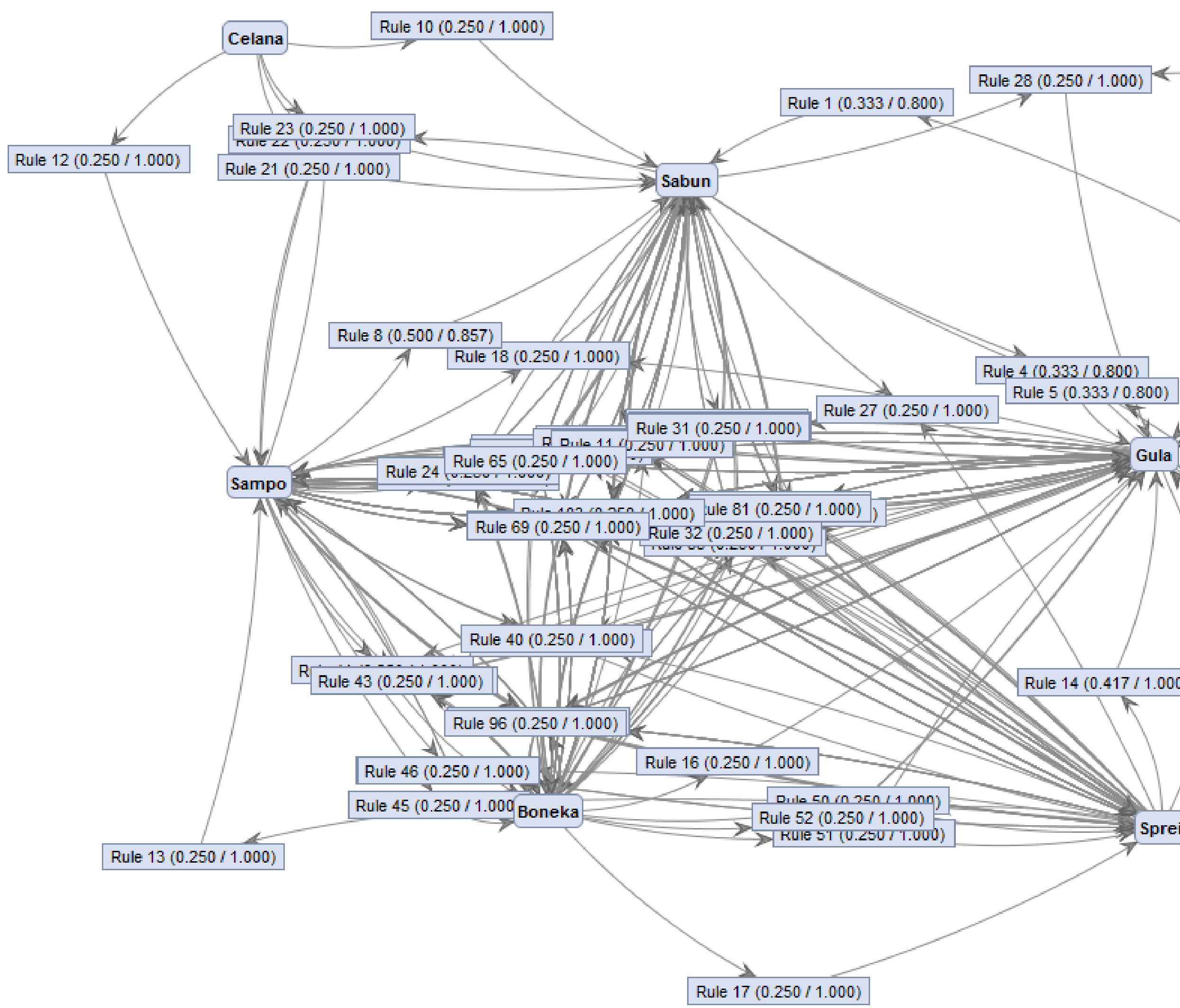
5. Aturan Asosiasi Pembelian Barang

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
3	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0
4	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0
6	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0
8	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0
9	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
10	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
11	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0

Pembelajaran dengan
Metode Asosiasi (*FP-Growth*)

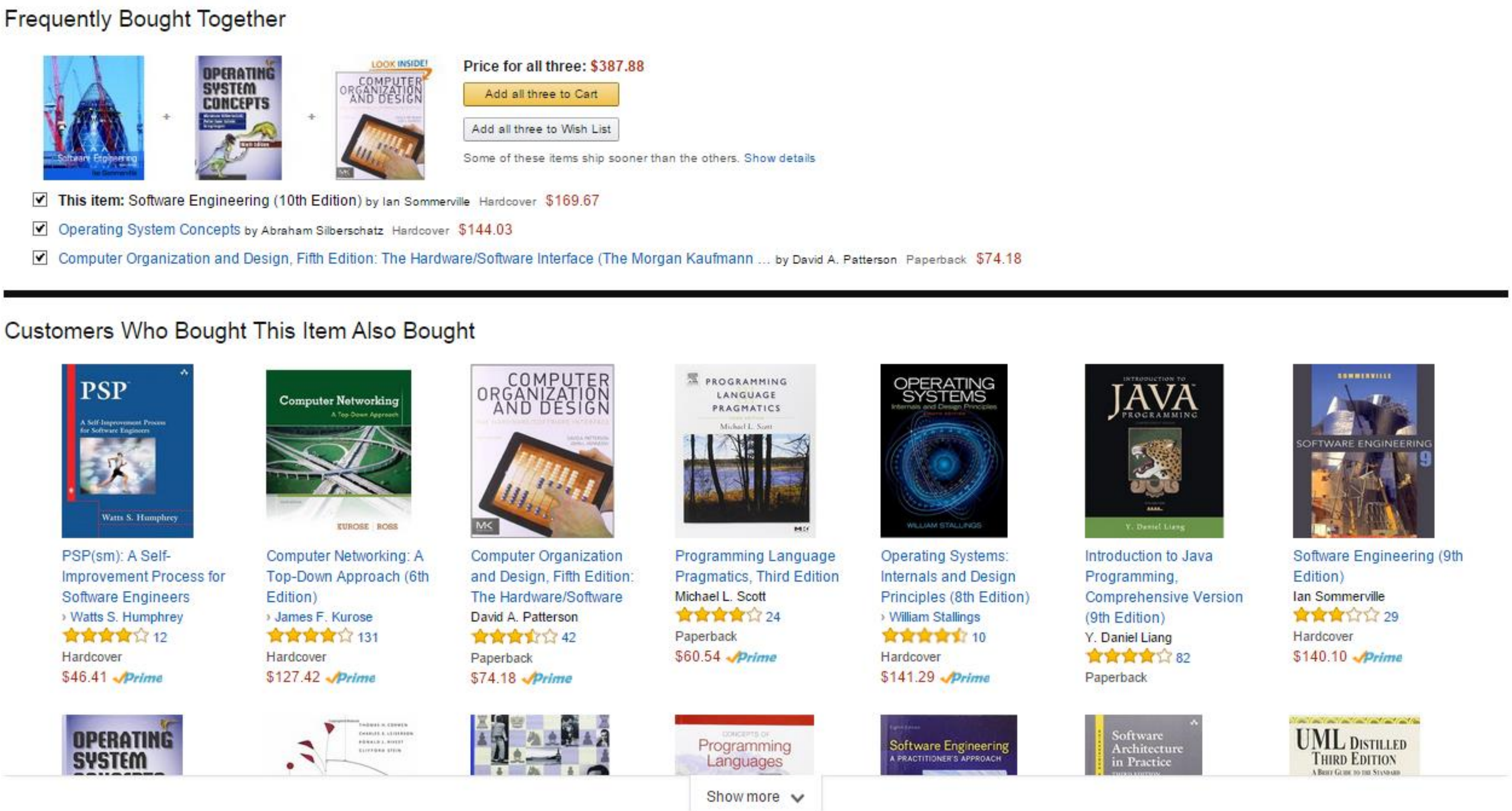
Pengetahuan Berupa Aturan Asosiasi



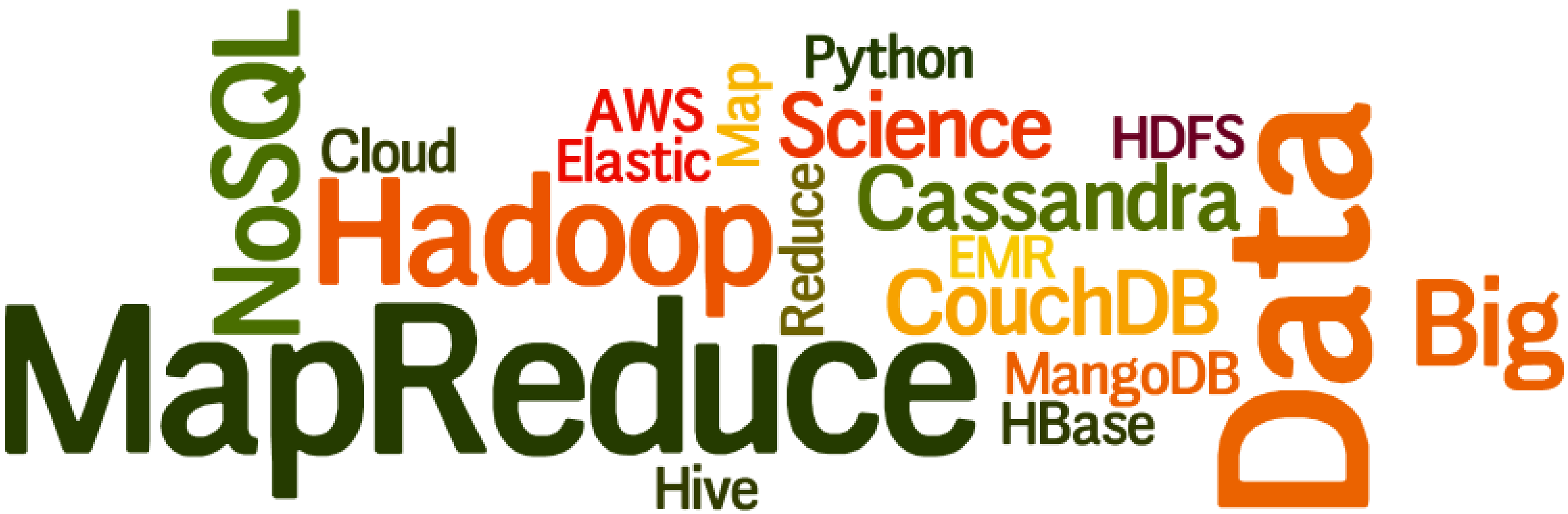
AssociationRules

```
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)
[Popok] --> [Gula] (confidence: 1.000)
[Boneka] --> [Gula] (confidence: 1.000)
[Boneka] --> [Sprei] (confidence: 1.000)
[Sampo, Gula] --> [Sabun] (confidence: 1.000)
[Sabun, Sprei] --> [Sampo] (confidence: 1.000)
[Sampo, Sprei] --> [Sabun] (confidence: 1.000)
[Celana] --> [Sabun, Sampo] (confidence: 1.000)
[Sabun, Celana] --> [Sampo] (confidence: 1.000)
[Sampo, Celana] --> [Sabun] (confidence: 1.000)
[Boneka] --> [Sabun, Sampo] (confidence: 1.000)
[Sabun, Boneka] --> [Sampo] (confidence: 1.000)
[Sampo, Boneka] --> [Sabun] (confidence: 1.000)
[Sabun, Sprei] --> [Gula] (confidence: 1.000)
```

Aturan Asosiasi di Amazon.com



Teknologi Pendukung Big Data?



BIG DATA LANDSCAPE

Infrastructure

NoSQL Databases
10gen, DATASTAX, Neo4j, HYPERTABLE, COUCHBASE, Cloudant, basho

Hadoop Related
HADAPT, infochimps, Hortonworks, cloudera, MAPR, Zettaset, amazon webservices, IBM

NewSQL Databases
MarkLogic, VoltDB, SQLFire, memsql, paradigm4, DRAWNTOSCALE

Collection / Transport
aspera, nodeable

MPP Databases
VERTICA, kognitio, PARACCEL, GREENPLUM

Management / Monitoring
oceansync, StackIQ, boundary, DATADOG

Cluster Services
LexisNexis, HPCC Systems, Acunu

Storage
Cleversafe, panasas, nimble storage

Crowdsourcing
CROWDCOMPUTING, CrowdFlower, amazon

Security
Stormpath, IMPERVA, codefortytwo software, DATAGUISE

Analytics

Analytics Solutions
Palantir, platfora, PERVASIVE, Datameer, PRECOG, DIGITAL REASONING

Data Visualization
Quid, ACTUATE, Kitenga, visual.ly, centrifuge, metaLayer

Statistical Computing
sas, REVOLUTION ANALYTICS, SKYTREE, p(k)

Social Media
bit.ly, tracx, bluefin, Dataminr

Sentiment Analysis
GENERAL SENTIMENT, crimson hexagon

Analytics Services
THINK BIG, accenture, McKinsey&Company

Big Data Search
elasticsearch, Autonomy

IT Analytics
splunk, sumologic

Location / People / Events
RapLeaf, Fliptop, Recorded Future, Placed

Real-Time
CONTINUUITY, feedzai

Crowdsourced Analytics
DataKind, kaggle

SMB Analytics
sumAll, custora

Applications

Ad Optimization
DataXu, aggregate knowledge, m6d, aiMatch, thetradedesk, bluekai, rocketfuel, TURN, 33across, MediaMath

Publisher Tools
VISUAL revenue, Yieldex, yieldbot

Application Service Provider
collective[i]

Industry Application
NEXT BIG SOUND, KNEWTON, Bloomberg, BILLGUARD, numberFire, zestcash, Climate Solutions

Marketing
LATTICE ENGINES, Sailthru, bloomreach, CLICKFOX

Data Sources

Data Marketplaces
DataMarket, Windows Azure, factual

Data Sources
DATASIFT, knoema, GNIP, infochimps

Ad Optimization
Withings, RunKeeper, JAWBONE, Nike+

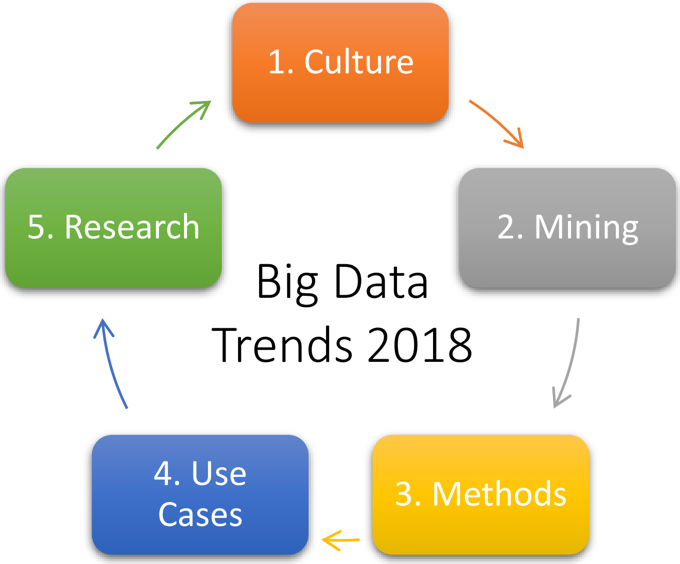
Cross Infrastructure / Analytics
SAP, sas, IBM, Google, ORACLE, Microsoft, amazon, 1010data, vmware, METAMARKETS, TERADATA, NetApp

Open Source Project

Framework	Query / Data Flow	Data Access	Coordination / Workflow	Real-Time	Statistical Tools	Machine Learning	Cloud Deployment	Messaging Framework
hadoop, HDFS, Apache Hadoop YARN, Spark	HIVE	cassandra, CouchDB, SciDB, mongoDB	ZooKeeper, talend, OOZIE	Storm	SciPy, R	mahout, Spark	WHIRR	kafka

Magic Quadrant for Advanced Analytics Platform *(Gartner, 2016)*

Big Data Analytics Solution 2015 *(The Forrester Wave)*

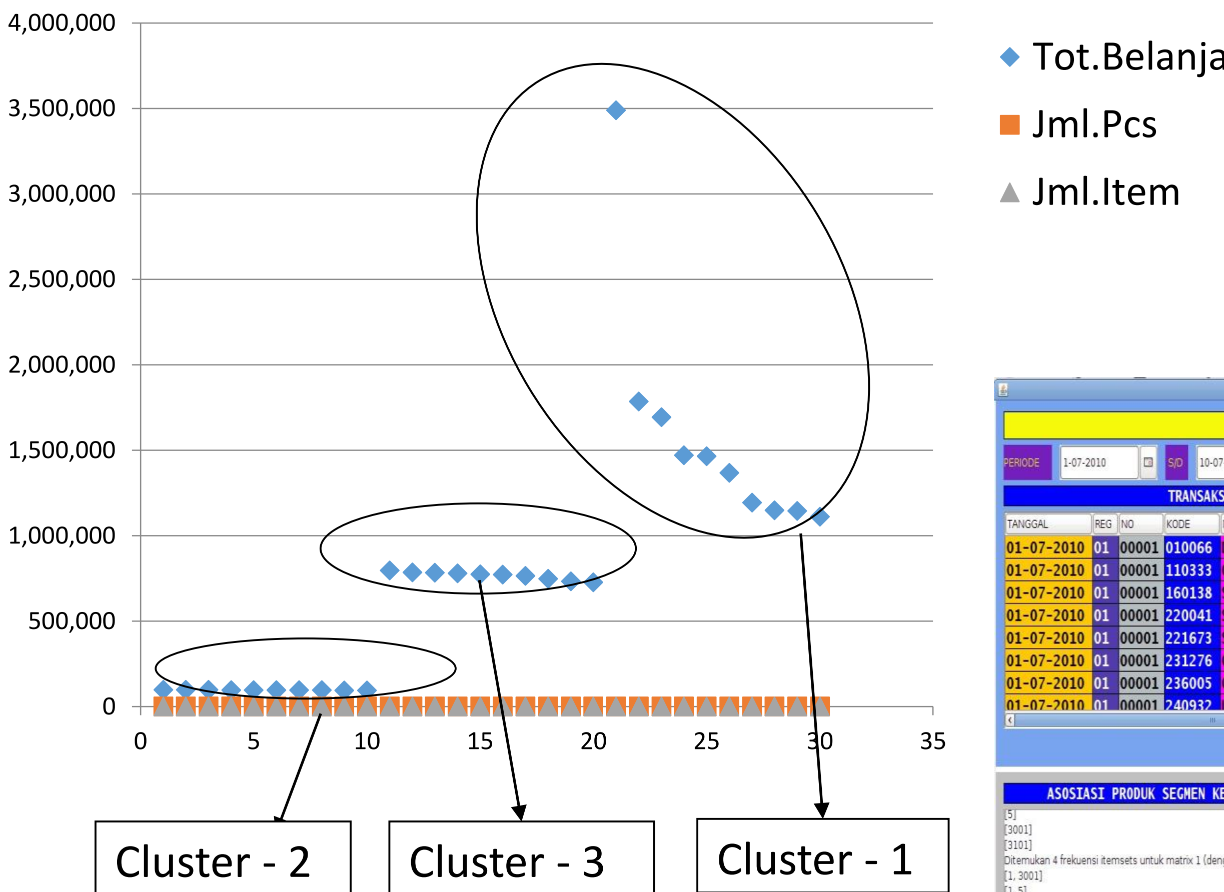


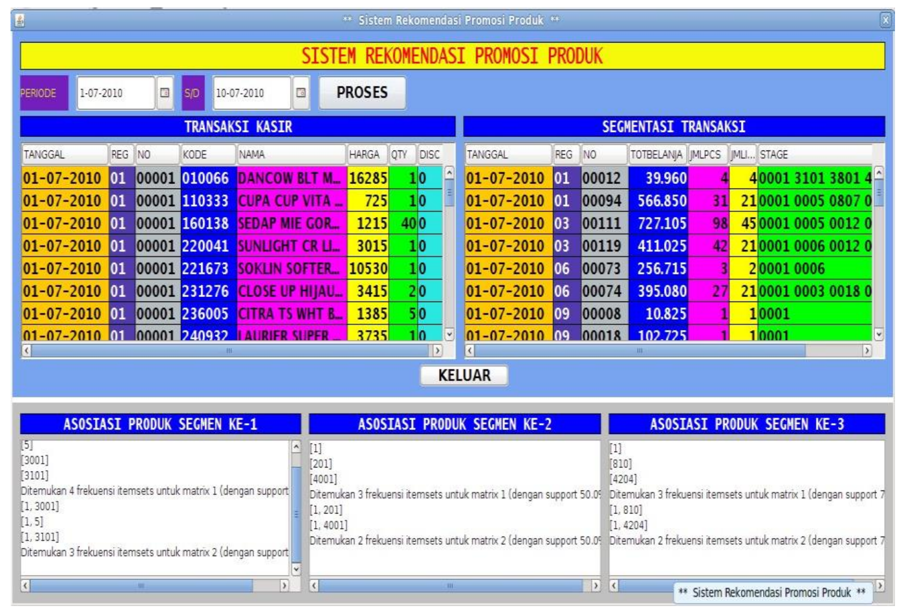
4. Big Data Use Cases

Private and Commercial Sector

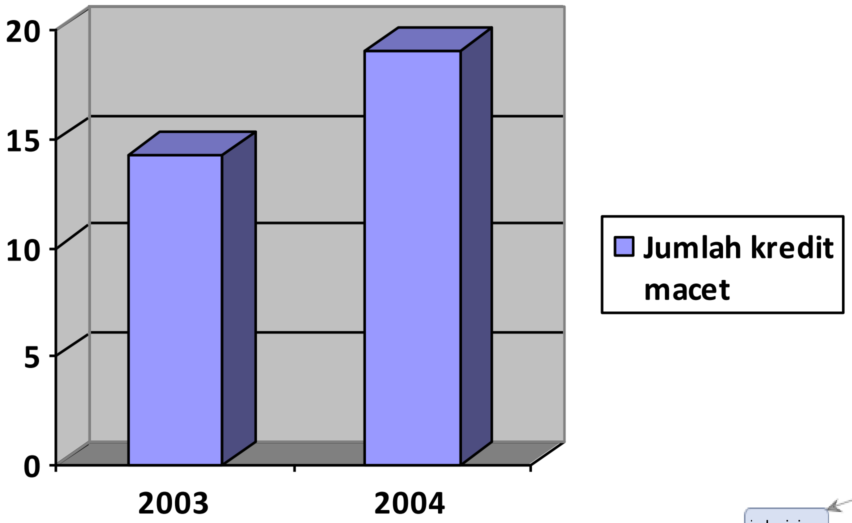
- Marketing: product recommendation, market basket analysis, product targeting, customer retention
- Finance: investment support, portfolio management, price forecasting
- Banking and Insurance: credit and policy approval, money laundry detection
- Security: fraud detection, access control, intrusion 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

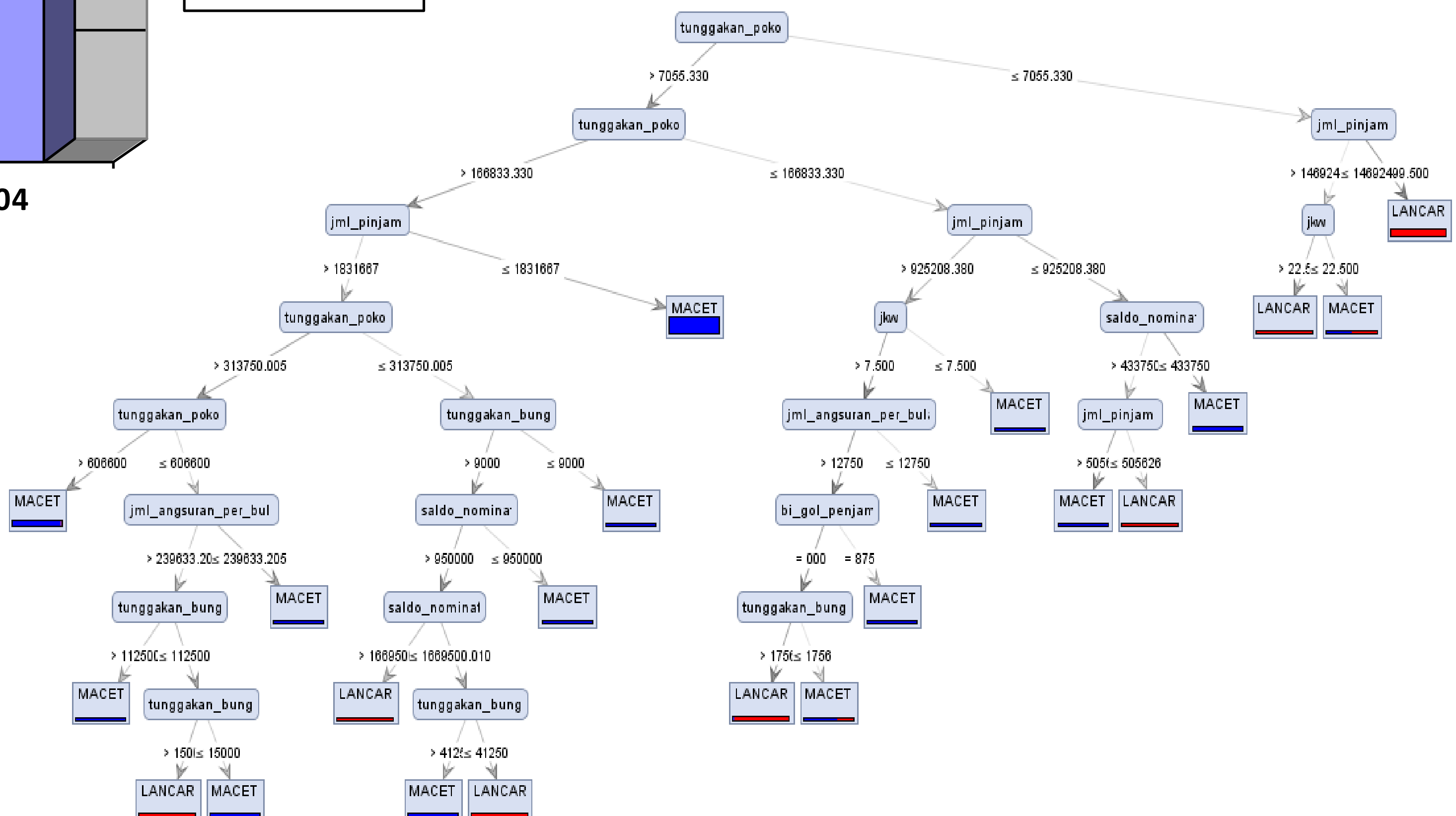
Use Case: Product Recommendation





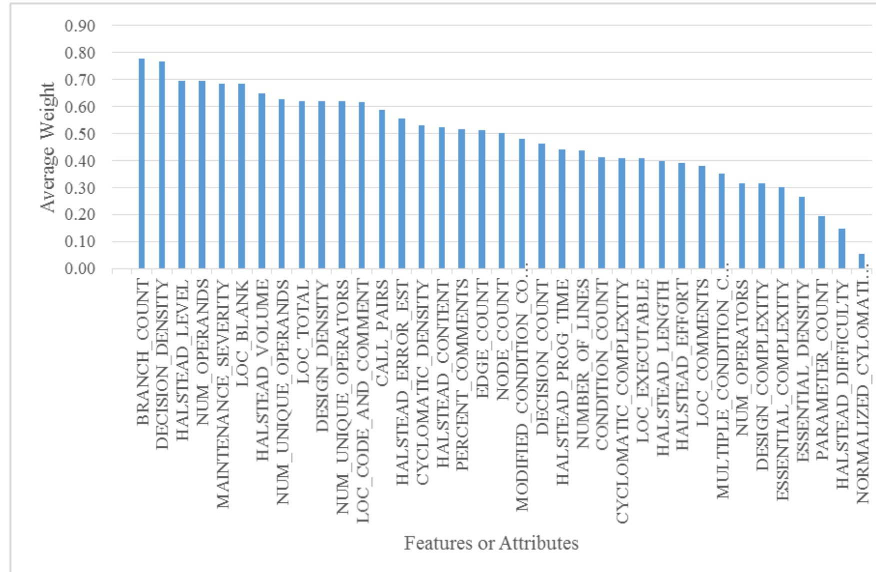
Use Case: Penentuan Kelayakan Kredit

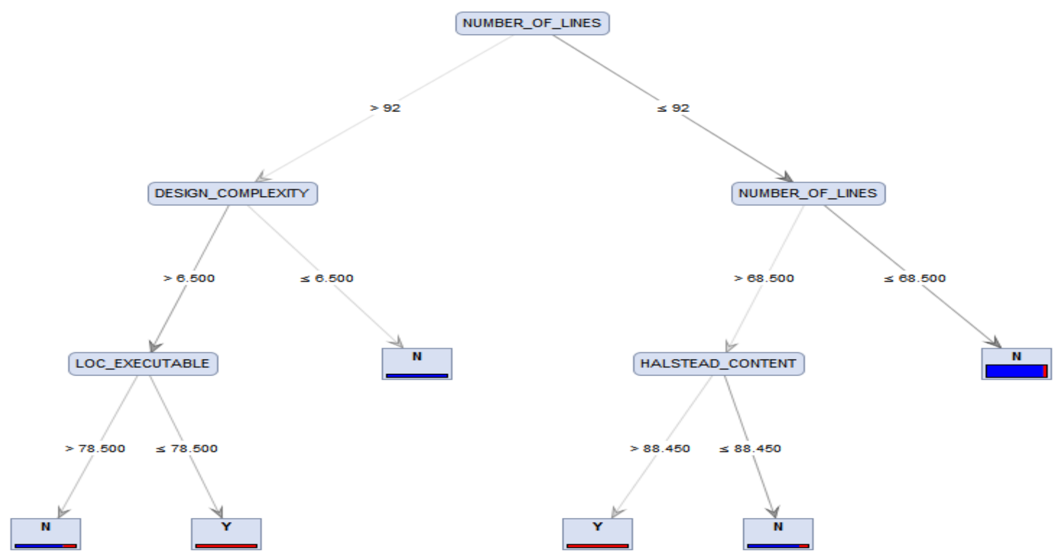




Use Case: Software Fault Prediction

- The cost of capturing and correcting defects is expensive
  - $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 detec software fault predictioı models is higher (71%) t software reviews (60%)

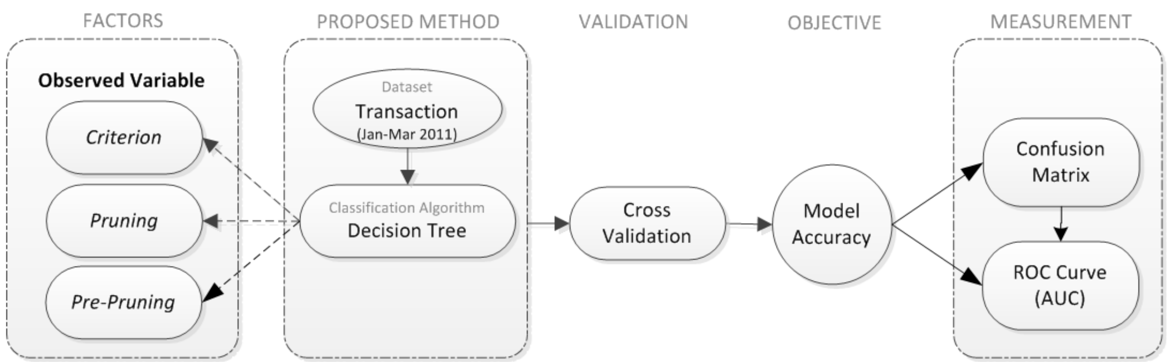


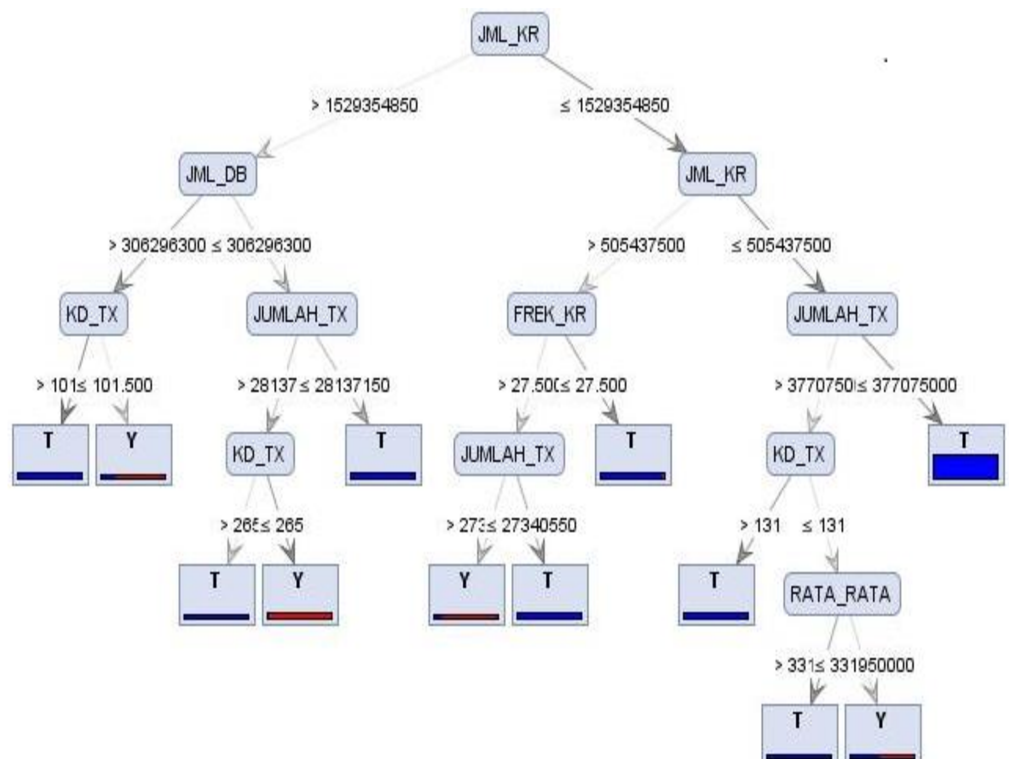


Public and Government Sector

- 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

Use Case: Deteksi Pencucian Uang

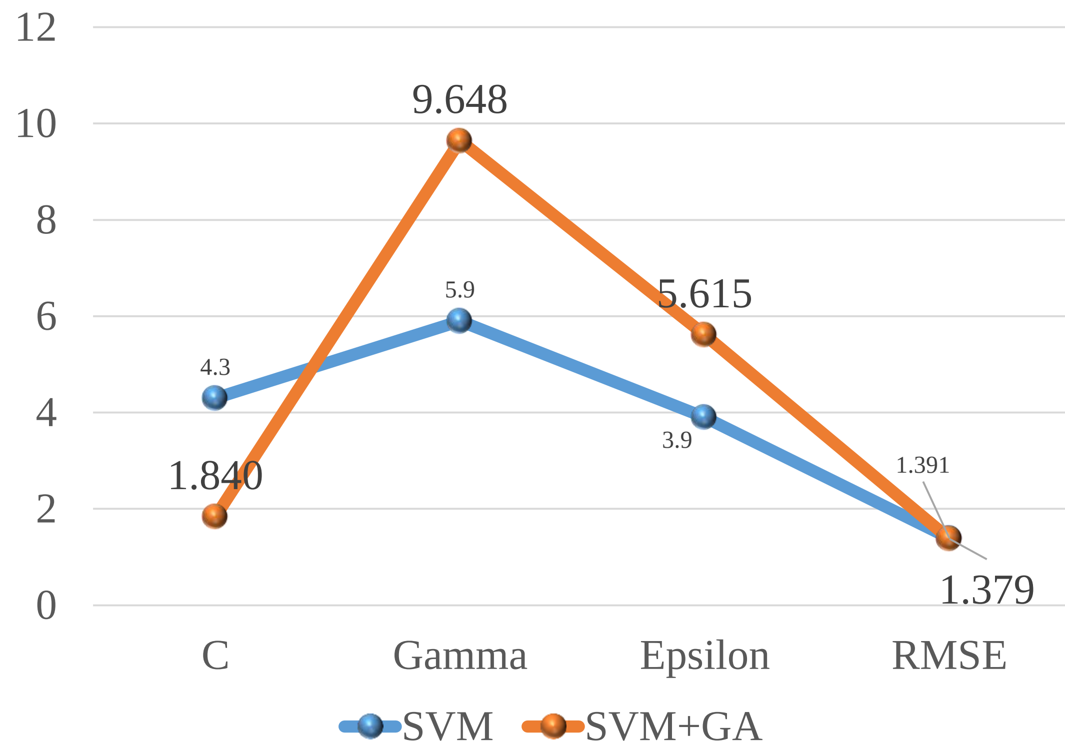




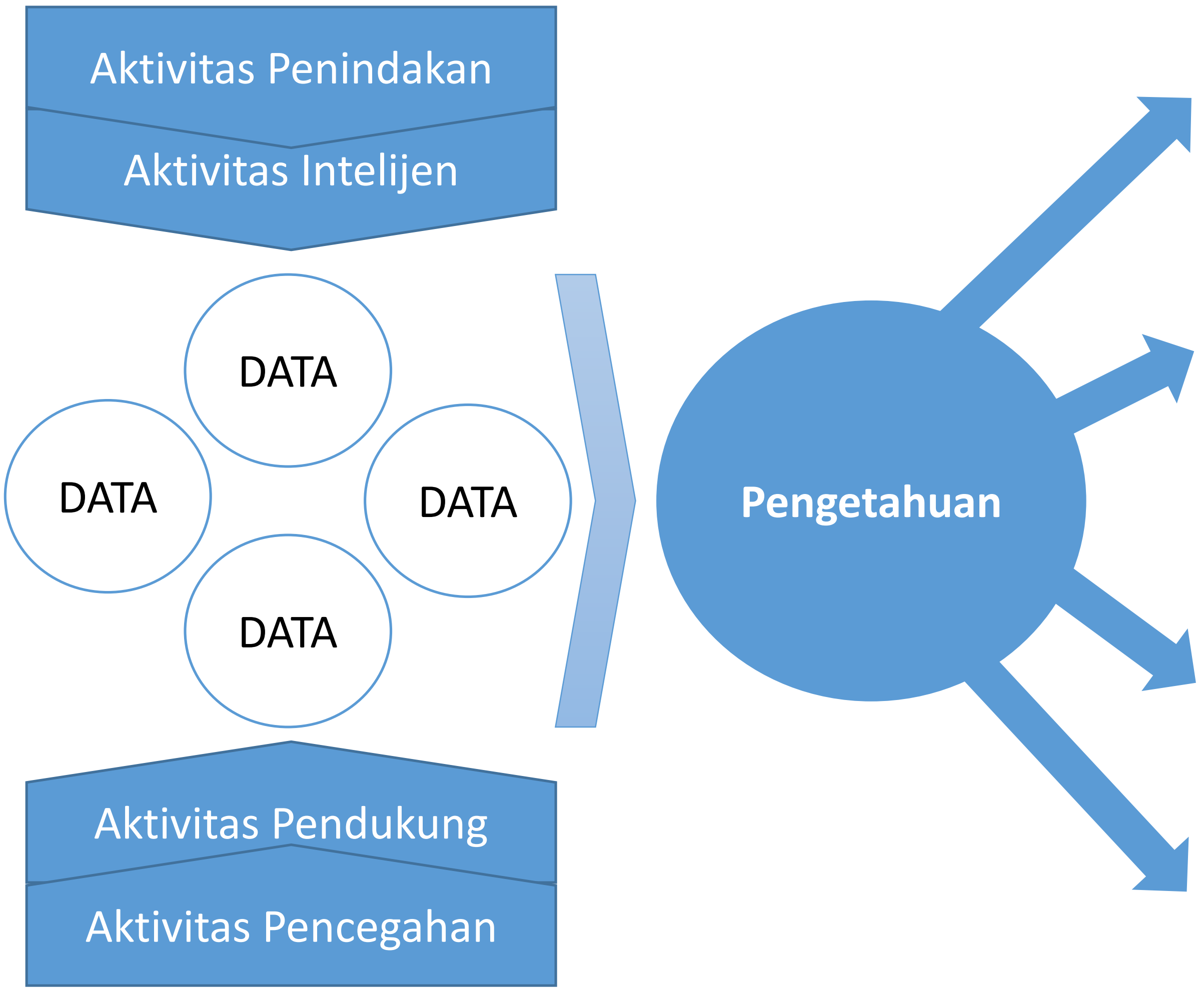
Use Case: Prediksi Kebakaran Hutan

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



Use Case: Prediksi Koruptor



Prediksi dan klastering calon tersangka koruptor

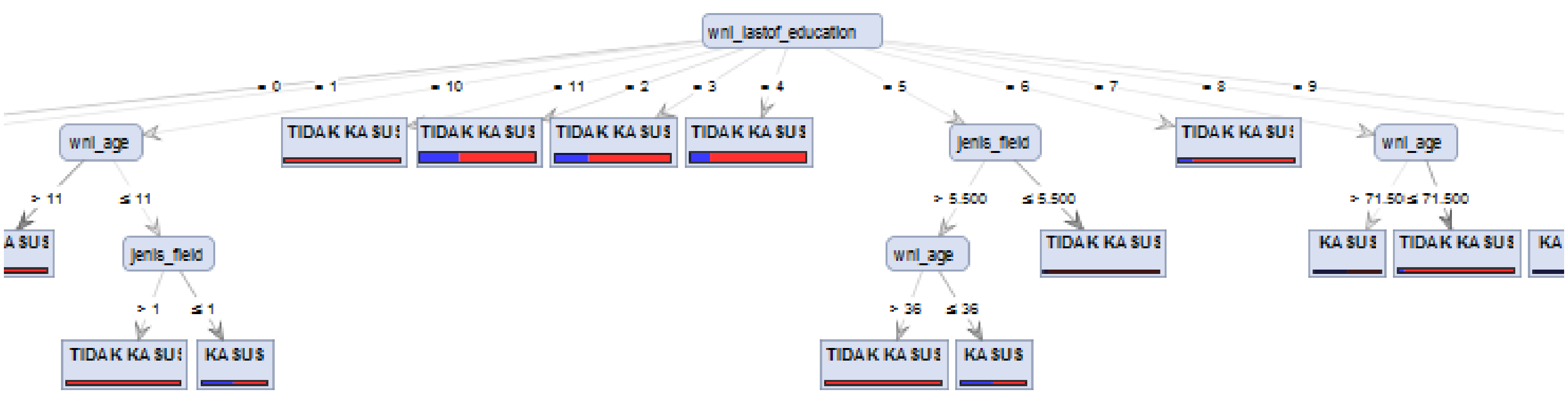
Asosiasi atribut tersangka koruptor

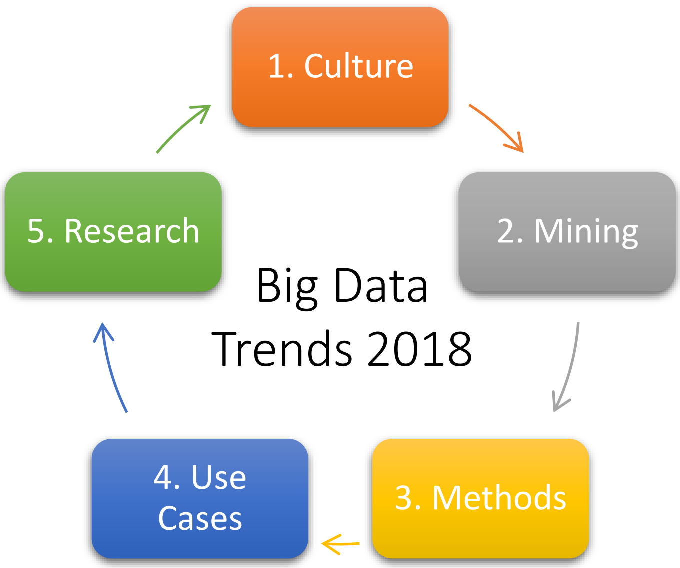
Prediksi pencucian uang

Estimasi jenis dan jumlah tahun hukuman

Use Case: Profiling dan Deteksi Kasus TKI

Row No.	status_kasus	wni_age	wni_lastof_...	gender_name	wni_marital...	wni_local_p...	self_report_...	jenis_field
1	KASUS	-183	3	Perempuan	5	32	PEA	3
2	KASUS	-181	0	Perempuan	5	32	Yordania	6
3	KASUS	-4	0	Perempuan	0	36	RRC	6
4	KASUS	-3	0	Perempuan	0	33	Suriah	6
5	KASUS	-1	0	Laki-laki	0	12	Libya	2
6	KASUS	-1	0	Perempuan	0	32	Libanon	6
7	KASUS	0	0	Laki-laki	0	11	Jepang	2
8	KASUS	0	0	Laki-laki	0	11	Jepang	5
9	KASUS	0	0	Laki-laki	0	11	Libya	2
10	KASUS	0	0	Laki-laki	0	11	Malaysia	3
11	KASUS	0	0	Laki-laki	0	11	Malaysia	6
12	KASUS	0	0	Laki-laki	0	11	Yaman	2
13	KASUS	0	0	Laki-laki	0	12	Amerika Seri...	3





5. Big Data Research

Komparasi Penelitian D3/D4 vs S1 vs S2 vs S3

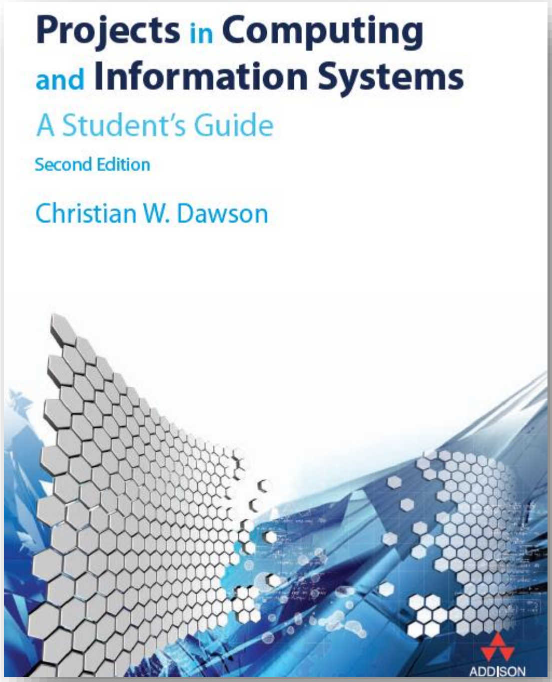
Aspek	Tugas Akhir (D3/D4)	Skripsi (D4/S1)	Tesis (S2)	Disertasi (S3)
Level Kontribusi	Penguasaan Kemampuan Teknis	Pengujian Teori	Pengembangan Teori	Penemuan Teori Baru
Bentuk Kontribusi	Implementasi dan pengembangan	Implementasi dan pengembangan	Perbaikan Secara Inkremental dan Terus Menerus	Substansial dan Invention
Target Publikasi	-	Domestic Conference	International Conference	International Journal

Komparasi Penelitian D3/D4 vs S1 vs S2 vs S3

- D3/D4:
  - Pengembangan Sistem Informasi Rumah Sakit untuk Rumah Sakit "Suka Sembuh"
  - *Karakter: menguasai skill teknis*
- S1:
  - Sistem Cerdas Berbasis Neural Network untuk Prediksi Harga Saham
  - *Karakter: menguji teori, ada software development*
- S2/S3:
  - Penerapan Algoritma Genetika untuk Pemilihan Arsitektur Jaringan Secara Otomatis pada Neural Network untuk Prediksi Harga Saham
  - *Karakter: mengembangkan teori (perbaikan metode), ada kontribusi ke teori/metode*

Apa Yang Dikejar di Penelitian?

Research is a considered activity, which aims to make an original contribution to knowledge *(Dawson, 2009)*



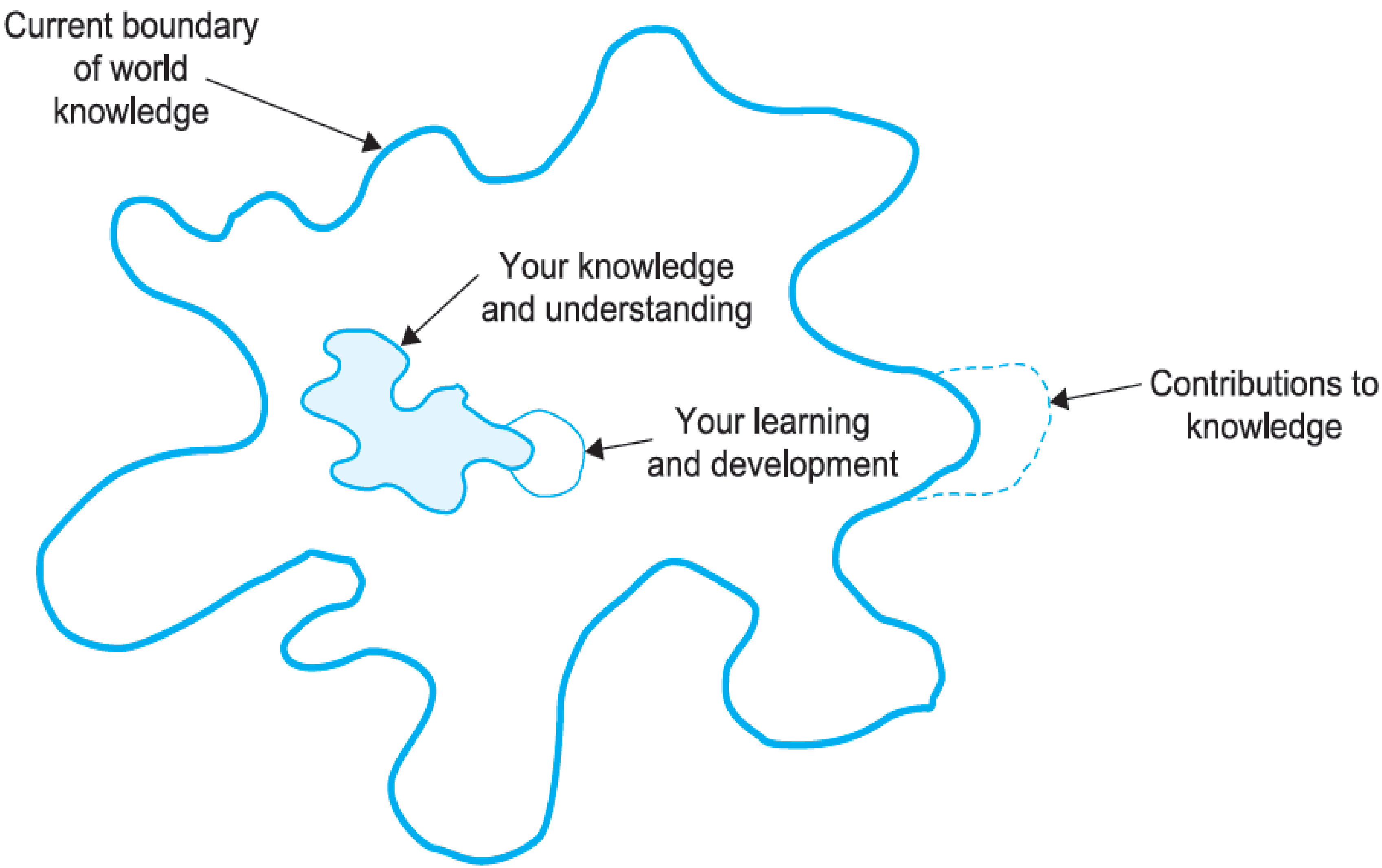
- Original Contribution: Kontribusi Orisinil
- To Knowlegde: Untuk Pengetahuan

Bentuk Kontribusi ke Pengetahuan

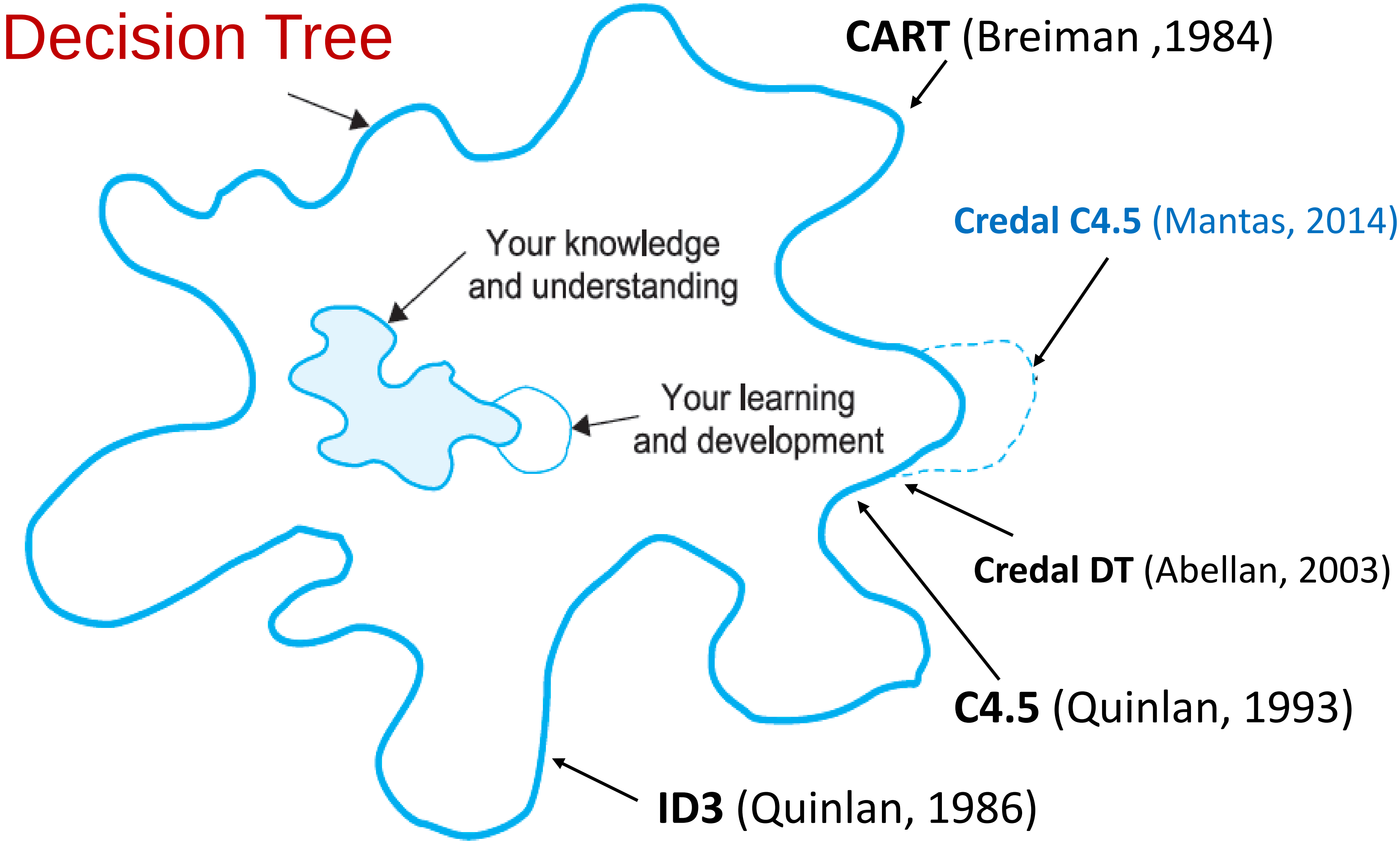
Kegiatan penyelidikan dan investigasi terhadap suatu masalah yang dilakukan secara berulang-ulang dan sistematis, dengan tujuan untuk menemukan atau merevisi teori, fakta, dan aplikasi

(Berndtsson et al., 2008)

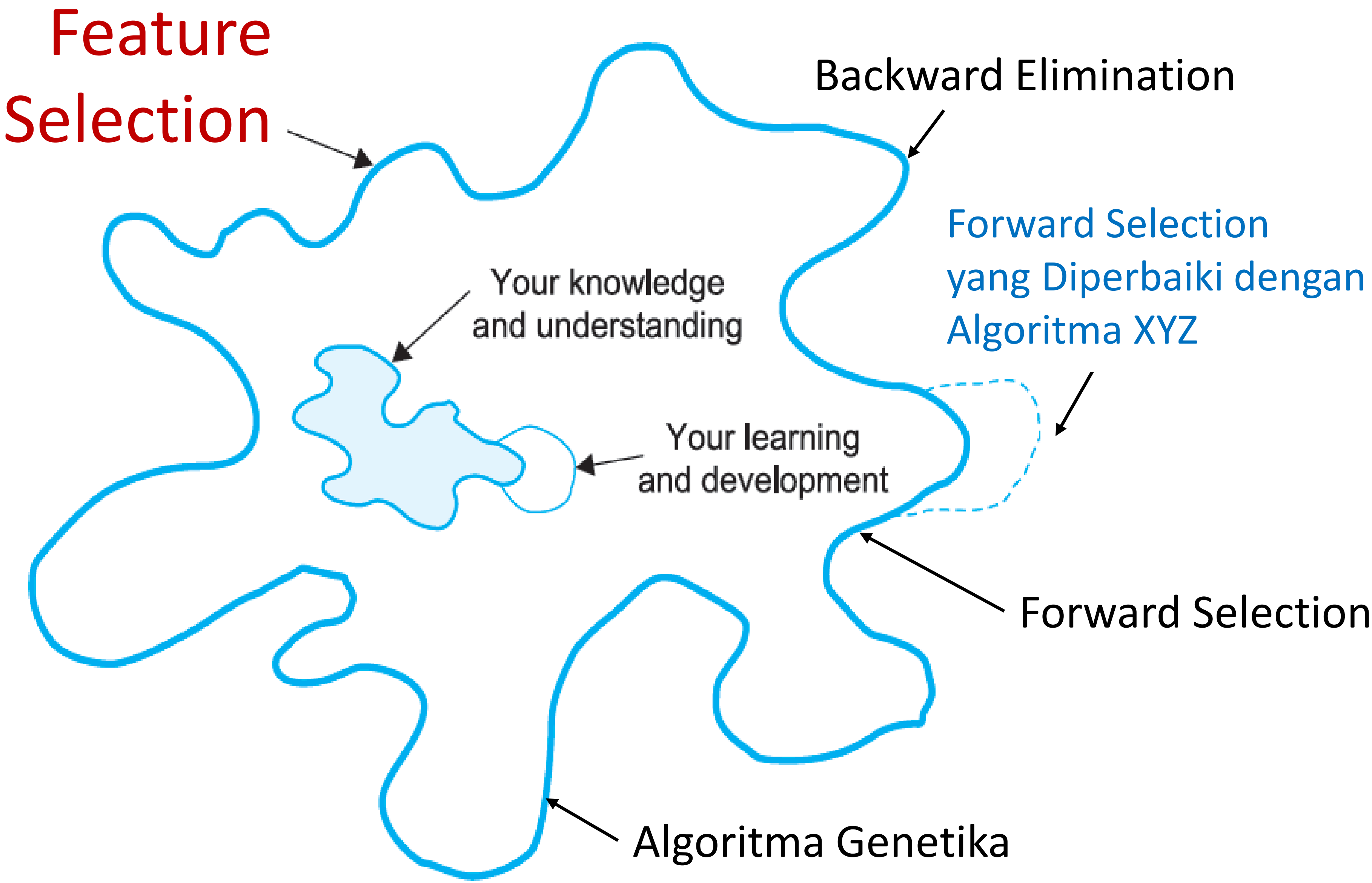
Kontribusi ke Pengetahuan



Contoh Kontribusi ke Pengetahuan



Contoh Kontribusi ke Pengetahuan



Penelitian Yang Memiliki Kontribusi?

- Penerapan Neural Network untuk Prediksi Harga Saham pada Perusahaan ABC
  - Pemilihan Arsitektur Jaringan pada Neural Network Secara Otomatis dengan Menggunakan Algoritma Semut untuk Prediksi Harga Saham

- Penerapan algoritma C4.5 untuk penentuan kelulusan mahasiswa tepat waktu: Studi Kasus STMIK XYZ
  - Modifikasi Penghitungan Gain dan Entropi untuk Peningkatan Akurasi pada Algoritma C4.5

Hanya penelitian dengan kontribusi ke pengetahuan yang bisa menembus jurnal-jurnal internasional terindeks

Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 35 (2008) 1817–1824
Expert Systems with Applications
www.elsevier.com/locate/eswa

Particle swarm optimization for parameter determination and feature selection of support vector machines

Shih-Wei Lin [a,*], Kuo-Ching Ying [b], Shih-Chieh Chen [c], Zne-Jung Lee [a]

[a] Information Management, Huafan University, Taiwan
[b] Industrial Engineering and Management Information, Huafan University, Taiwan
[c] Industrial Management, National Taiwan University of Science and Technology, Taiwan

Abstract

Support vector machine (SVM)
the SVM training procedure, alon
neously determines the parameter v
swarm optimization (PSO) based
developed.
Several public datasets are emp
approach. The developed approac
and other approaches. Experiment
of grid search and many other app
the PSO + SVM approach is valu

Contents lists available at ScienceDirect
Expert Systems with Applications
journal homepage: www.elsevier.com/locate/eswa

Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data

Carlos J. Mantas, Joaquín Abellán
Department of Computer Science & Artificial Intelligence,

ARTICLE INFO

Keywords:
Imprecise probabilities
Imprecise Dirichlet Model
Uncertainty measures
Credal decision trees
C4.5 algorithm
Noisy data

Contents lists available at ScienceDirect
Expert Systems with Applications
journal homepage: www.elsevier.com/locate/eswa

An extended support vector machine forecasting framework for customer churn in e-commerce

Xiaobing Yu [a,*], Shunsheng Guo [b], Jun Guo [b], Xiaorong Huang [b]

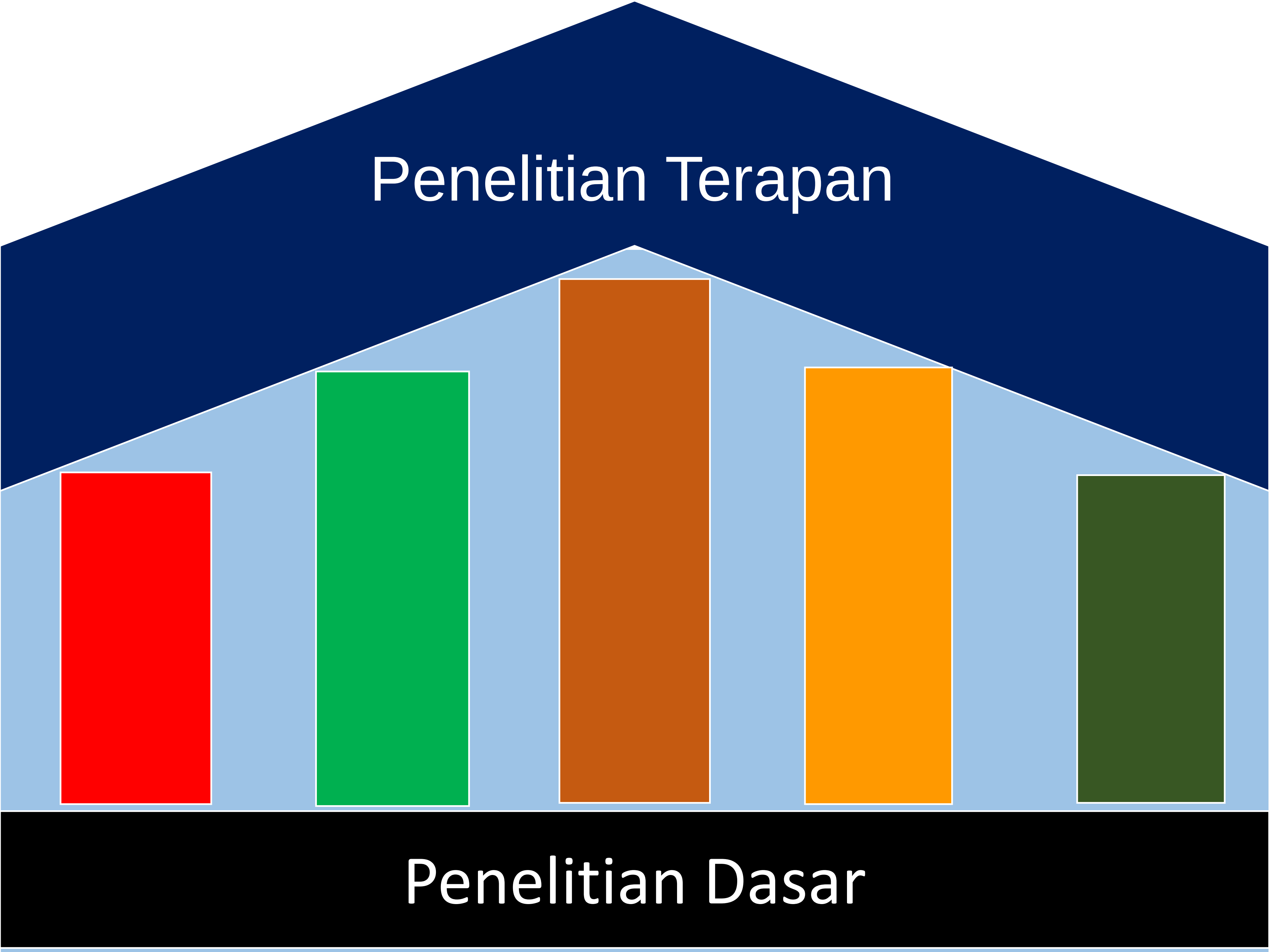
[a] School of Economics and Management, Nanjing University of Information Science & Technology, Nanjing 210044, PR China
[b] School of Mechanic and Electronic Engineering, Wuhan University of Technology, Wuhan 430070, PR China

ARTICLE INFO

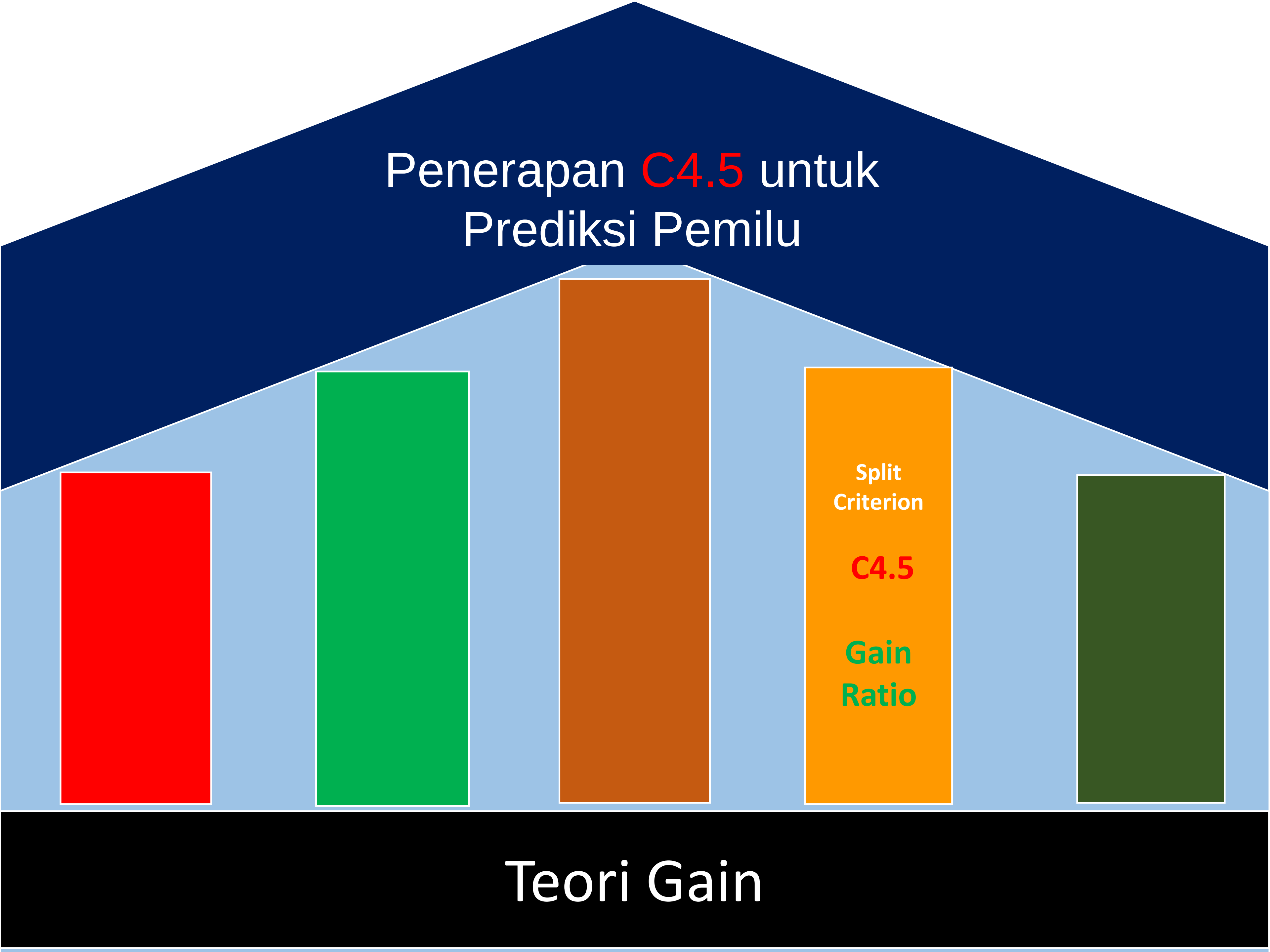
Keywords:
Customer churn
Support vector machine
E-commerce
Kernel function

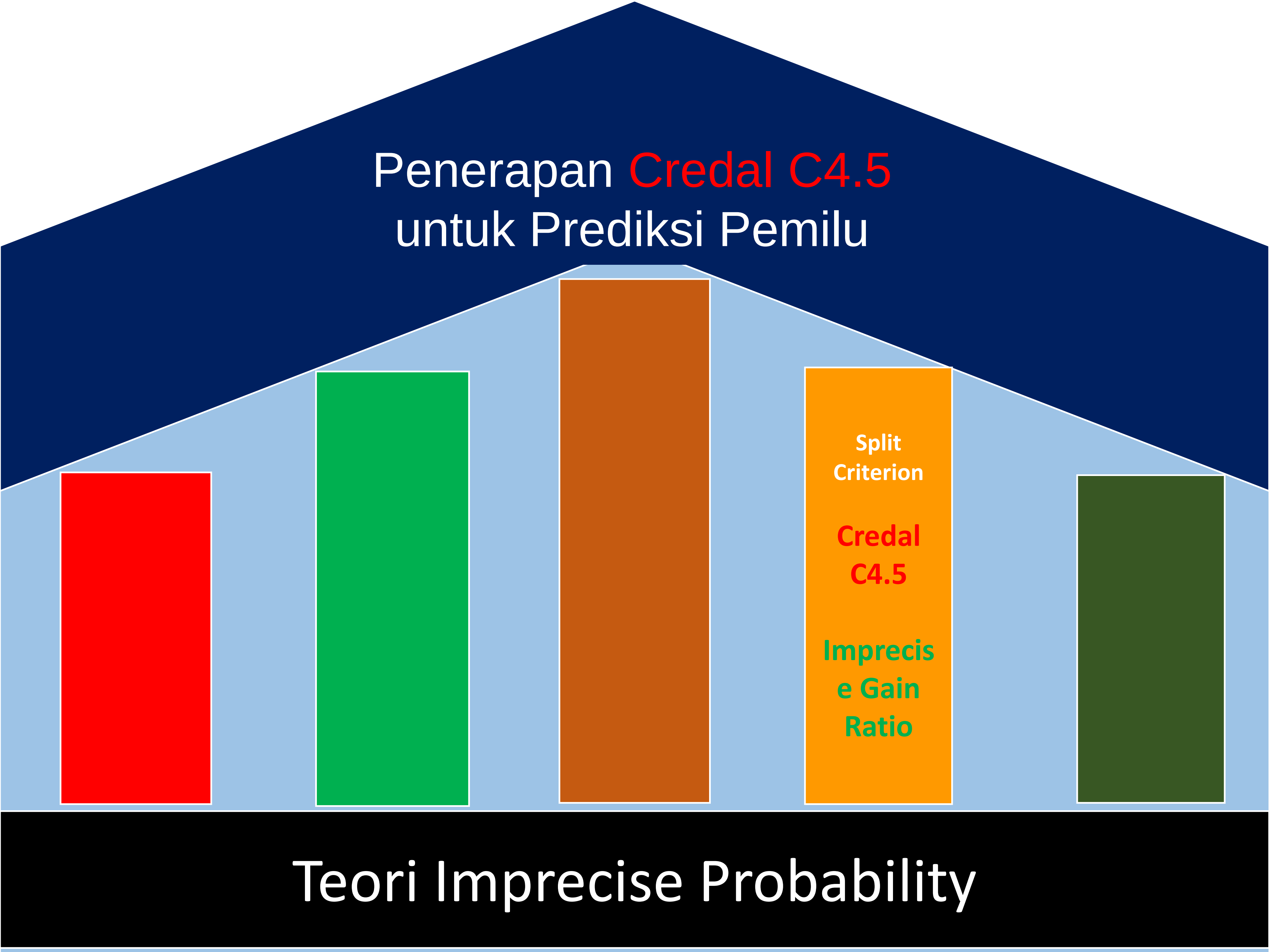
ABSTRACT

In order to accurately forecast and prevent customer churn in e-commerce, a customer churn forecasting framework is established through four steps. First, customer behavior data is collected and converted into data warehouse by extract transform load (ETL). Second, the subject of data warehouse is established and some samples are extracted as train objects. Third, alternative predication algorithms are chosen to train selected samples. Finally, selected predication algorithm with extension is used to forecast other customers. For the imbalance and nonlinear of customer churn, an extended support vector machine (ESVM) is proposed by introducing parameters to tell the impact of churner, non-churner and nonlinear. Artificial neural network (ANN), decision tree, SVM and ESVM are considered as alternative predication algorithms to forecast customer churn with the innovative framework. Result shows that ESVM performs best among them in the aspect of accuracy, hit rate, coverage rate, lift coefficient and treatment time. This novel ESVM can process large scale and imbalanced data effectively based on the framework.

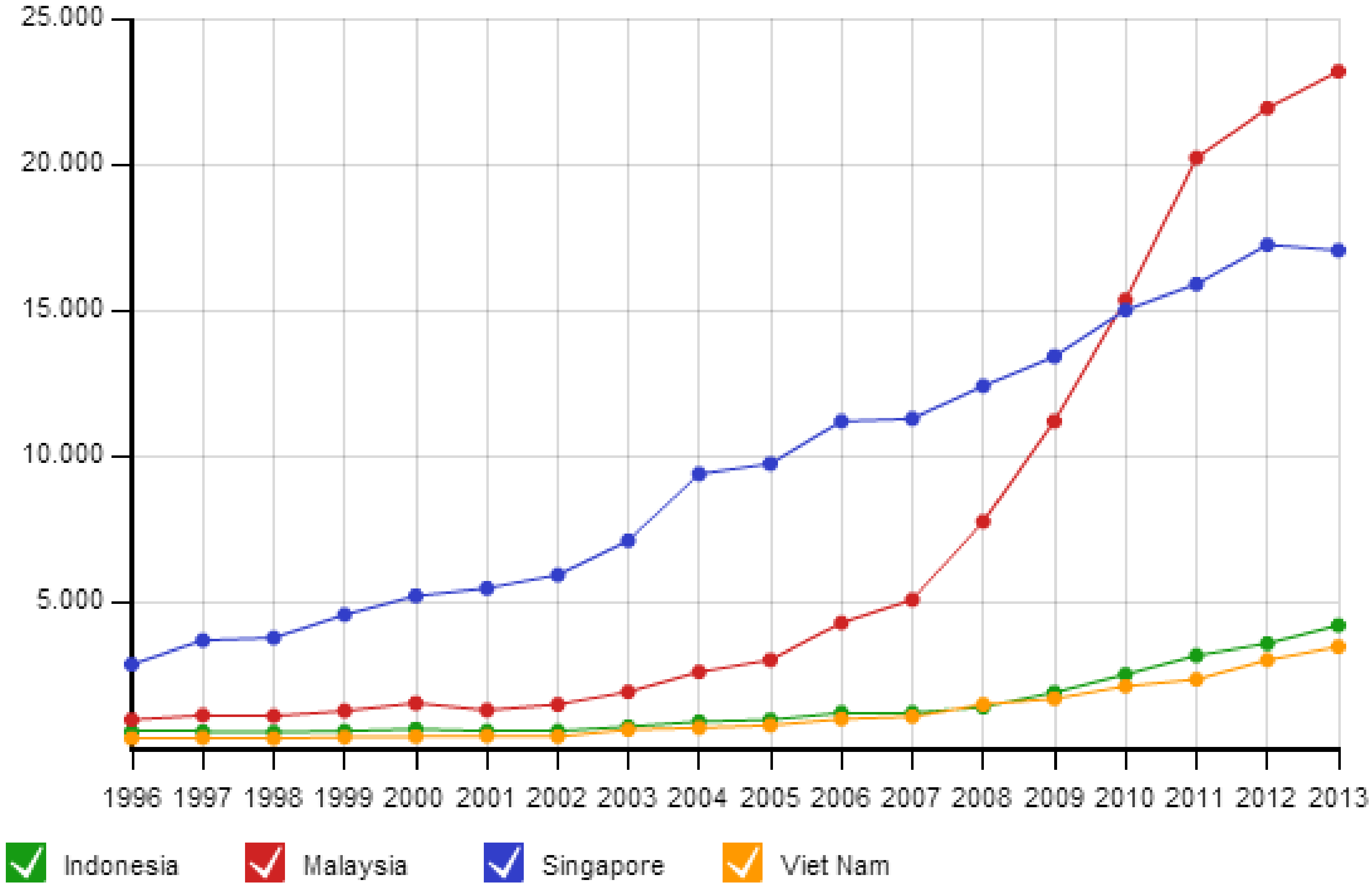


Penelitian Terapan
Penelitian Dasar


Penerapan C4.5 untuk
Prediksi Pemilu
Split
Criterion
C4.5
Gain
Ratio
Teori Gain


Penerapan Credal C4.5
untuk Prediksi Pemilu
Split Criterion
Credal C4.5
Imprecise Gain Ratio
Teori Imprecise Probability

Statistik Jumlah Publikasi (ScimagoJR.Com)



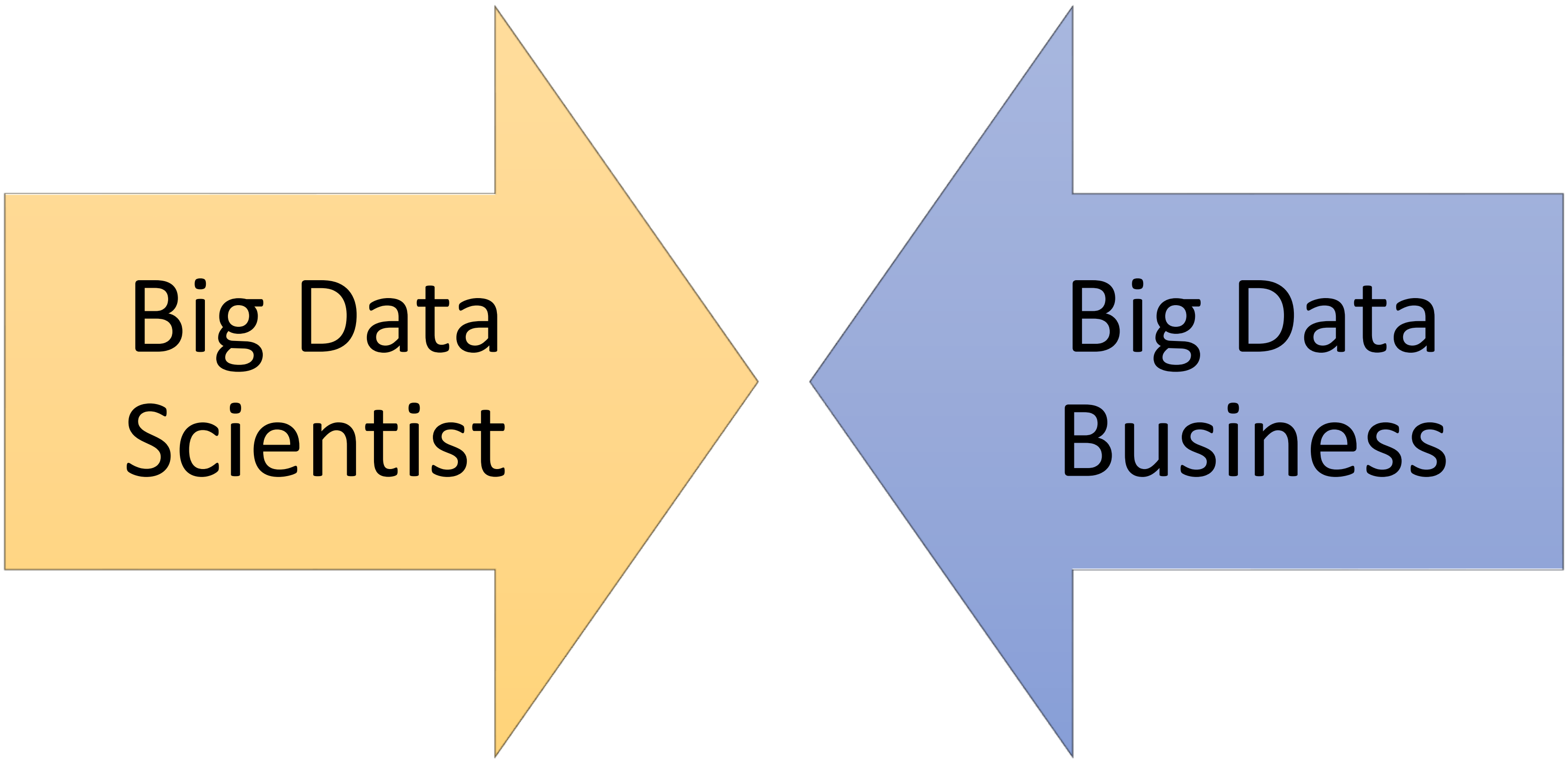
Rangking Publikasi Ilmiah (ScimagoJR.Com)

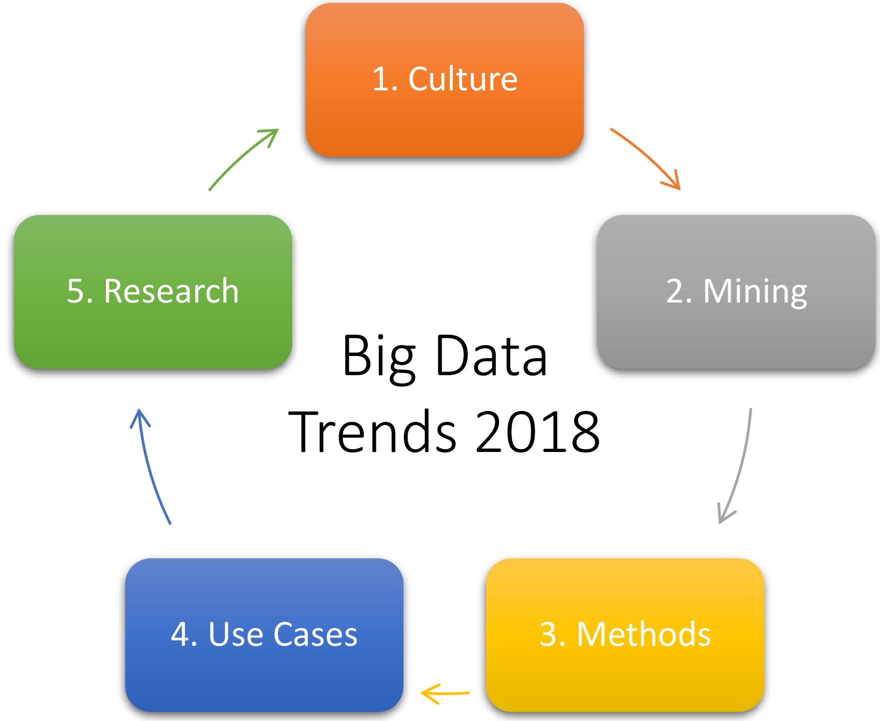
	Country	Documents	Citable documents	Citations	Self-Citations	Citations per Document	H index
1	United States	7.063.329	6.672.307	129.540.193	62.480.425	20,45	1.380
2	China	2.680.395	2.655.272	11.253.119	6.127.507	6,17	385
3	United Kingdom	1.918.650	1.763.766	31.393.290	7.513.112	18,29	851
4	Germany	1.782.920	1.704.566	25.848.738	6.852.785	16,16	740
5	Japan	1.776.473	1.734.289	20.347.377	6.073.934	12,11	635
6	France	1.283.370	1.229.376	17.870.597	4.151.730	15,60	681
7	Canada	993.461	946.493	15.696.168	3.050.504	18,50	658
8	Italy	959.688	909.701	12.719.572	2.976.533	15,26	588
9	Spain	759.811	715.452	8.688.942	2.212.008	13,89	476
10	India	750.777	716.232	4.528.302	1.585.248	7,99	301
11	Australia	683.585	643.028	9.338.061	2.016.394	16,73	514
12	Russian Federation	586.646	579.814	3.132.050	938.471	5,52	325
13	South Korea	578.625	566.953	4.640.390	1.067.252	10,55	333
14	Netherlands	547.634	519.258	10.050.413	1.701.502	21,25	576
15	Brazil	461.118	446.892	3.362.480	1.151.280	10,09	305
16	Taiwan	398.720	389.411	3.259.864	790.103	10,41	267
17	Switzerland	395.703	377.016	7.714.443	1.077.442	22,69	569
18	Sweden	375.891	361.569	6.810.427	1.104.677	20,11	511
19	Poland	346.611	339.712	2.441.439	652.956	8,25	302
20	Turkey	306.926	291.814	1.935.431	519.675	8,24	210
21	Belgium	299.077	285.735	4.696.153	701.283	18,16	454
22	Israel	224.674	215.590	3.663.004	530.340	17,78	414

Rangking Publikasi Ilmiah (ScimagoJR.Com)

49	Slovenia	50.565	49.471	403.209	83.402	9,53	153
50	Bulgaria	45.348	44.609	319.449	56.183	7,80	138
51	Nigeria	40.952	40.124	174.002	42.457	6,23	89
52	Tunisia	38.334	36.859	169.981	39.062	6,77	85
53	Colombia	35.890	34.768	228.686	36.843	10,61	133
54	Serbia	28.882	28.312	81.010	23.288	8,75	68
55	Morocco	27.253	26.175	157.219	29.432	7,11	99
56	Venezuela	27.138	26.445	204.691	29.729	8,42	130
57	Algeria	25.714	25.387	105.945	20.698	6,49	78
58	Belarus	24.801	24.466	122.850	24.438	5,08	106
59	Lithuania	24.755	24.434	151.748	37.377	8,61	109
60	Cuba	24.606	23.847	123.183	28.193	5,81	93
61	Indonesia	20.166	19.740	146.670	16.149	10,94	112
62	Jordan	19.847	19.507	107.550	15.257	7,24	82
63	Bangladesh	19.481	19.037	115.329	22.662	8,37	97
64	Estonia	19.141	18.774	204.306	38.547	13,58	130
65	United Arab Emirates	19.051	18.331	100.247	11.207	7,56	87
66	Kenya	16.727	16.044	206.886	34.874	15,09	131
67	Viet Nam	16.474	16.116	125.927	18.500	11,79	107
68	Kuwait	13.775	13.425	93.290	12.879	7,67	83
69	Lebanon	13.677	12.847	97.316	10.182	9,70	97
70	Philippines	13.163	12.796	141.070	15.727	13,38	116
71	Puerto Rico	11.209	10.953	150.252	11.819	15,34	129

Scientist vs Businessman





Romi Satria Wahono
http://romisatriawahono.net
romi@romisatriawahono.net
081586220090