Kuliah 13

Text & Web Mining

Data Mining – Ilmu Komputer IPB

Data terstruktur

- Sejauh ini kita berurusan dengan data terstruktur,

Attribute → Value	Outlook → Sunny
Attribute → Value	Temperature → Hot
Attribute → Value	Windy → Yes
…	Humidity → High
Attribute → Value	Play → Yes

- Umumnya data mining menggunakan data semacam ini

Tipe data yang kompleks

- Berkembangnya data kompleks
  - **Data spasial**: daat geografik, citra medis dan satelit
  - **Data Multimedia** : gambar, audio, dan video
  - **Data Time-series** : data perbankan & data saham
  - **Text data**: kata yang mendeskripsikan objek
  - **World-Wide-Web**: teks yang sangat tidak terstruktur dan data multimedia

Fokus

Basisdata Teks

- Dalam prakteknya terdapat banyak basis data teks:
  - artikel berita
  - paper riset
  - buku
  - perpustakaan digital
  - e-mail
  - halaman web

- Berkembang dengan cepat baik dari segi jumlah maupun kepentingan (80%)

Text Mining

- Text mining merujuk pada data mining yang menggunakan dokumen teks sebagai data
- Hampir semua tugas Text Mining menggunakan metode Information Retrieval (IR) untuk pra-proses dokumen teks.
- Metode ini sedikit berbeda daripada metode pra-proses data yang digunakan dalam tabel relasional
- Web search juga berakar pada IR

Definisi Text Mining

- Discover useful and previously unknown "gems" of information in **large text collections**

Patterns

Trends

Associations

Definisi Text Mining

Text Mining adalah proses yang secara otomatis mengekstrak informasi komprehensif yang bermakna, berguna dan belum diketahui sebelumnya dari repositori dokumen tekstual

Text Mining
=
Data Mining (diaplikasinya ke data text)
+
basic linguistics

Definisi

- “yang tidak diketahui sebelumnya” ?
  - Definisi ketat
    - Informasi yang bahkan penulisnya tidak mengetahui
      - Contoh: menemukan metode baru untuk pertumbuhan rambut yang merupakan efek samping dari suatu prosedur
  - Definisi longgar
    - Menemukan kembali informasi yang telah ditulis pengarang dalam teksnya
      - Contoh: secara otomatis mengekstrak nama produk dari sebuah halaman web

Text Mining Tasks

- **Diberikan:**
  - Sumber dokumen tekstual
  - Kueri terbatas (berbasis teks) yang didefinisikan dengan baik
- **Temukan:**
  - Kalimat dengan informasi relevan
  - Ekstrak informasi relevan & abaikan informasi yang tidak relevan
  - Hubungkan informasi & keluaran yang saling berhubungan dalam format yang sudah ditetapkan sebelumnya

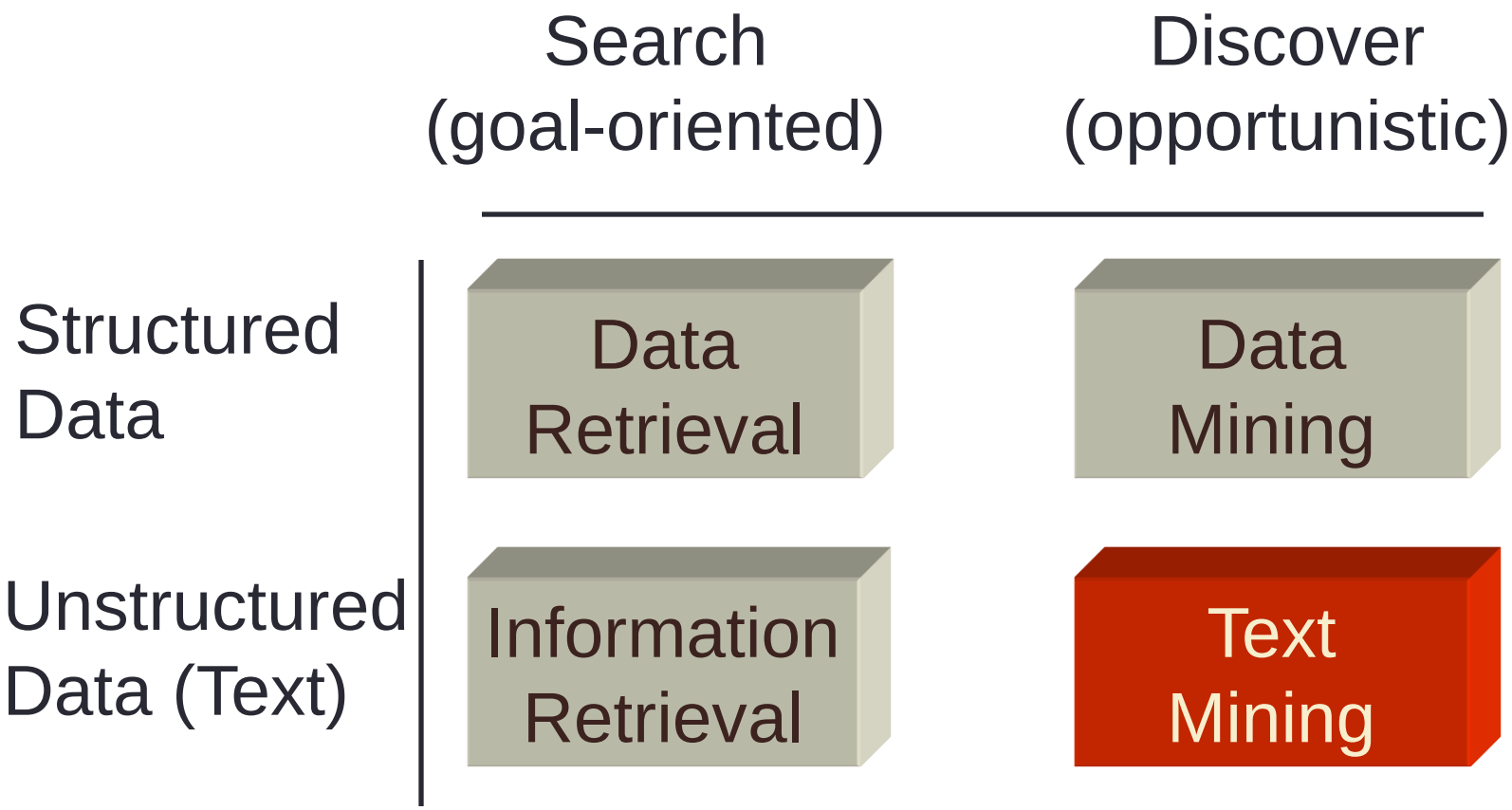
Tasks yang terkait atau dapat diselesaikan dengan text mining

- Search dan retrieval
- Analisis semantic
- Clustering
- Kategorisasi
- Feature extraction
- Pembuatan Ontology
- Dynamic focusing

DM vs TM

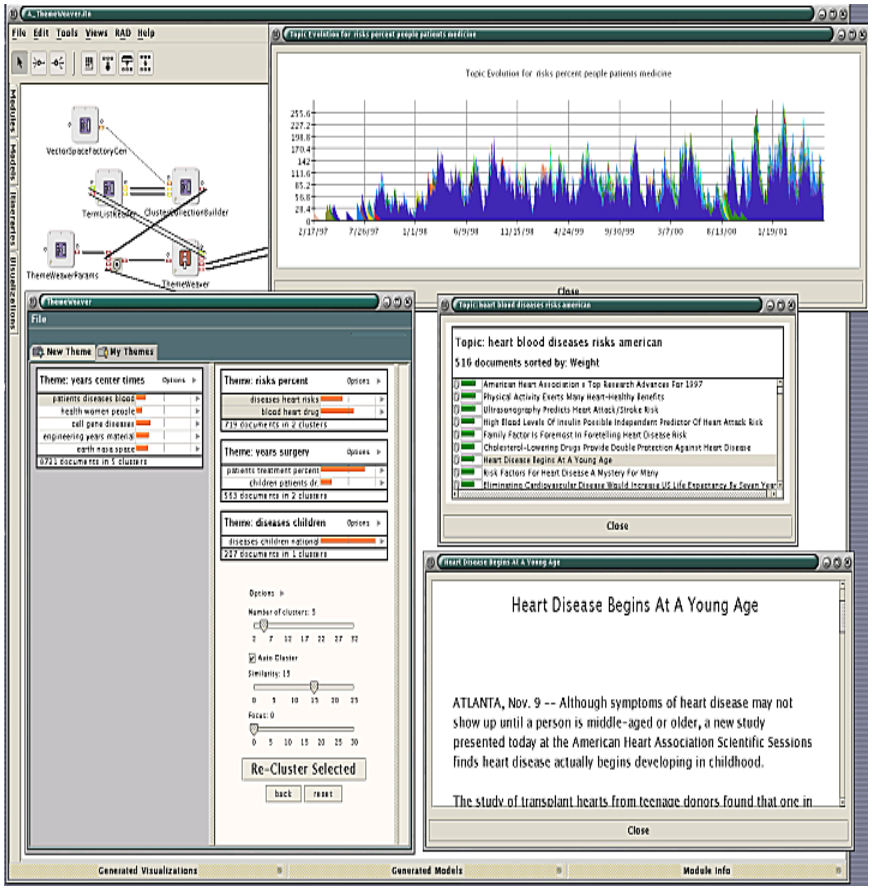
	Data Mining	Text Mining
Objek yang diteliti	Data numerik dan kategorikal	Teks
Struktur objek	Basis data relasional	Free form texts
Tujuan	Memprediksi outcome dari situasi mendatang	Temu kembali informasi yang relevan, distill the meaning, menkategorikan dan penyampaian target
Metode	Machine learning: Symbolic Knowledge Acquisition Technology (SKAT), DT, NN, GA, MBR, MBA	Indexing, pemrosesan neural network khusus, linguistik, ontologies
Ukuran market saat ini	100,000 analysts pada perusahaan besar dan menengah	100,000,000 pekerja korporat dan pengguna individual
Kematangan	Implementasi yang luas sejak 1994	Implementasi yang luas mulai 2000

“Search” vs “Discover”



Aplikasi Text Mining

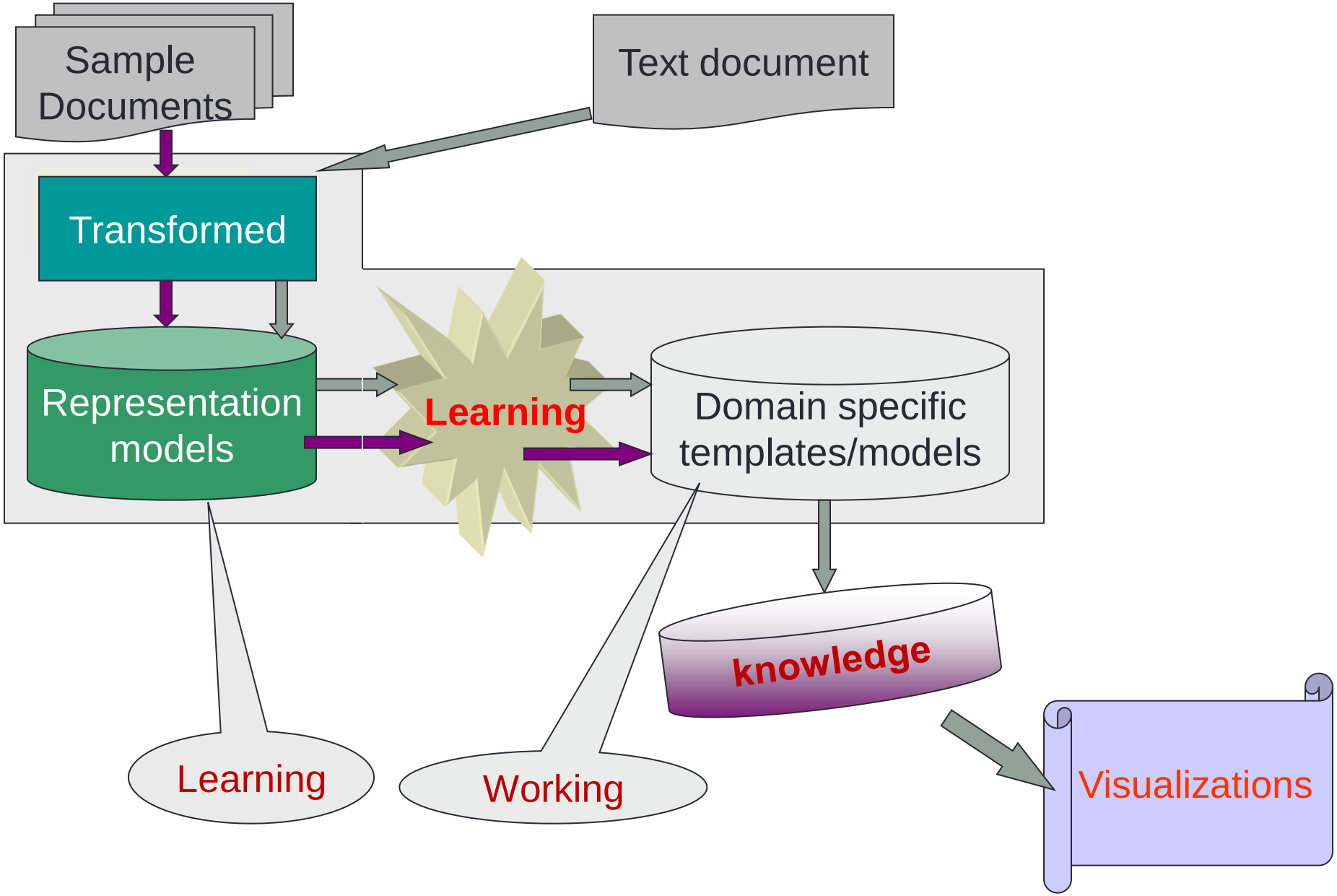
- **Pemasaran:** Menemukan kelompok pembeli yang potensial berdasarkan profil teks pengguna
  - contoh. amazon
- **Industri:** Mengidentifikasi situs web kelompok pesaing
  - Produk pesaing dan harganya
- **Pencarian kerja:** mengidentifikasi parameter dalam pencarian pekerjaan
  - www.flipdog.com



Aplikasi Text Mining

- Search engines
- Enterprise portals
- Knowledge management systems
- e-Business systems
- Vertical applications:
  - Kategorisasi dan routing e-mail
  - Kategorisasi catatan call center
  - Sistem Customer-relationship management (CRM)

Text Mining


Sample Documents
Text document
Transformed
Representation models
Learning
Domain specific templates/models
knowledge
Learning
Working
Visualizations

Karakteristik teks: Outline

- Basis data tekstual berukuran besar
- Berdimensi tinggi
- Beberapa mode input
- Ketergantungan
- Ambiguitas
- Noisy data
- Teks yang tidak terstruktur dengan baik

Karakteristik teks

- Basis data tekstual ukuran besar
  - Pertimbangan efisiensi
    - Lebih dari 2,000,000,000 halaman web
    - Hampir semua publikasi juga tersedia alam bentuk elektronik
- Dimensi tinggi (Sparse input)
  - Pertimbangan setiap word/phrase sebagai sebuah dimensi
- Beberapa mode input
  - Sebagai contoh: Web mining: informasi mengenai pengguna dibangkitkan oleh semantics, mencari pola dan outside knowledge base.

Karakteristik teks

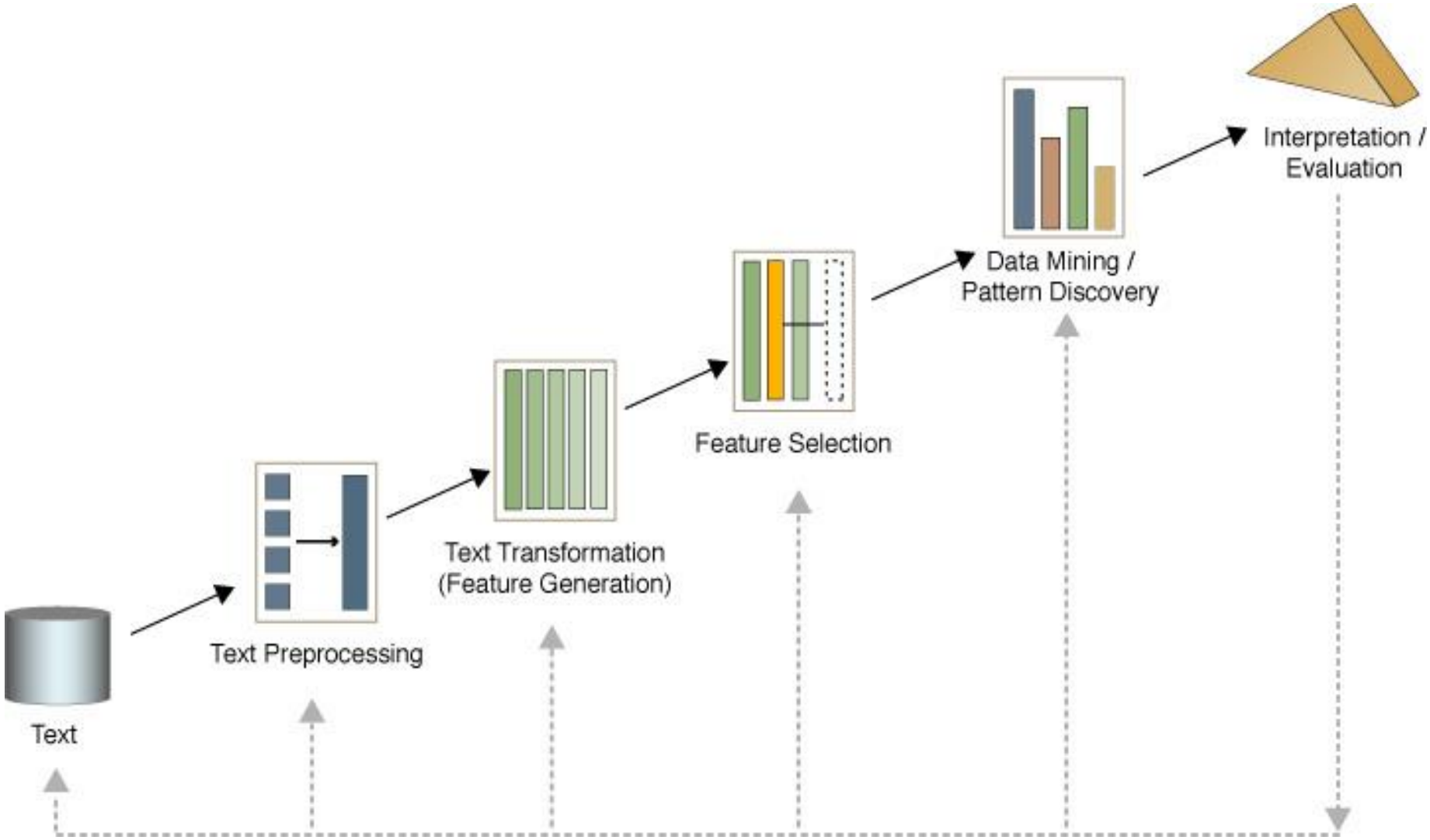
- Dependensi
  - Informasi yang relevan adalah hubungan conjunction yang kompleks dari word/phrase
    - Contoh: Kategorisasi dokumen.

      disambigu pronoun.
- Ambiguitas
  - Ambiguitas kata
    - Pronouns (he, she …)
    - “buy”, “purchase”
  - Ambiguitas semantik
    - The king saw the rabbit with his glasses. (8 makna)

Karakteristik teks

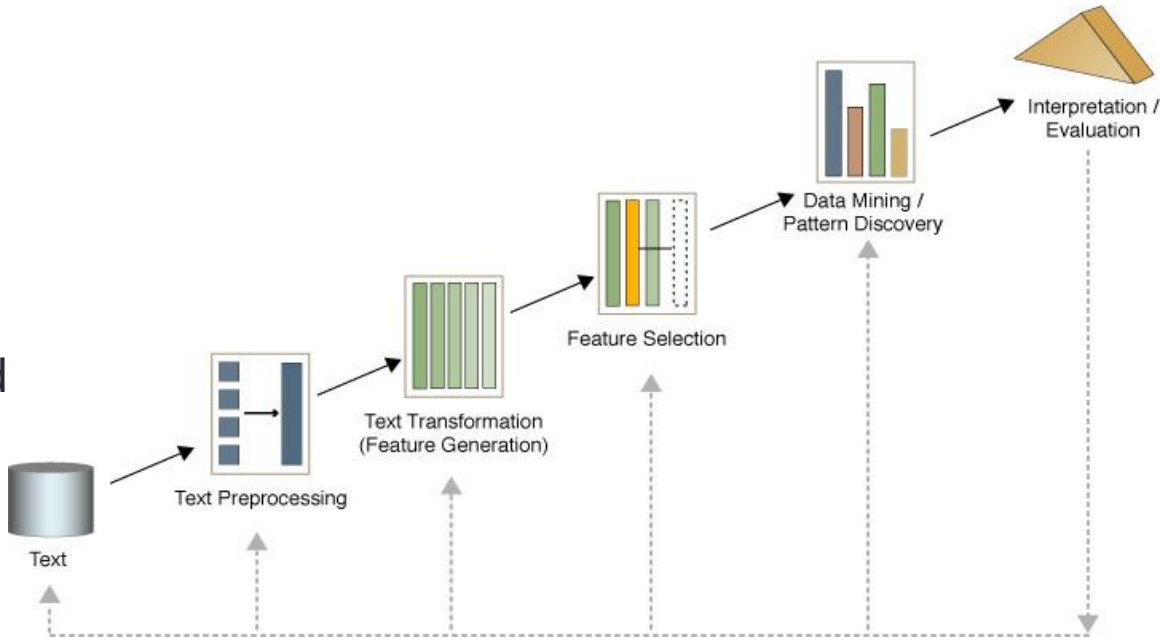
- Data yang mengandung noise
  - Contoh: kesalahan ejaan
- Teks yang tidak terstruktur dengan baik
  - Chat rooms
    - “r u available ?”
    - “Hey whazzzzzz up”
  - Speech

Proses text mining



Proses text mining

- Praproses Text
  - Analisis teks Syntactic/Semantic
- Pembangkitan fitur
  - Bag of words
- Seleksi fitur
  - Simple counting
  - Statistics
- Text/Data Mining
  - Classification- Supervised learning
  - Clustering- Unsupervised learning
- Anlisis hasil



Analisis teks syntactic/semantic

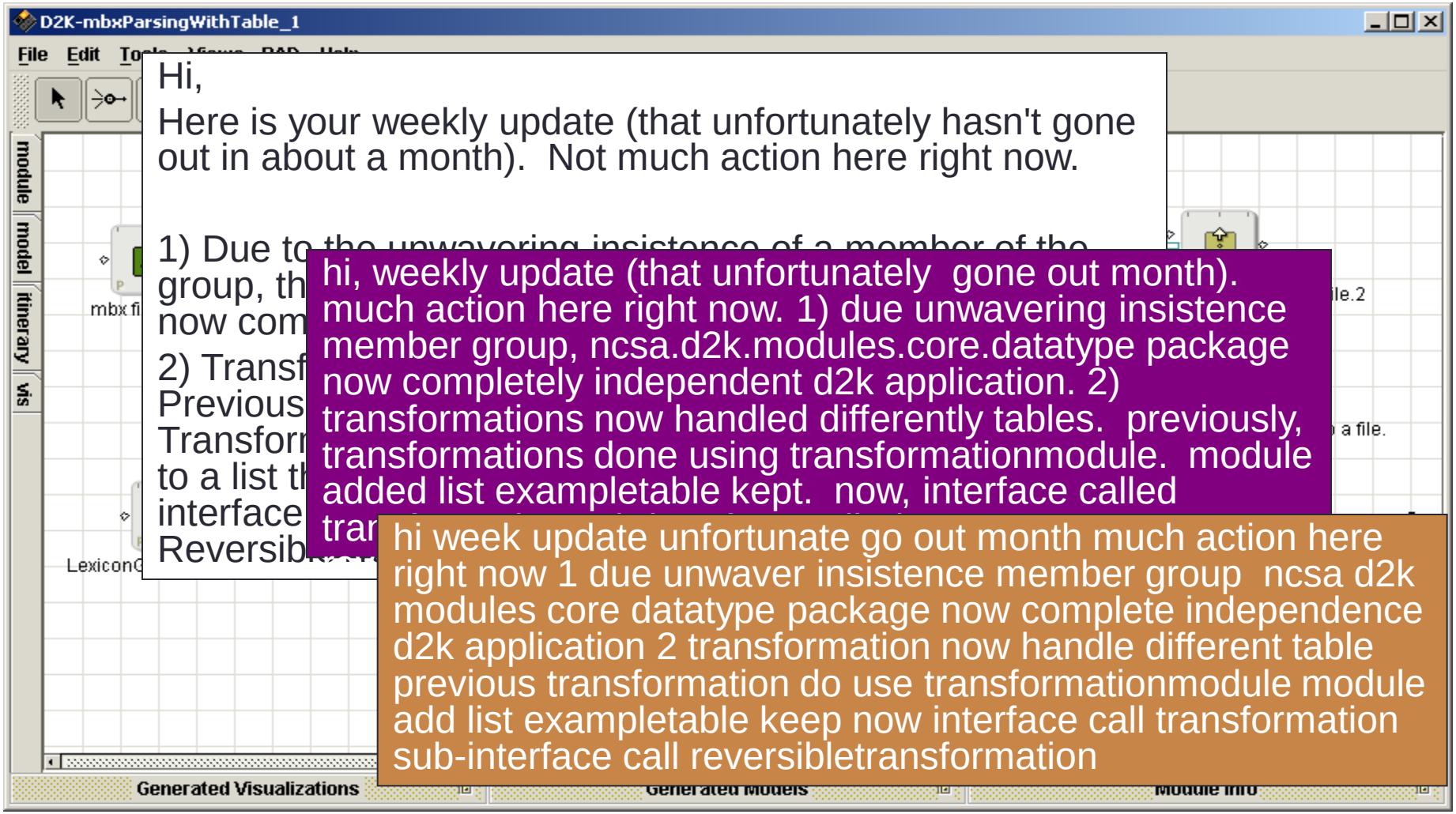
- Part of Speech (pos) tagging
  - Tentukan pos yang sesuai untuk setiap word

    Contoh: John *(noun)* gave *(verb)* the *(det)* ball *(noun)*
  - *~98% accurate.*
- Word sense disambiguation
  - Berbasis konteks atau berbasis proksimity
  - Sangat akurat
- Parsing
  - Membangkitkan parse tree (graf) untuk setiap kalimat
  - Setiap kalimat adalah graf yang berdiri sendiri

Pembangkitan fitur: Bag of words

- Dokumen teks direpresentasikan oleh word yang terdapat di dalamnya (dan kemunculannya)
  - Contoh., “Lord of the rings” $\rightarrow$ {“the”, “Lord”, “rings”, “of”}
  - Sangat efisien
  - Membuat proses learning menjadi lebih sederhana dan mudah
  - Urutan kata tidak begitu penting untuk aplikasi tertentu.
- Stemming: mengidentifikasi kata dengan root-nya
  - contoh., flying, flew $\rightarrow$ fly
  - Mengurangi dimensionalitas
- Stop words: Kata-kata umum yang tidak membantu proses text mining
  - Contoh: “the”, “a”, “an”, “you” …

Pembangkitan fitur: D2K Example



Pembangkitan fitur: XML

- Search engine yang berorientasi pada keyword saat ini tidak dapat menangani kueri seperti ini
  - Find all books authored by “Scooby-Doo”.

- XML: Extensible Markup Language
  - Dokumen XML memiliki struktur bersarang dimana setiap elemen berasosiasi dengan sebuah tag.
  - Tags menjelaskan semantik dari elemen.

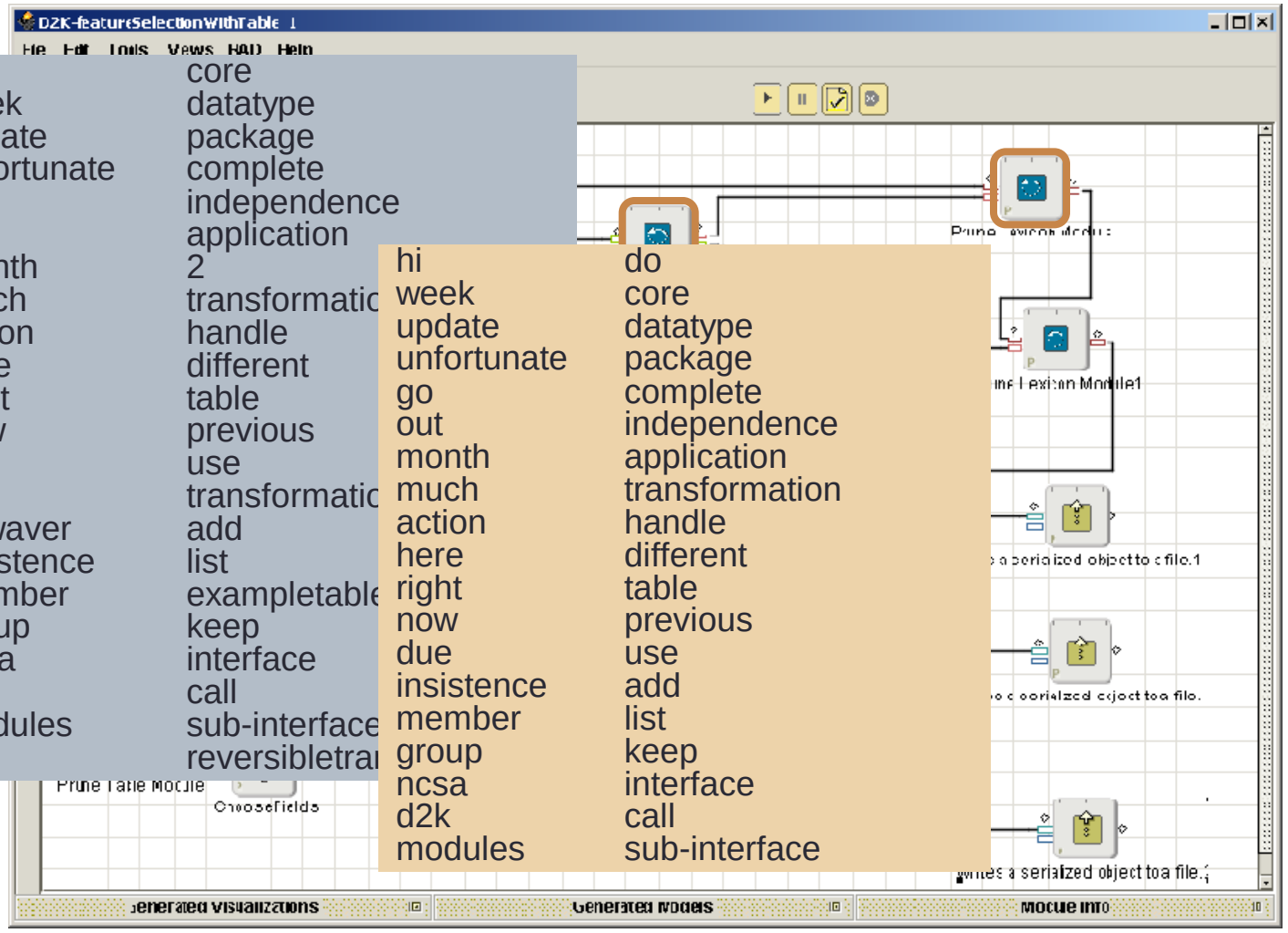
```
<book> <title> The making of a bad movie </title>
        <author> <name> Scooby-Doo </name>
       <affiliation> Cartoons </affiliation> </author>
</book>
```

Seleksi fitur

- Mengurangi dimensi
  - Proses learning merupakan tugas yang sulit pada data dimensi yang tinggi

- Fitur tidak relevan
  - Tidak semua fitur membantu!
    - contoh, the existence of a noun in a news article is unlikely to help classify it as “politics” or “sport”

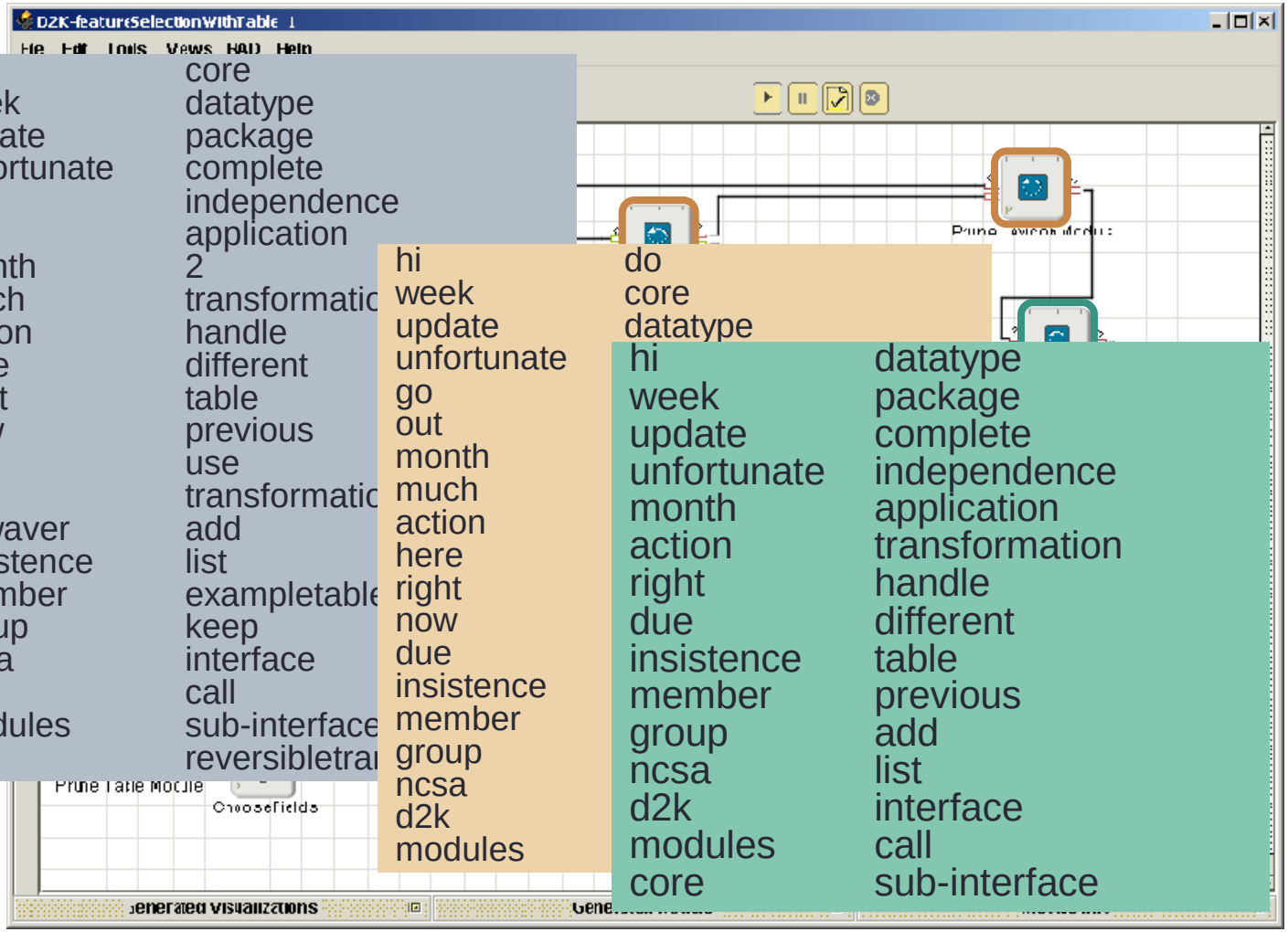
Seleksi fitur: D2K Example I



hi	core
week	datatype
update	package
unfortunate	complete
go	independence
out	application
month	2
much	transformatic
action	handle
here	different
right	table
now	previous
1	use
due	transformatic
unwaver	add
insistence	list
member	exampletable
group	keep
ncsa	interface
d2k	call
modules	sub-interface
do	reversibletra

hi	do
week	core
update	datatype
unfortunate	package
go	complete
out	independence
month	application
much	transformation
action	handle
here	different
right	table
now	previous
due	use
insistence	add
member	list
group	keep
ncsa	interface
d2k	call
modules	sub-interface

Seleksi fitur: D2K Example II



Text Mining: definisi klasifikasi

- **Diberikan:** koleksi dari record dengan label (training set)
  - Setiap record mengandung sekumpulan fitur (atribut), dan kelas true (label)
- **Tentukan:** model untuk kelas sebagai fungsi dari nilai-nilai fitur
- **Tujuan:** record yang belum ada sebelumnya diberi label seakurat mungkin
  - test set digunakan untuk menghitung akurasi model. Biasanya, Data yang diberikan dibagi ke dalam training set dan test set, training set digunakan untuk membangun model dan test set digunakan untuk validasi model

Text Mining: definisi Clustering

- **Diberikan:** sekumpulan dokumen dan similarity measure diantara dokumen
- **Tentukan:** clusters sedemikian sehingga:
  - Dokumen di dalam satu cluster mirip dengan dokumen lainnya.
  - Dokumen dalam cluster yang terpisah tidak mirip dengan dokumen lainnya
- **Tujuan:**
  - Temukan kumpulan dokumeny yang tepat

Similarity Measures:

- Euclidean Distance jika atribut adalah kontinu
- Problem-specific Measure lainnya
  - contoh berapa banyak kata yang umum ada dalam dokumen-dokumen ini.

Contoh

GREAT Camera., Jun 3, 2004

Reviewer: **jprice174** from Atlanta, Ga.

I did a lot of research last year before I bought this camera... It kinda hurt to leave behind my beloved nikon 35mm SLR, but I was going to Italy, and I needed something smaller, and digital.

The pictures coming out of this camera are amazing. The 'auto' feature takes great pictures most of the time. And with digital, you're not wasting film if the picture doesn't come out. …

Ringkasan:

Feature1: **picture**

Positive: 12

- The pictures coming out of this camera are amazing.
- Overall this is a good camera with a really good picture clarity.

…

Negative: 2

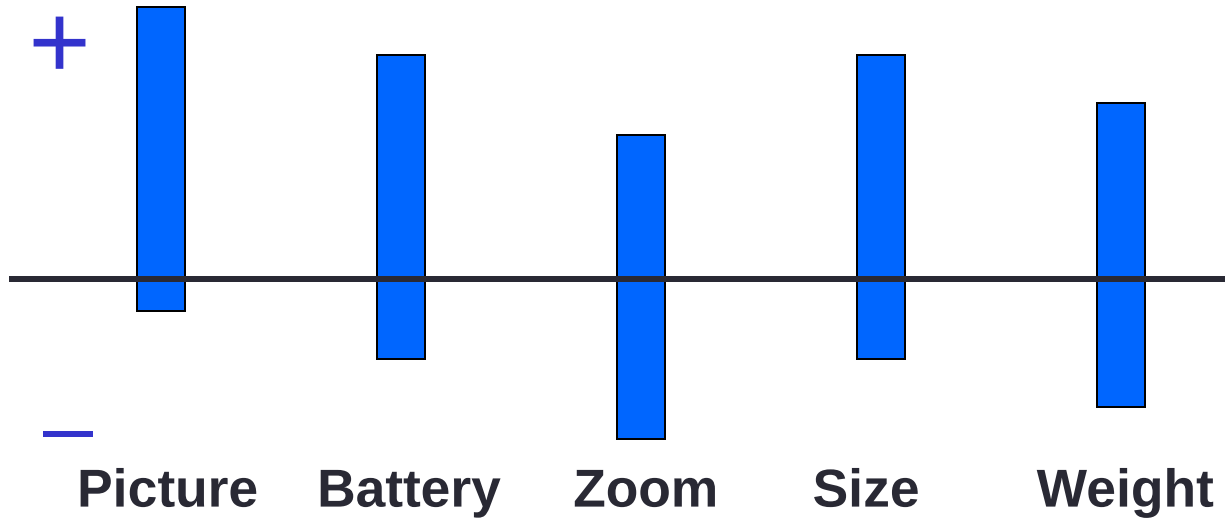
- The pictures come out hazy if your hands shake even for a moment during the entire process of taking a picture.
- Focusing on a display rack about 20 feet away in a brightly lit room during day time, pictures produced by this camera were blurry and in a shade of orange.

Feature2: **battery life**

…

Perbandingan visual

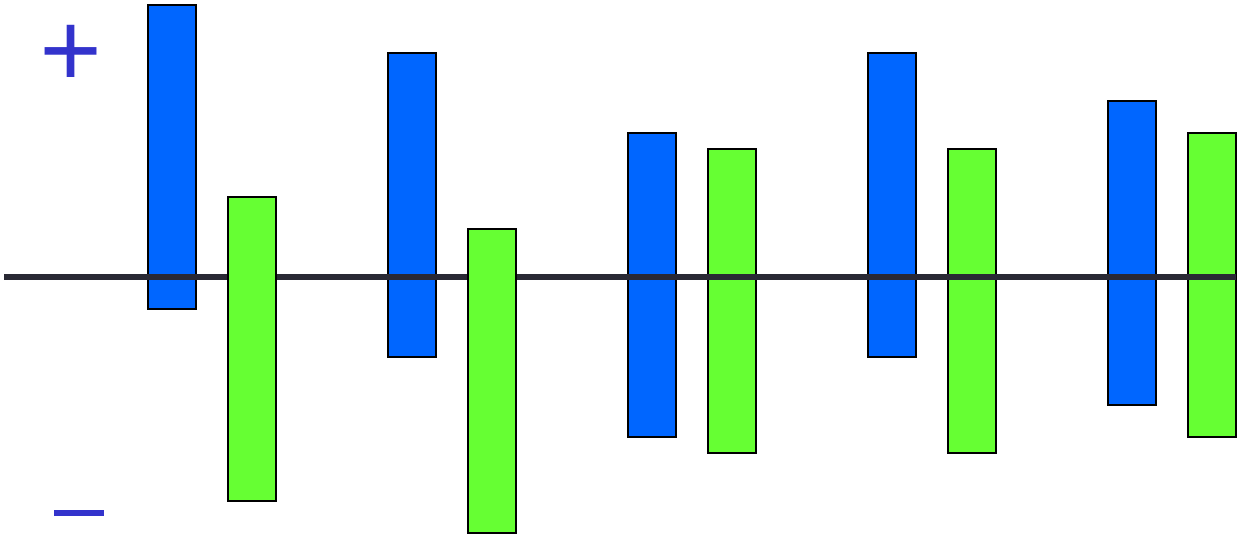
Ringkasan dari review dari Digital camera 1



Perbandingan dari review dari

Digital camera 1

Digital camera 2



Ekstraksi informasi

Posting from Newsgroup
Telecommunications. Solaris Systems
Administrator. 55-60K. Immediate need.

3P is a leading telecommunications firm
in need of a energetic individual to
fill the following position in the
Atlanta office:

SOLARIS SYSTEM ADMINISTRATOR
Salary: 50-60K with full benefits
Location: Atlanta, Georgia no relocation
assistance provided

FILLED TEMPLATE
job title: SOLARIS SYSTEM ADMINISTRATOR
salary: 55-60K
city: Atlanta
state: Georgia
platform: SOLARIS
area: Telecommunications

Document

I am a Windows NT software engineer seeking a permanent position in a small quiet town 50 - 100 miles from New York City.

I have over nineteen years of experience in all aspects of development of application software, with recent focus on design and implementation of systems involving multi-threading, client/server architecture, and anti-piracy. For the past five years, I have implemented Windows NT services in Visual C++ (in C and C++). I also have designed and implemented multithreaded applications in Java. Before working with Windows NT, I programmed in C under OpenVMS for 5 years.

Filled Template

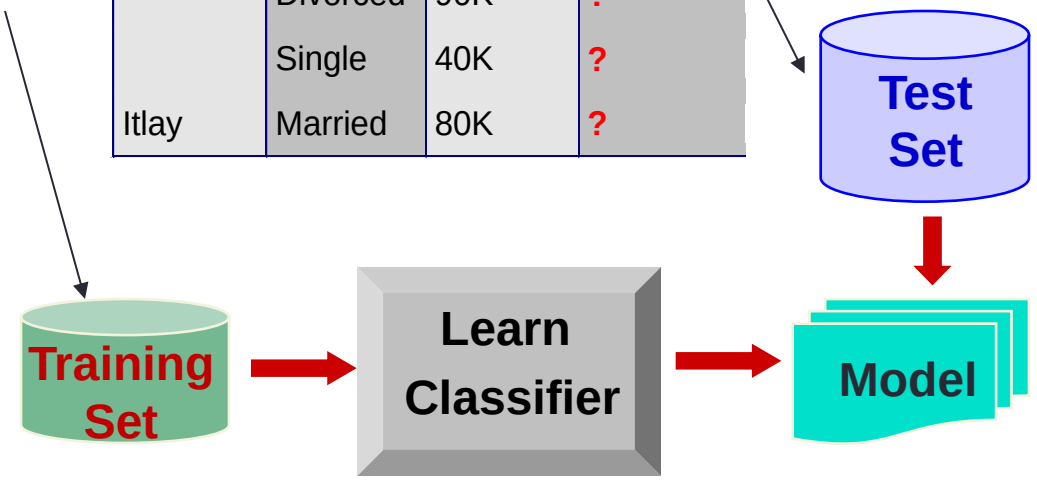
```
title: Windows NT software engineer
location: New York City
language: Visual C++, C, C++, Java
platform: Windows NT, OpenVMS
area: multi-threading, client/server,
      anti-piracy
years of experience: nineteen years
```

Klasifikasi: Contoh

categorical categorical continuous class

Ex#	Country	Marital Status	Income	Hooligan
1	England	Single	125K	Yes
2	England	Married		Yes
3	England	Single	70K	Yes
4	Italy	Married	40K	No
5	USA	Divorced	95K	No
6	England	Married	60K	Yes
7	England		20K	Yes
8	Italy	Single	85K	Yes
9	France	Married	75K	No
10	Denmark	Single	50K	No

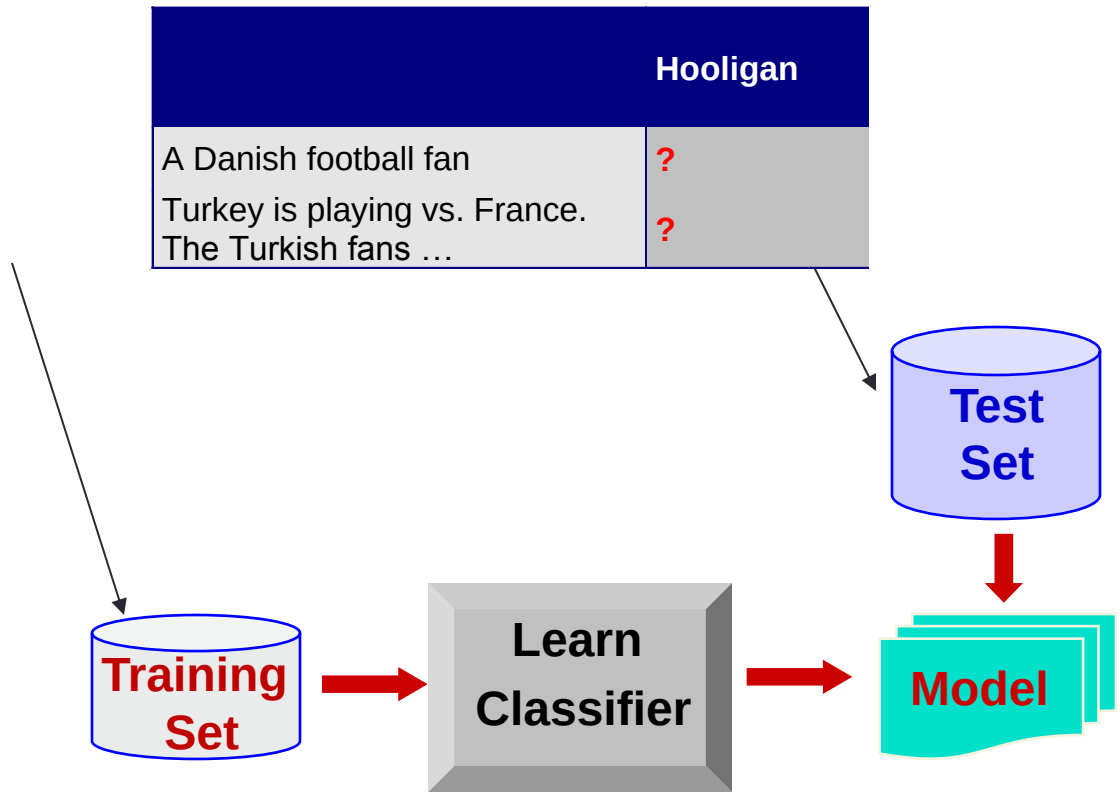
Country	Marital Status	Income	Hooligan
England	Single	75K	?
Turkey	Married	50K	?
England	Married	150K	?
	Divorced	90K	?
	Single	40K	?
Itlay	Married	80K	?

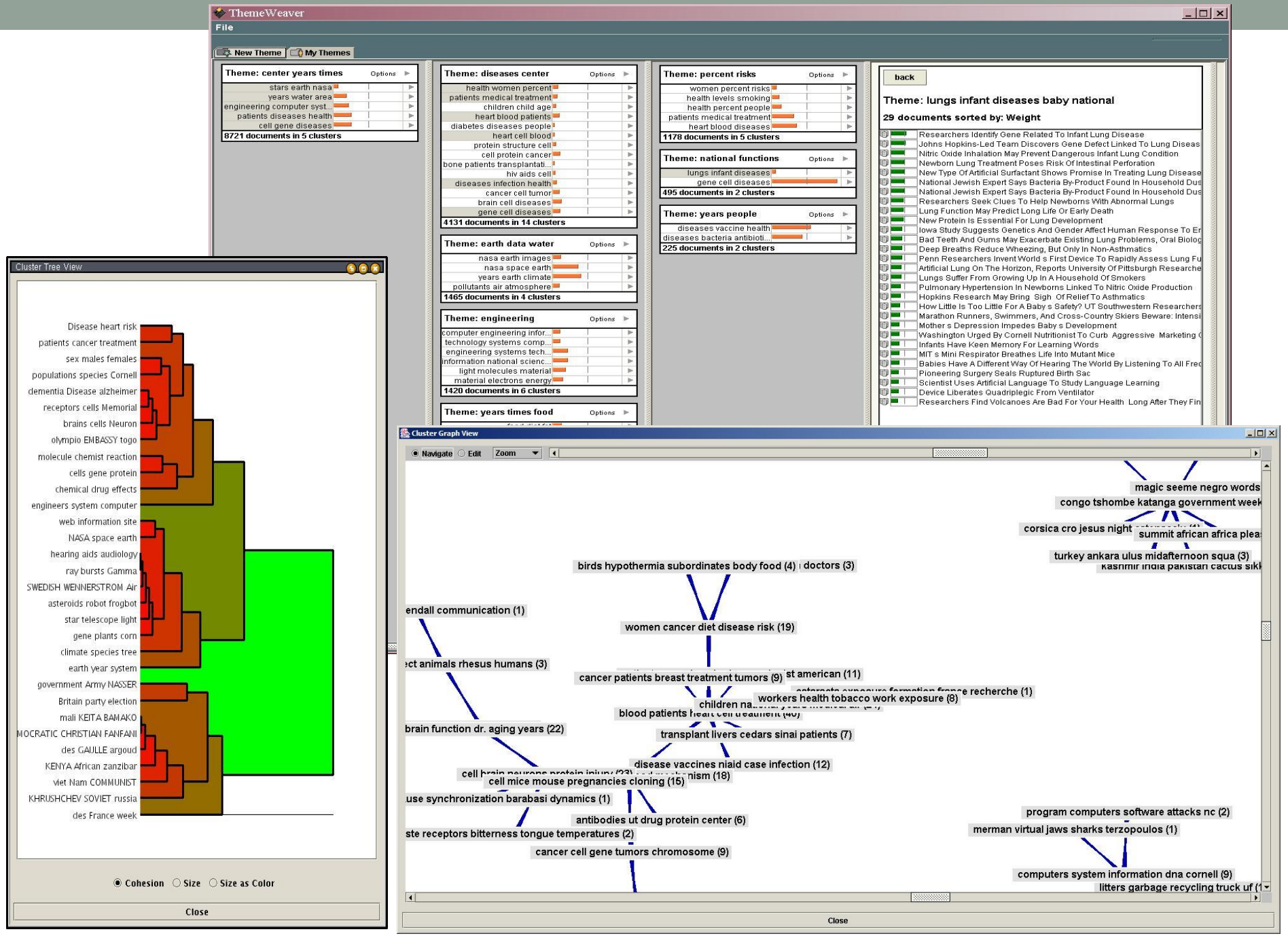


Klasifikasi teks: contoh

Ex#	text	Hooligan (class)
1	An English football fan …	Yes
2	During a game in Italy …	Yes
3	England has been beating France …	Yes
4	Italian football fans were cheering …	No
5	An average USA salesman earns 75K	No
6	The game in London was horrific	Yes
7	Manchester city is likely to win the championship	Yes
8	Rome is taking the lead in the football league	Yes

text	Hooligan
A Danish football fan	?
Turkey is playing vs. France. The Turkish fans …	?



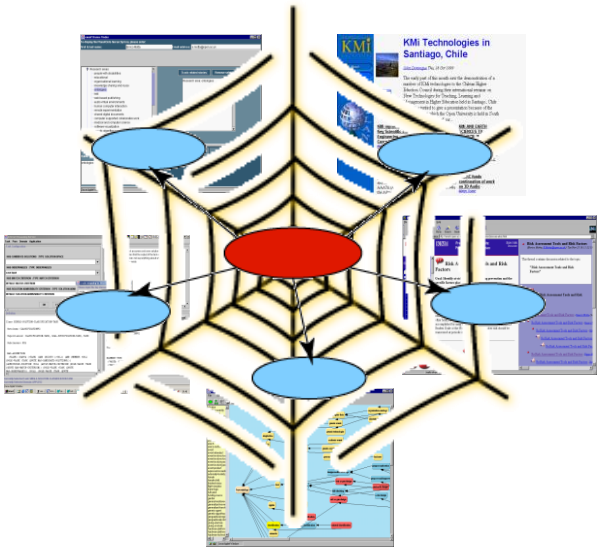

ThemeWeaver
File
New Theme
My Themes
Theme: center years times
Options
stars earth nasa
years water area
engineering computer syst...
patients diseases health
cell gene diseases
8721 documents in 5 clusters
Theme: diseases center
Options
health women percent
patients medical treatment
children child age
heart blood patients
diabetes diseases people
heart cell blood
protein structure cell
cell protein cancer
bone patients transplantati...
hiv aids cell
diseases infection health
cancer cell tumor
brain cell diseases
gene cell diseases
4131 documents in 14 clusters
Theme: earth data water
Options
nasa earth images
nasa space earth
years earth climate
pollutants air atmosphere
1465 documents in 4 clusters
Theme: engineering
Options
computer engineering infor...
technology systems comp...
engineering systems tech...
information national scienc...
light molecules material
material electrons energy
1420 documents in 6 clusters
Theme: years times food
Options
Theme: percent risks
Options
women percent risks
health levels smoking
health percent people
patients medical treatment
heart blood diseases
1178 documents in 5 clusters
Theme: national functions
Options
lungs infant diseases
gene cell diseases
495 documents in 2 clusters
Theme: years people
Options
diseases vaccine health
diseases bacteria antibioti...
225 documents in 2 clusters
back
Theme: lungs infant diseases baby national
29 documents sorted by: Weight
Researchers Identify Gene Related To Infant Lung Disease
Johns Hopkins-Led Team Discovers Gene Defect Linked To Lung Diseas
Nitric Oxide Inhalation May Prevent Dangerous Infant Lung Condition
Newborn Lung Treatment Poses Risk Of Intestinal Perforation
New Type Of Artificial Surfactant Shows Promise In Treating Lung Disease
National Jewish Expert Says Bacteria By-Product Found In Household Dus
National Jewish Expert Says Bacteria By-Product Found In Household Dus
Researchers Seek Clues To Help Newborns With Abnormal Lungs
Lung Function May Predict Long Life Or Early Death
New Protein Is Essential For Lung Development
Iowa Study Suggests Genetics And Gender Affect Human Response To Er
Bad Teeth And Gums May Exacerbate Existing Lung Problems, Oral Biolog
Deep Breaths Reduce Wheezing, But Only In Non-Asthmatics
Penn Researchers Invent World s First Device To Rapidly Assess Lung Fu
Artificial Lung On The Horizon, Reports University Of Pittsburgh Researche
Lungs Suffer From Growing Up In A Household Of Smokers
Pulmonary Hypertension In Newborns Linked To Nitric Oxide Production
Hopkins Research May Bring Sigh Of Relief To Asthmatics
How Little Is Too Little For A Baby s Safety? UT Southwestern Researchers
Marathon Runners, Swimmers, And Cross-Country Skiers Beware: Intensi
Mother s Depression Impedes Baby s Development
Washington Urged By Cornell Nutritionist To Curb Aggressive Marketing
Infants Have Keen Memory For Learning Words
MIT s Mini Respirator Breathes Life Into Mutant Mice
Babies Have A Different Way Of Hearing The World By Listening To All Fre
Pioneering Surgery Seals Ruptured Birth Sac
Scientist Uses Artificial Language To Study Language Learning
Device Liberates Quadriplegic From Ventilator
Researchers Find Volcanoes Are Bad For Your Health Long After They Fin
Cluster Tree View
Disease heart risk
patients cancer treatment
sex males females
populations species Cornell
dementia Disease alzheimer
receptors cells Memorial
brains cells Neuron
olympio EMBASSY togo
molecule chemist reaction
cells gene protein
chemical drug effects
engineers system computer
web information site
NASA space earth
hearing aids audiology
ray bursts Gamma
SWEDISH WENNERSTROM Air
asteroids robot frogbot
star telescope light
gene plants corn
climate species tree
earth year system
government Army NASSER
Britain party election
mali KEITA BAMAKO
MOCRATIC CHRISTIAN FANFANI
des GAULLE argoud
KENYA African zanzibar
viet Nam COMMUNIST
KHRUSHCHEV SOVIET russia
des France week
Cohesion
Size
Size as Color
Close
Cluster Graph View
Navigate
Edit
Zoom
magic seeme negro words
congo tshombe katanga government week
corsica cro jesus night
summit african africa plea
turkey ankara ulus midafternoon squa (3)
birds hypothermia subordinates body food (4)
doctors (3)
endall communication (1)
women cancer diet disease risk (19)
ect animals rhesus humans (3)
cancer patients breast treatment tumors (9)
st american (11)
workers health tobacco work exposure (8)
brain function dr. aging years (22)
transplant livers cedars sinai patients (7)
disease vaccines niaid case infection (12)
cell mice mouse pregnancies cloning (15)
use synchronization barabasi dynamics (1)
antibodies ut drug protein center (6)
ste receptors bitterness tongue temperatures (2)
cancer cell gene tumors chromosome (9)
program computers software attacks nc (2)
merman virtual jaws sharks terzopoulos (1)
computers system information dna cornell (9)
Close

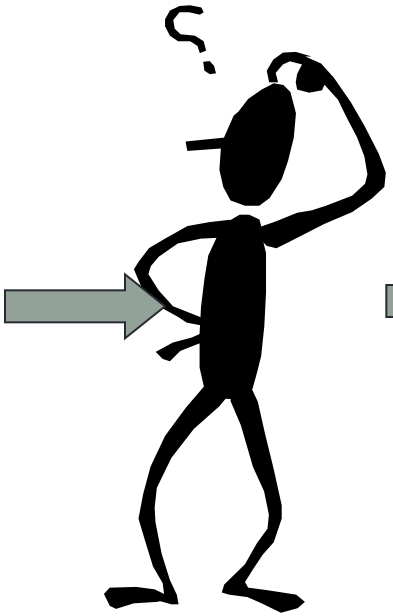
Web Mining

Data mining – Ilmu Komputer IPB

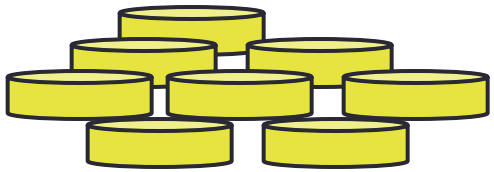
Web Mining

www





pengetahuan



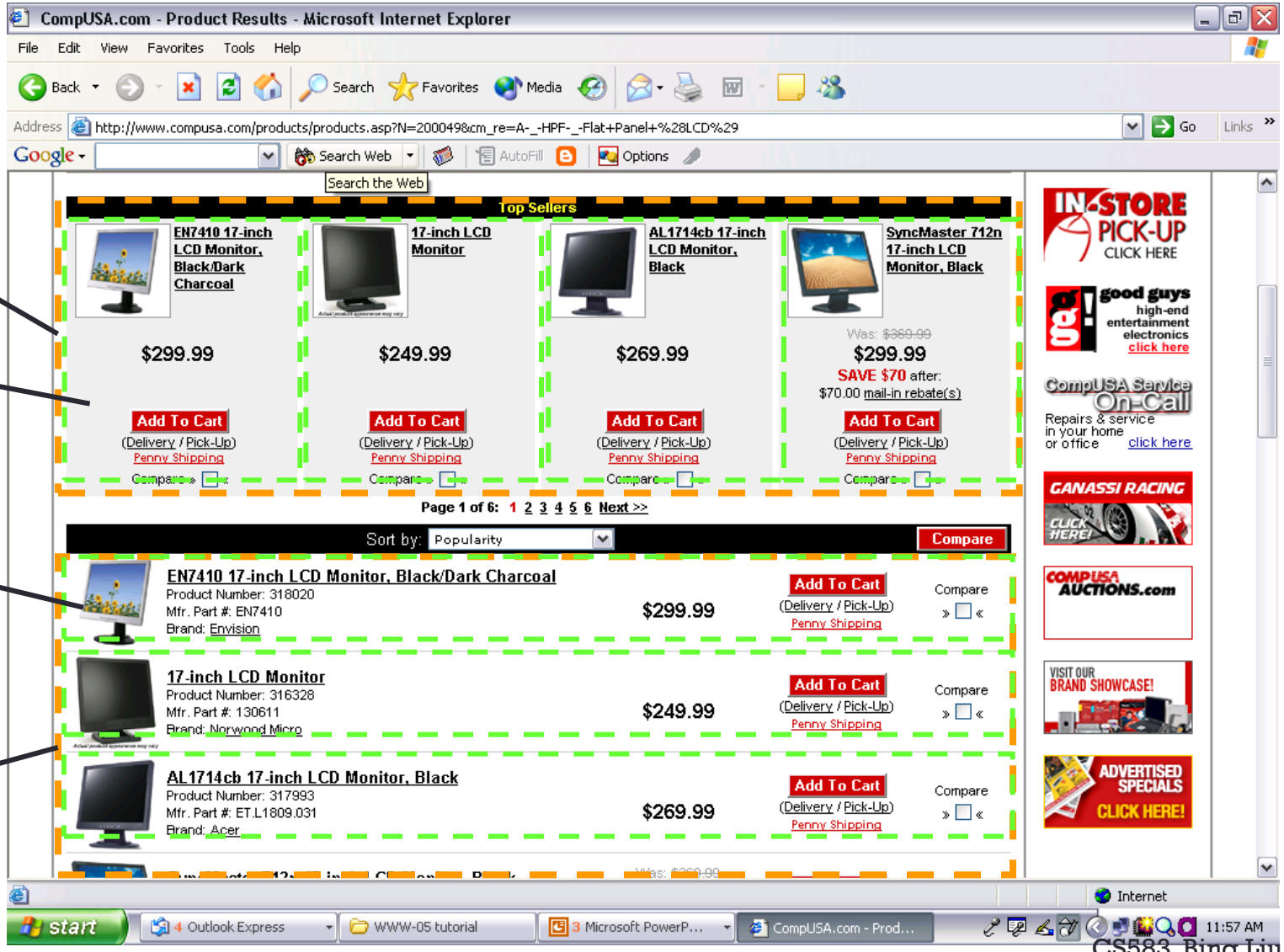
Contoh: Ekstraksi data web

Data bagian 1

Record data

Record data

Data bagian 2



Ekstraksi item data (contoh bagian 1)

image1	EN7410 17-inch LCD Monitor Black/Dark charcoal		$299.99		Add to Cart	(Delivery / Pick-Up)	Penny Shopping	Compare
image2	17-inch LCD Monitor		$249.99		Add to Cart	(Delivery / Pick-Up)	Penny Shopping	Compare
image3	AL1714 17-inch LCD Monitor, Black		$269.99		Add to Cart	(Delivery / Pick-Up)	Penny Shopping	Compare
image4	SyncMaster 712n 17-inch LCD Monitor, Black	Was: $369.99	$299.99	Save $70 After: $70 mail-in-rebate(s)	Add to Cart	(Delivery / Pick-Up)	Penny Shopping	Compare

Apa itu Web Mining?

Mencari informasi menarik dan berguna dari konten dan penggunaan Web

- Web search : Google, Yahoo, MSN, Ask, …
- Specialized search: contoh: http://www.google.com/shopping (comparison shopping), job ads (Flipdog)
- eCommerce :
  - Rekomentasi:contoh Netflix, Amazon
  - Memperbaiki tingkat konversi: produk terbaik berikutnya yang ditawarkan
- Iklan, contoh. Google Adsense
- Fraud detection: click fraud detection, …
- Meningkatkan rancangan dan kinerja Web site

Web Mining

- Web mining – teknik-teknik data mining untuk mencari dan mengekstrak secara otomatis informasi dari dokumen/layanan Web (Etzioni, 1996).
- Penelitian Web mining research – integrasi dari beberapa komunitas penelitian (Kosala and Blockeel, Juli 2000) seperti:
  - Basis data (DB)
  - Information retrieval (IR)
  - Sub area machine learning (ML)
  - Natural language processing (NLP)

Web Mining

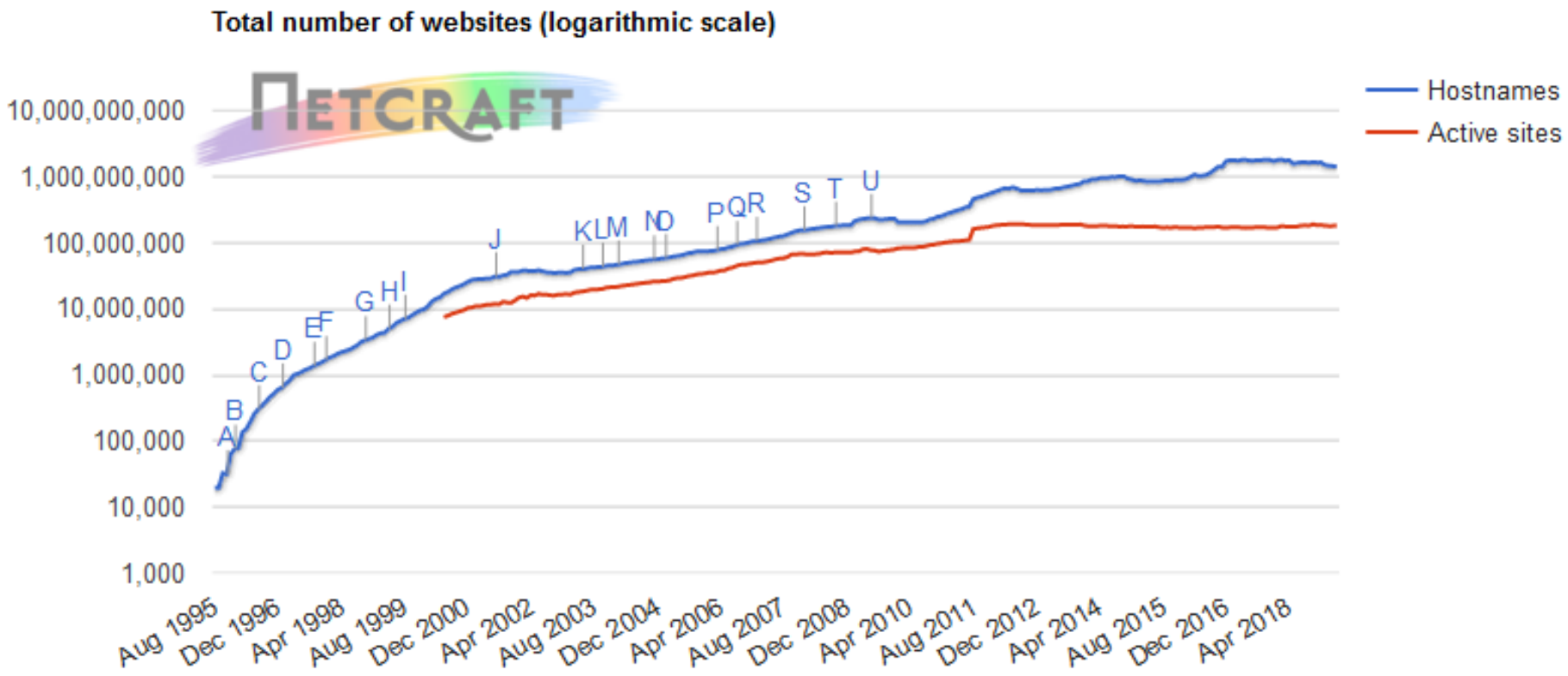
- World Wide Web dapat memberikan lebih banyak tantangan dan kesempatan daripada area lainnya.
- Walaupun demikian, terdapat tantangan serius:
  - www berukuan besar
  - Kompleksitas halaman web lebih besar dari koleksi dokumen teks lainnya
  - Sangat dinamis
  - Keragaman yang luas dari penggunanya
  - Hanya sedikit porsi informasi yang benar-benar berguna

Seberapa besar ukuan Web ?

Mengapa menambang Web?

- **Web kaya akan informasi tekstual**
  - Book/CD/Video stores (contoh, Amazon)
  - Informasi restoran (contoh, Zagat)
  - Harga mobil (contoh, Carpoint)
- **Banyaknya data pada pola akses pengguna**
  - Web logs mengandung urutan URL yang diakses oleh pengguna
- **Menggali informasi yang “previously unknown”**
  - People who ski also frequently break their leg.
  - Restaurants that serve sea food in California are likely to be outside San-Francisco

April 2019 Web Server Survey: 1,445,266,139 sites



http://news.netcraft.com/archives/category/web-server-survey/

Fitur unik pada Web

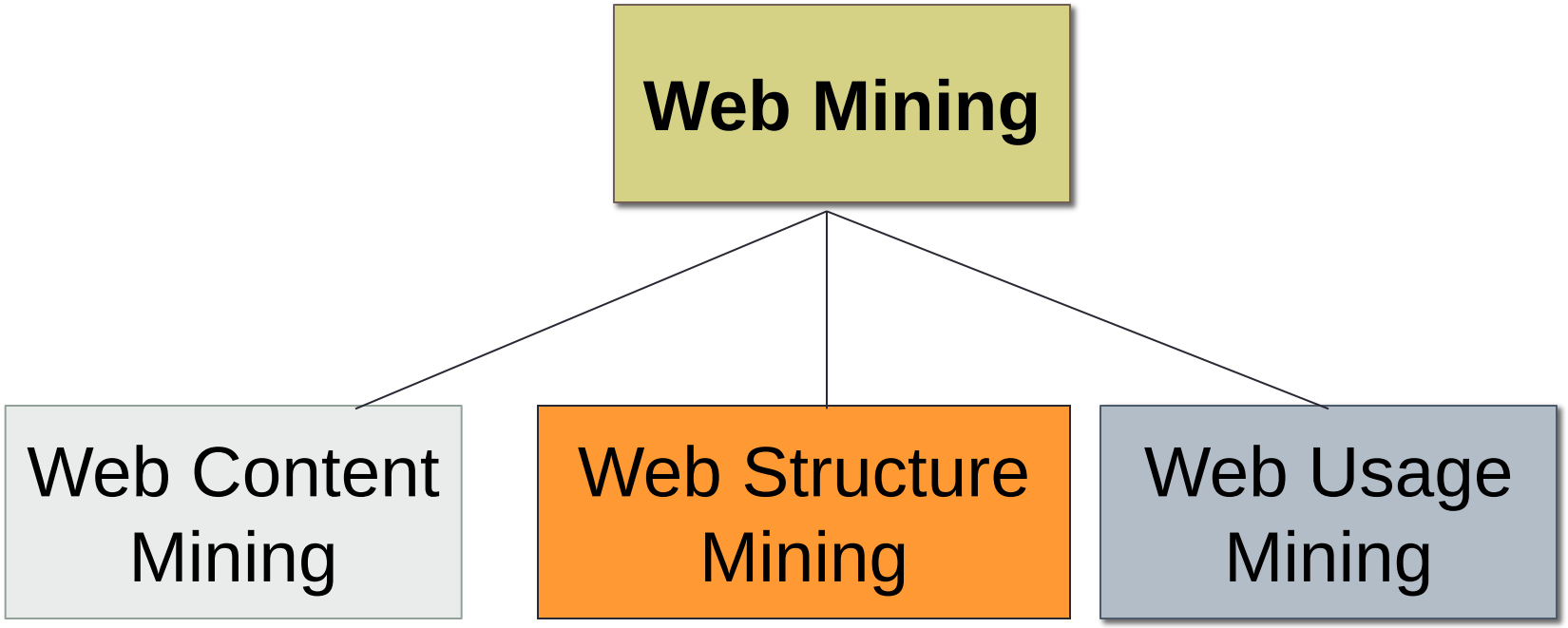
- Web adalah koleksi dokumen berukuran besar yang mengandung:
  - Informasi Hyper-link
  - Informasi akses dan penggunaan
- Web sangat dinamis
  - Halaman Web secara konstan dibangkitkan (dibuang)

Tantangan: Membangun algoritme Web mining baru untuk
- Eksploitasi hyper-links dan pola akses.
- Beradaptasi terhadap sumber dokumennya

Web Mining vs Data Mining

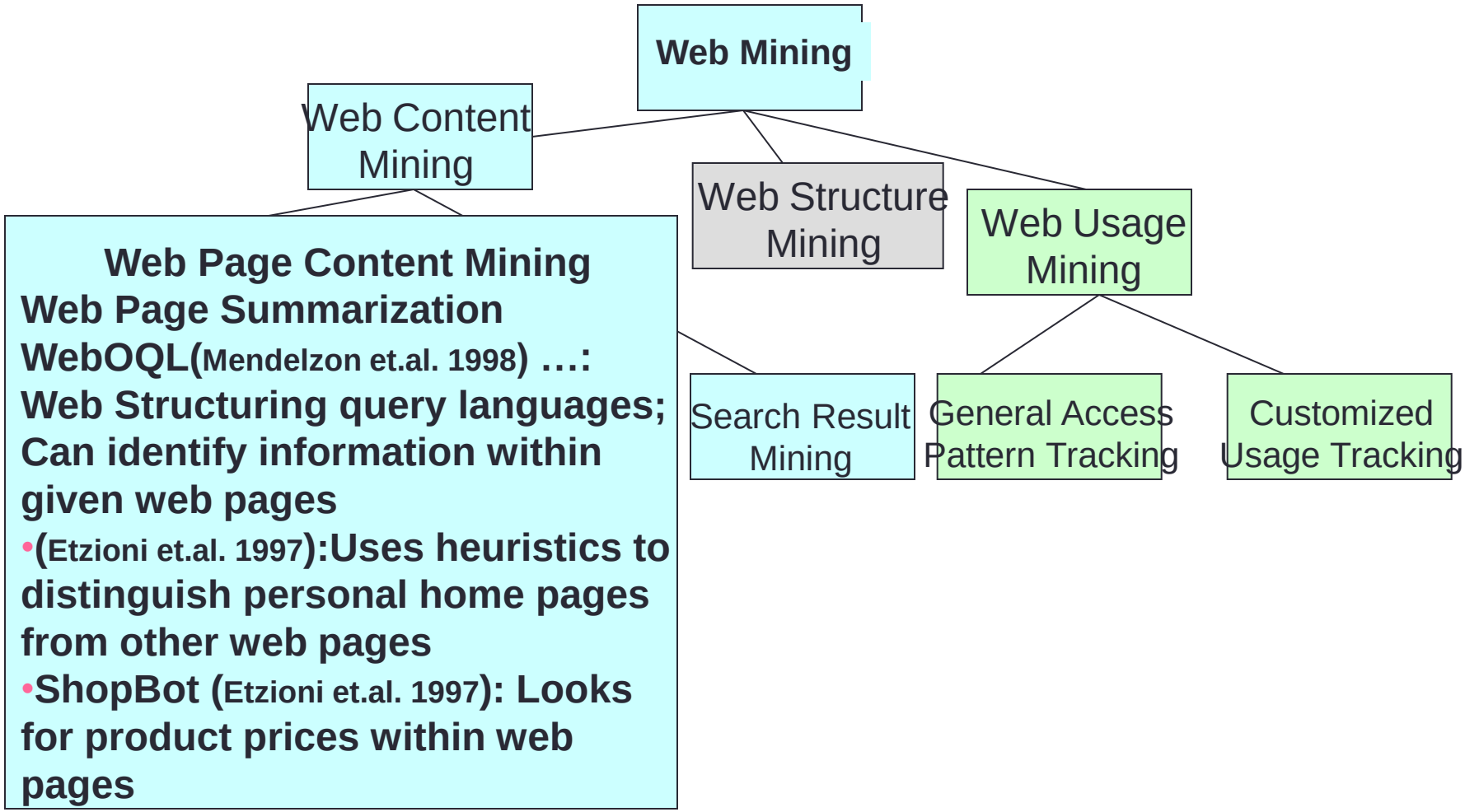
Struktur	• Web bukan merupakan relasi • Informasi tekstual dan struktur linkage
Skala	• Data penggunaan berukuran besar dan berkembang secara cepat • Data yang dibangkitkan per hari dapat dibandingkan dengan data warehouse konvensional terbesar
Kecepatan	• Seringkali perlu bereaksi untuk mengembangkan pola penggunaan secara real-time (contoh: merchandising) • Manusia tidak terlibat dalam proses

Taksonomi Web Mining

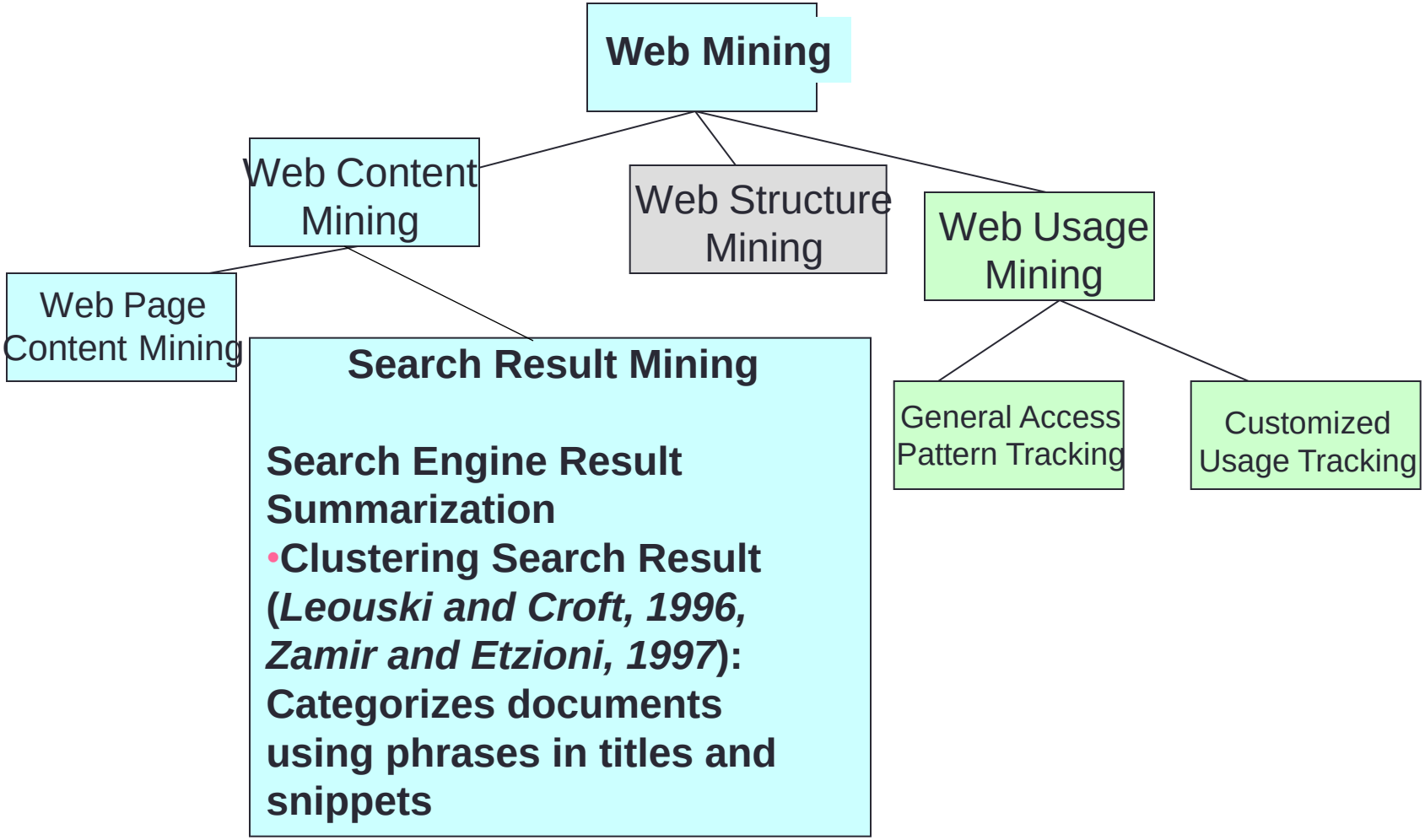


Taksonomi Web Mining

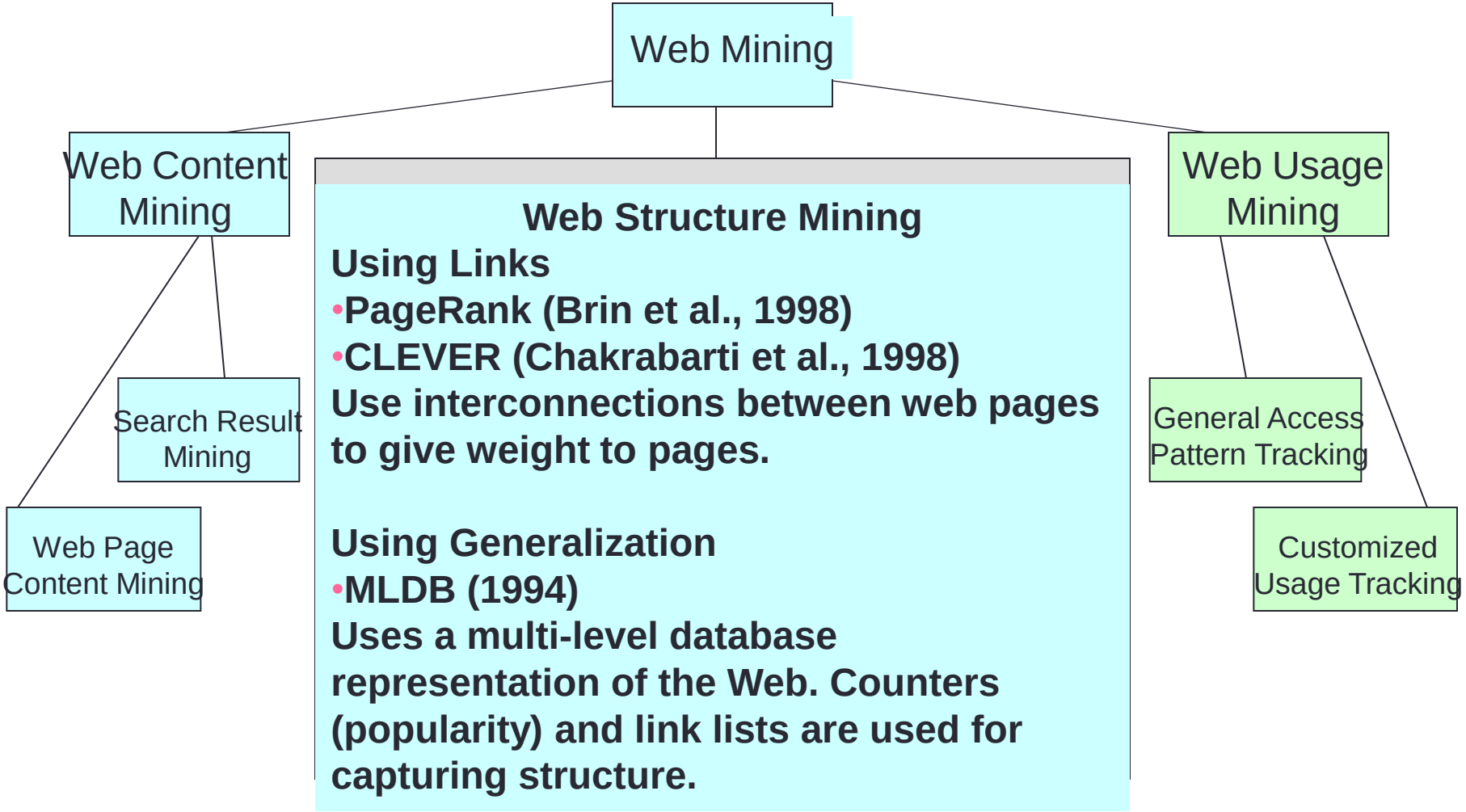
Mining the World Wide Web



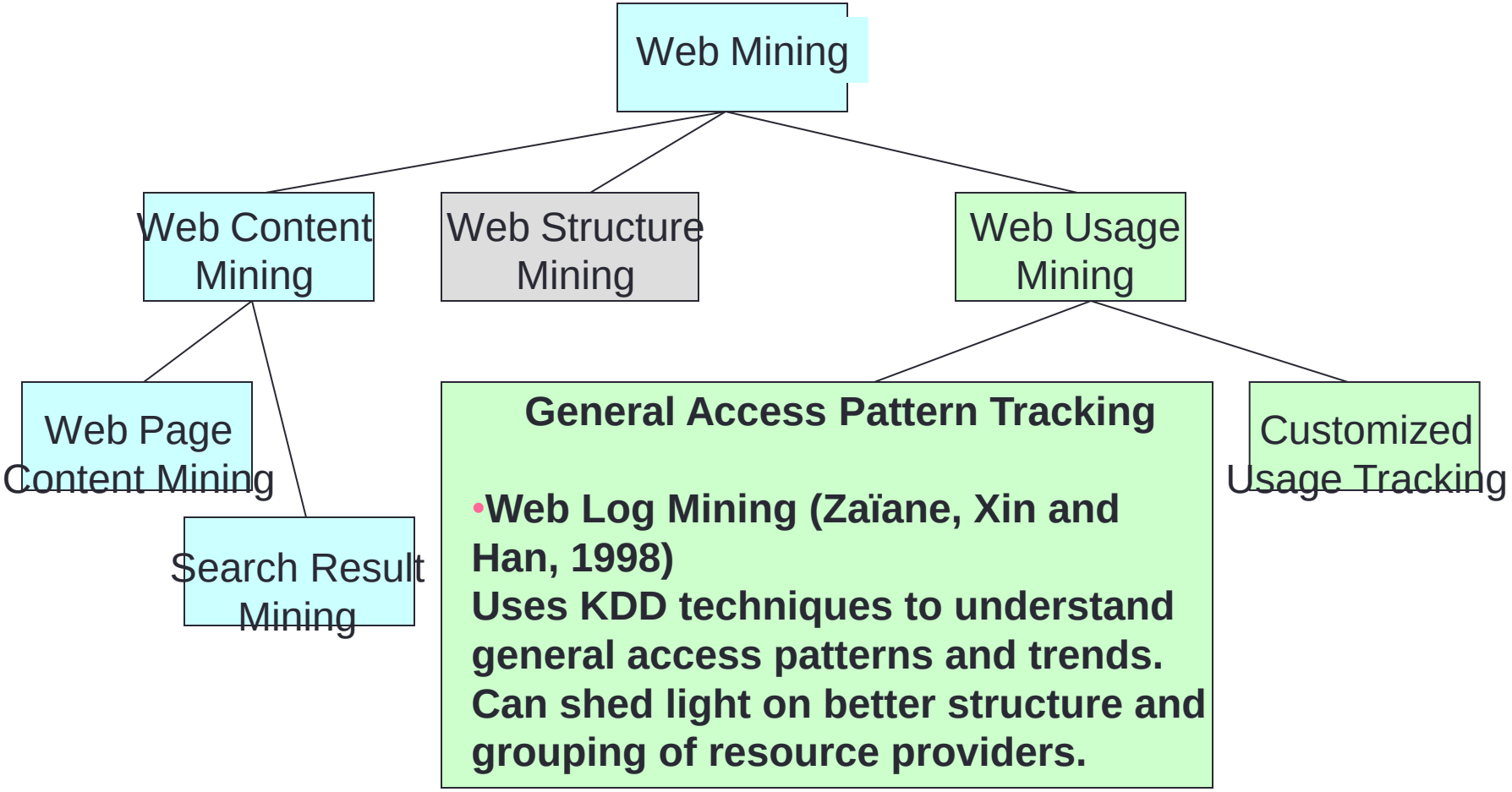
Mining the World Wide Web



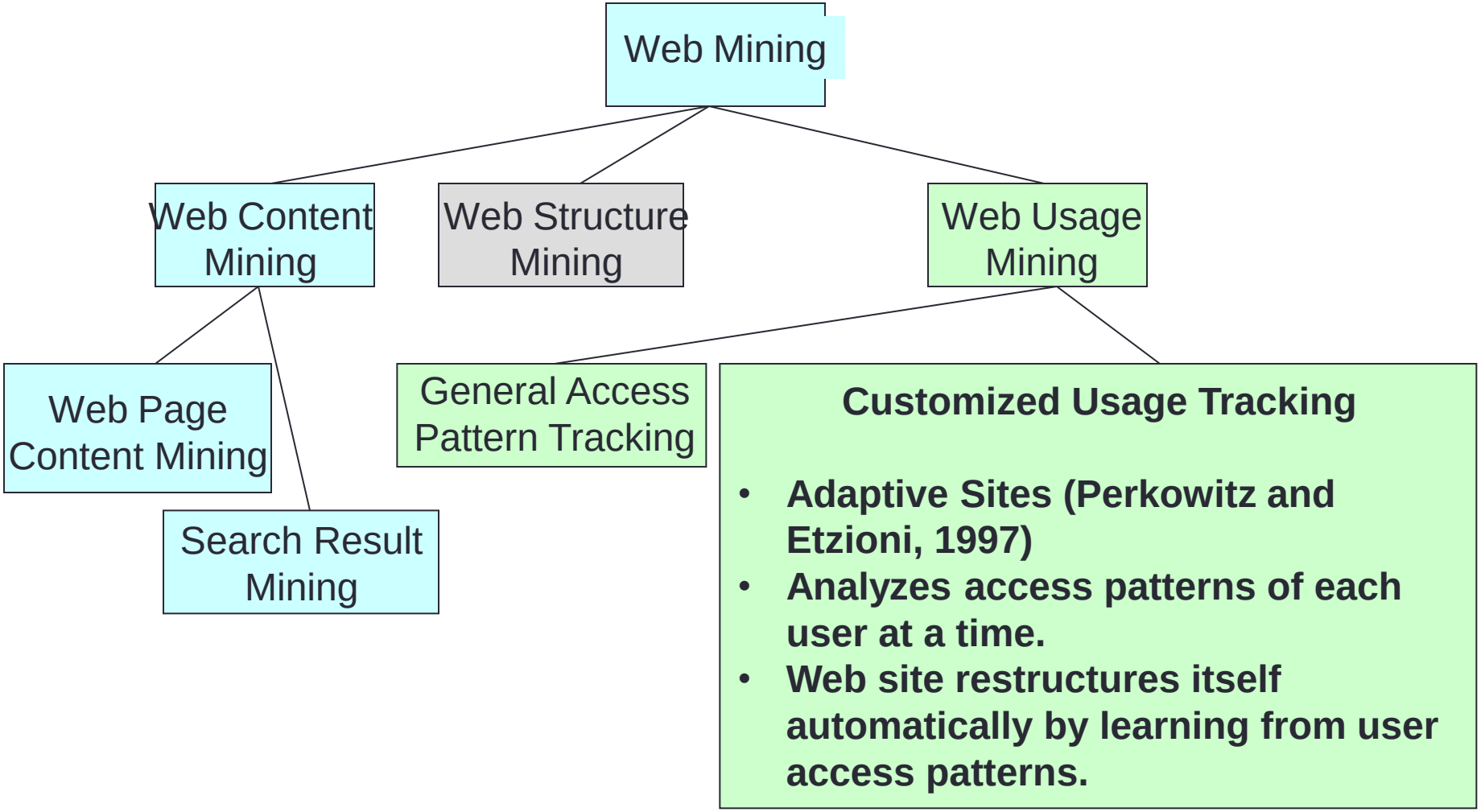
Mining the World Wide Web



Mining the World Wide Web



Mining the World Wide Web



Pendekatan-pendekatan Web Content Mining

- Pendekatan Information Retrieval
  - Membantu dan meningkatkan pencarian dan filtering informasi untuk pengguna biasanya berdasarkan pada profil pengguna yang disimpulkan atau diminta.
- Pendekatan basis data
  - Untuk memodelkan data pada Web dan mengintegrasikannya sedemikian sehingga kueri yang kompleks selain berbasis keywords dapat dilakukan.

Web Content Mining

	IR View	DB View
View of Data	Unstructured Semi-structured	Semi-structured Web site as DB
Main Data	Text documents Hypertext documents	Hypertext documents
Representation	Bag of words, n-grams Terms, phrases Concepts or ontology Relational	Edge-labeled graph Relational
Methods	Machine Learning Statistics	ILP Association rules
Applications	Categorization Clustering Finding extraction rules Finding patterns in text User modeling	Finding frequent substructures Web site schema discovery

Isu dalam Web Content Mining

- Pengembangan alat cerdas untuk IR
  - Mencari kata kunci & frasa kunci
  - Menemukan aturan gramatikal & collocation
- Klasifikasi/kategorisasi hyperteks
  - Mengekstra frasa kunci dari dokumen html
- Ekstraksi model/aturan pembelajaran
  - Hierarchical clustering
  - Memprediksi keterhubungan kata
- Membangun web Query system (WebOQL, XMLQL)
- Mining multimedia data

Web Structure Mining

View of Data	Links structure
Main Data	Links structure
Representation	Graph
Methods	Proprietary algorithms
Applications	Categorization Clustering

Web Structure Mining

- Untuk menemukan struktur link dari hyperlinks pada level antardokumen untuk membangun ringkasan struktur tentang situs web
  - Arah 1: berbasis hyperlinks, mengkategorikan halaman Web & informasi yang dibangun
  - Arah 2: menemukan struktur dari dokumen web itu sendiri
  - Arah 3: menemukan kealamiahan hierarki/jaringan hyperlinks pada situs web tertentu

Web Structure Mining

- Menemukan halaman web yang authorative
  - Menemukembalikan halaman yang tidak hanya relevan, tapi juga berkualitas tinggi/authorative terhadap topik

- Hyperlinks dapat merujuk authority
  - Web mengandung juga hyperlinks dari satu halaman ke halaman lain
  - Hyperlinks mengandung anotasi manusia berjumlah besar
  - Hyperlink yang merujuk ke halaman lain, dapat dipertimbangkan sebagai kesukaan pengarang terhadap halaman lain

Web Usage Mining

View of Data	Interactivity
Main Data	Server logs Browser logs
Representation	Relational table Graph
Methods	Machine learning Statistics Association rules
Applications	Site construction, adaptation & management Marketing User modeling

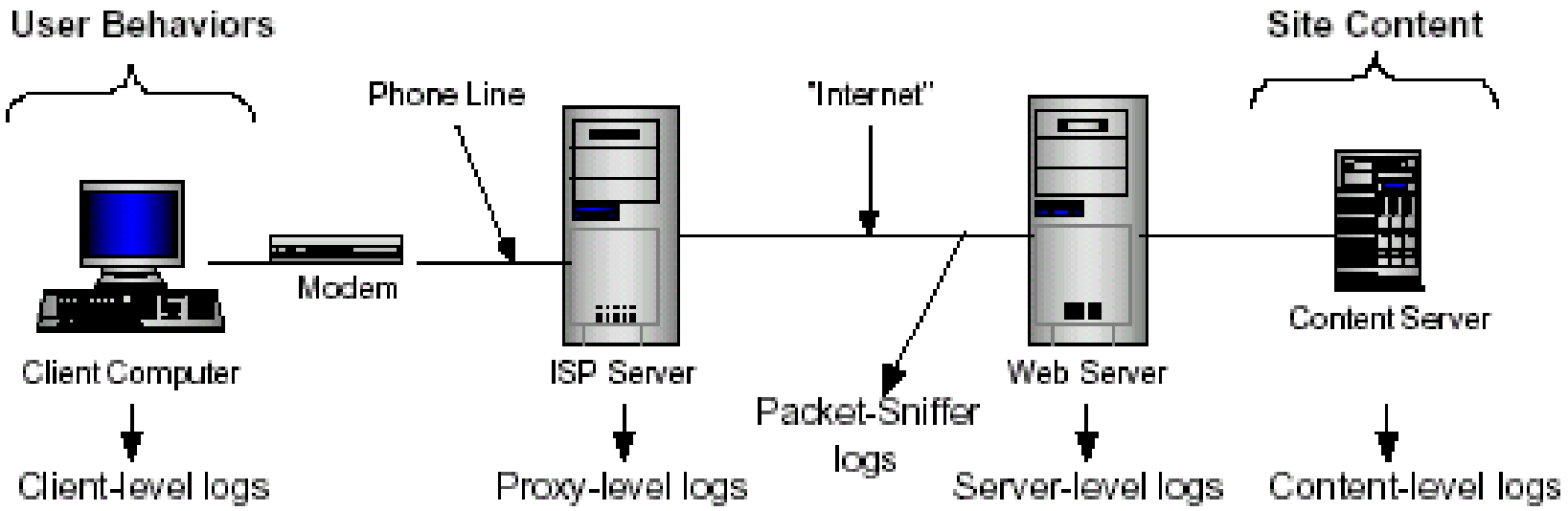
Web Usage Mining

- **Web usage mining** juga disebut Web log mining
  - Teknik mining untuk menemukan pola penggunaan yang menarik dari data sekunder yang diturunkan dari interaksi pengguna ketika menjelajahi web

Web Usage Mining

- Aplikasi
  - Menargetkan kostumer yang potensial untuk produk elektronik
  - Memperluas kualitas dan pengantaran Internet Information Services kepada pengguna akhir.
  - Memperbaiki performa sistem web server
  - Mengidentifikasi lokasi iklan yang potensial
  - Memfasilitasi personalisasi/situs adaptif
  - Memperbaki desain situs
  - Deteksi fraud/intrusion
  - Memprediksi aksi pengguna

Potential Data Sources



Log Data – Analisis Sederhana

- Analisis statistik dari pengguna
  - Panjang path
  - Viewing time
  - Banyaknya page views
- Analisis statistik dari site
  - Halaman yang paling seting dilihat
  - Invalid URL yang paling umum terjadi

Web Log – Data Mining Applications

- Association rules
  - Find pages that are often viewed together
- Clustering
  - Cluster users based on browsing patterns
  - Cluster pages based on content
- Classification
  - Relate user attributes to patterns

Format umum data Log

- Remotehost: browser hostname or IP #
- Remote log name of user (almost always "-" meaning "unknown")
- Authuser: authenticated username
- Date: Date and time of the request
- "request”: exact request lines from client
- Status: The HTTP status code returned
- Bytes: The content-length of response

SERVER LOGS

What's in a Typical Server Log?

<ip_addr><base_url> - <date><method><file><protocol><code><bytes><referrer><user_agent>

```
203.30.5.145 www.acr-news.org - [01/Jun/1999:03:09:21 -0600] "GET /Calls/OWOH.html
HTTP/1.0" 200 3942 "http://www.lycos.com/cgi-
bin/pursuit?query=advertising+psychology&maxhits=20&cat=dir" "Mozilla/4.5 [en] (Win98; I)"

203.30.5.145 www.acr-news.org - [01/Jun/1999:03:09:23 -0600] "GET
/Calls/Images/earthani.gif HTTP/1.0" 200 10689 "http://www.acr-news.org/Calls/OWOH.html"
"Mozilla/4.5 [en] (Win98; I)"

203.30.5.145 www.acr-news.org - [01/Jun/1999:03:09:24 -0600] "GET /Calls/Images/line.gif
HTTP/1.0" 200 190 "http://www.acr-news.org/Calls/OWOH.html" "Mozilla/4.5 [en] (Win98; I)"

203.30.5.145 www.acr-news.org - [01/Jun/1999:03:09:25 -0600] "GET /Calls/Images/red.gif
HTTP/1.0" 200 104 "http://www.acr-news.org/Calls/OWOH.html" "Mozilla/4.5 [en] (Win98; I)"

203.252.234.33 www.acr-news.org - [01/Jun/1999:03:32:31 -0600] "GET / HTTP/1.0" 200 4980 ""
"Mozilla/4.06 [en] (Win95; I)"

203.252.234.33 www.acr-news.org - [01/Jun/1999:03:32:35 -0600] "GET /Images/line.gif
HTTP/1.0" 200 190 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"

203.252.234.33 www.acr-news.org - [01/Jun/1999:03:32:35 -0600] "GET /Images/red.gif
HTTP/1.0" 200 104 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"

203.252.234.33 www.acr-news.org - [01/Jun/1999:03:32:35 -0600] "GET /Images/earthani.gif
HTTP/1.0" 200 10689 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"

203.252.234.33 www.acr-news.org - [01/Jun/1999:03:33:11 -0600] "GET /CP.html HTTP/1.0" 200
3218 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"
```

Fields

- **Client IP:** 128.101.228.20
- **Authenticated User ID:** - -
- **Time/Date:** [10/Nov/1999:10:16:39 -0600]
- **Request:** "GET / HTTP/1.0"
- **Status:** 200
- **Bytes:** -
- **Referrer:** “-”
- **Agent:** "Mozilla/4.61 [en] (WinNT; I)"

Ringkasan tipe data kompleks

- Area yang sedang berkembang dalam menambang tipe data yang kompleks:
  - **Text mining** dapat dilakukan cukup efektif, khususnya jika dokumen adalah semi terstruktur
  - **Web mining** lebih sulit dilakukan karena kurangnya struktur demikian (semi terstruktur)
    - Data mencakup dokumen teks, dokumen hypertext, link structure, dan logs
    - Mengandalkan unsupervised learning, kadang-kadang diikuti dengan supervised learning seperti klasifikasi