INGENIERÍA INDUSTRIAL
UNIVERSIDAD DE CHILE

FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

Chapter 1: Introduction

Juan Domingo Velásquez Silva

Cristobal Gaspar Ignacio Pizarro Venegas

Summary: Introduction

- Initial Concepts
  - The World Wide Web
  - E-business
  - Toward new portal generation
  - Objective of the lecture
- The KDD process
- Web Data and Web Mining
- Web Intelligence

Section 1.1

Initial Concepts

The World Wide Web

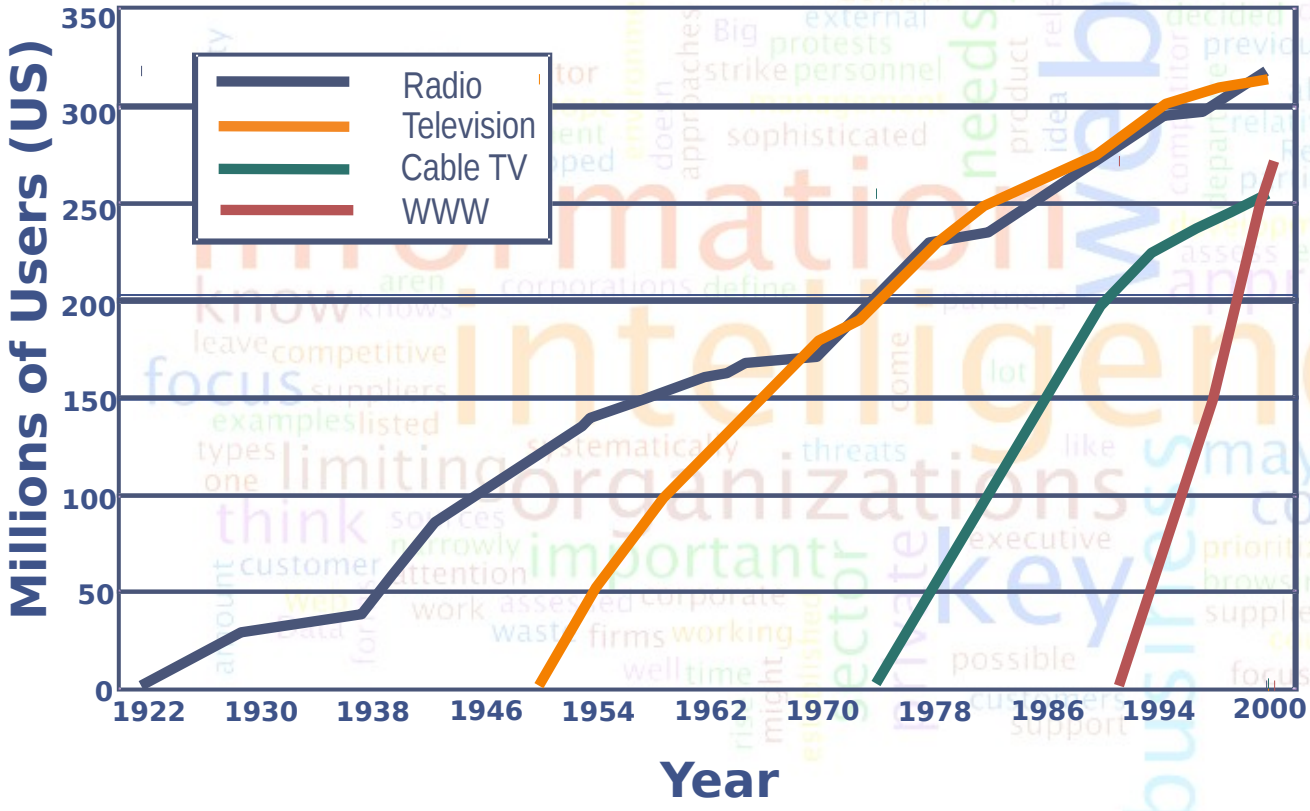
Tim Berners-Lee (1993)

"The World Wide Web (W3) is the universe of network-accessible information, an embodiment of human knowledge. It is an initiative started at CERN, now with many participants. It has a body of software, and a set of protocols and conventions. W3 uses hypertext and multimedia techniques to make the web easy for anyone to roam, browse, and contribute to"

The NET

- Toward the high speedway Of the information.
- **¿Who is the owner?**
- **Network of Network**
- **Exponential growth**

Technology adoption curve



Time to reach 200 Million Users:

Radio ….55 Years

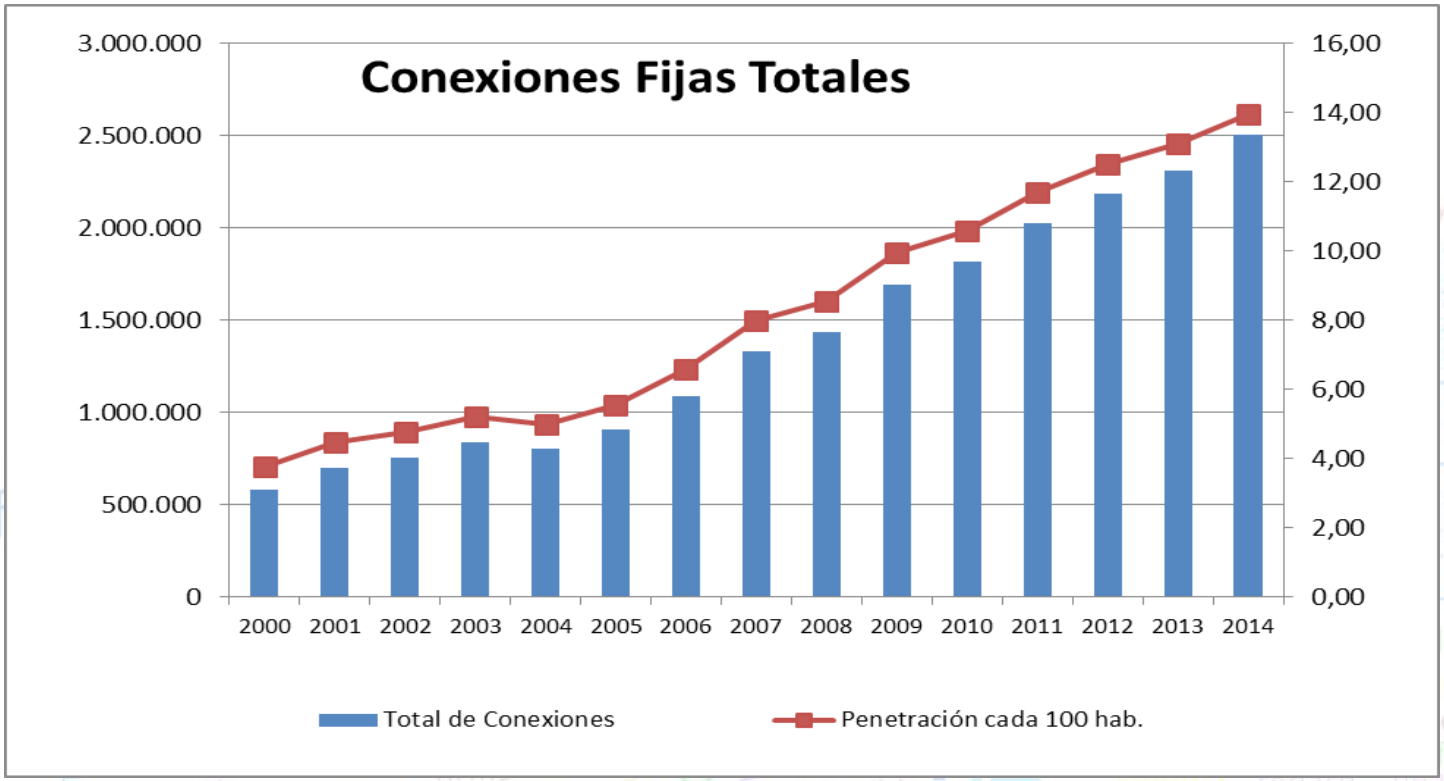
TV ……. 25 Years

Cable ... 12 Years

WWW…. 5 Years

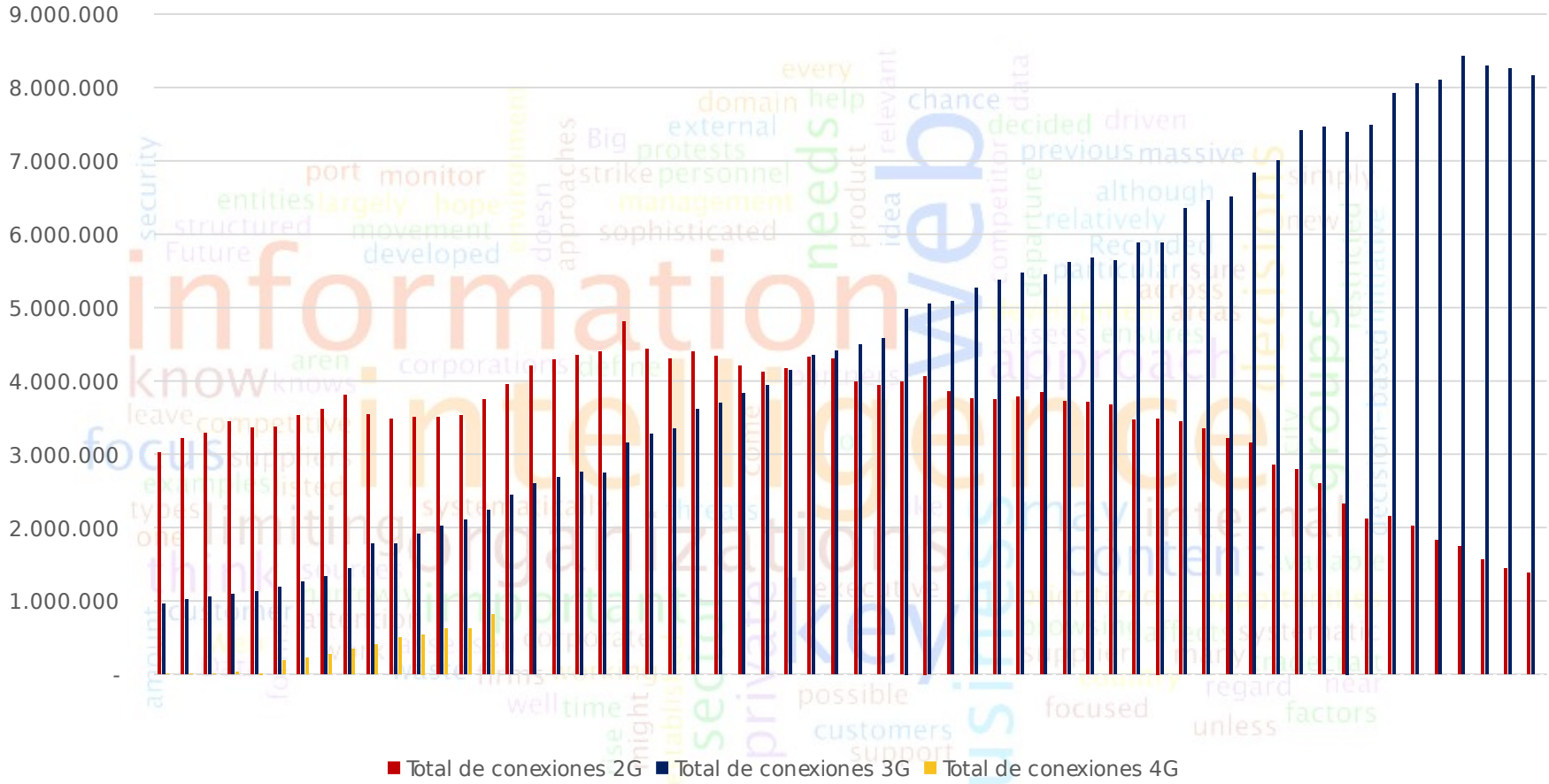
Source: Morgan Stanley Technology Research

Internet in Chile



Internet
Series conexiones internet fija (Fecha Publicación 16 de junio de 2015)
(Período Información Primer Trimestre 2002 – Marzo 2015)

Number of Mobile Internet Connections

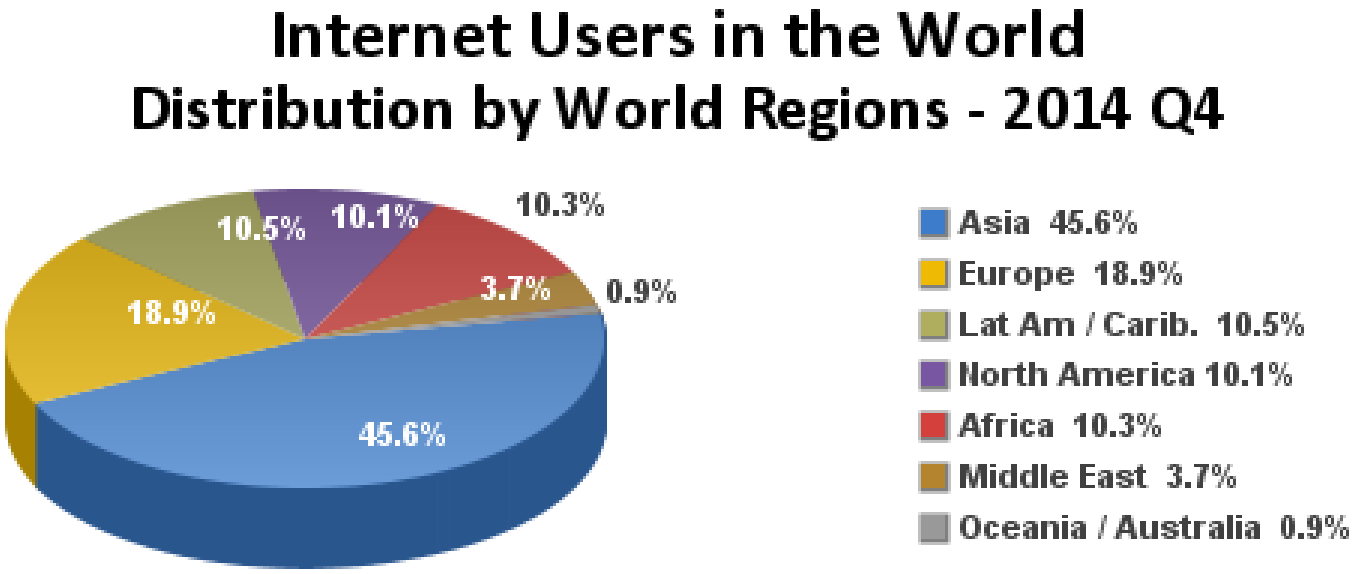


Series conexiones internet móvil (Fecha Publicación 16 de junio de 2015)
(Período Información Primer Trimestre 2002 – Marzo 2015)

Chile

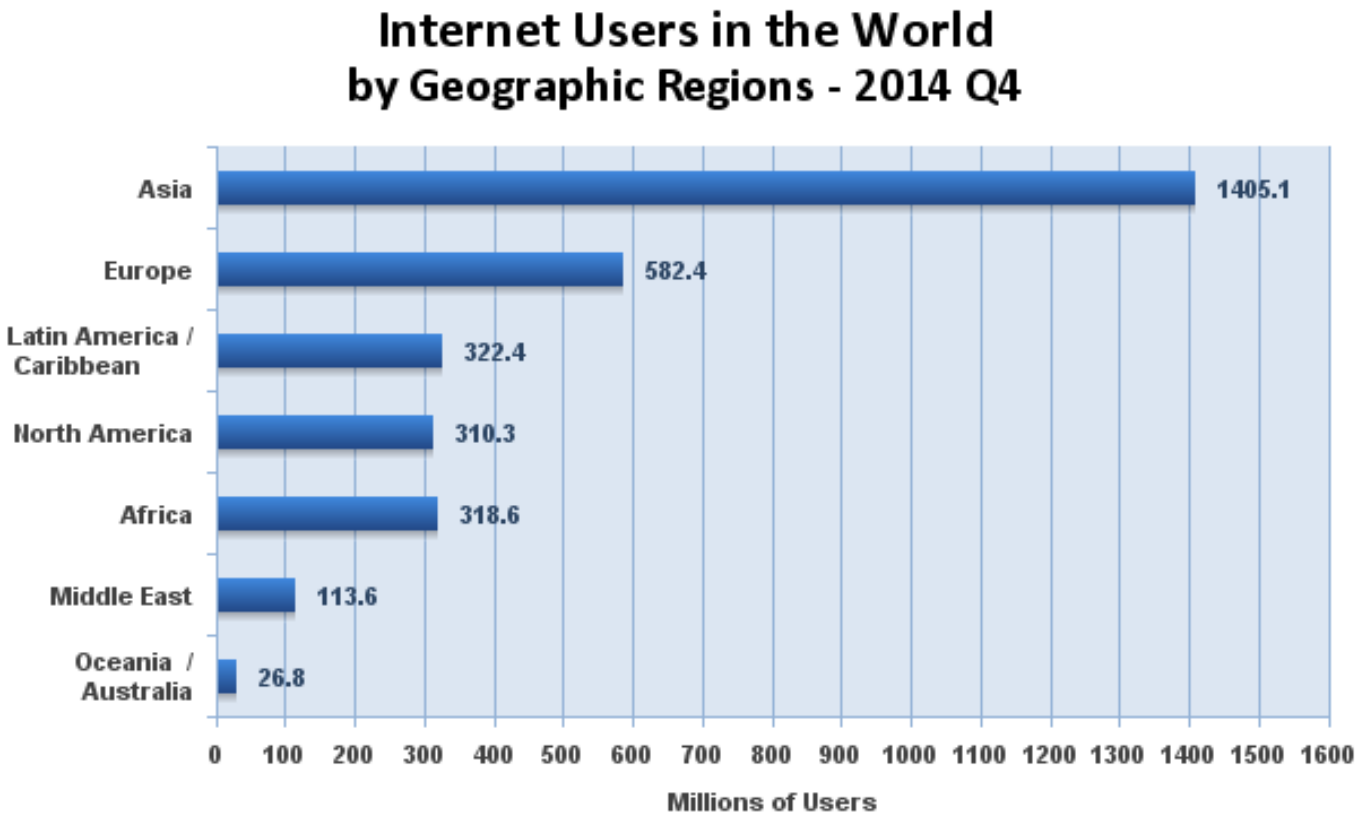
- 2008, in Chile we had 7.9 million of Internet users
- The e-commerce got sales by USM$ 14.558 during 2008, growing a 20% respect 2007,being:
  - 97% B2B y B2G.
  - B2C: 3%

Internet in the World



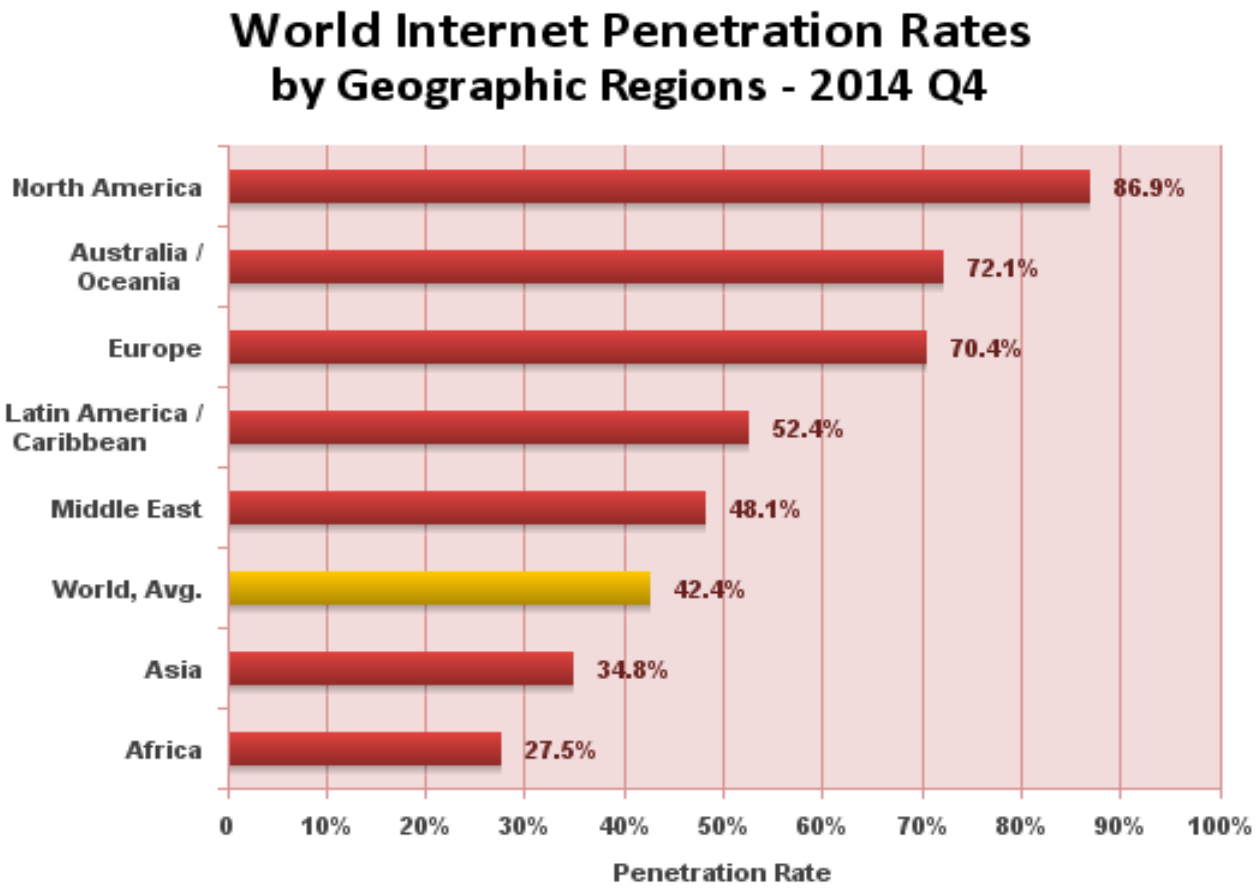
Source: Internet World Stats - www.internetworldstats.com/stats.htm
Basis: 3,079,339,857 Internet users on Dec 31, 2014

Internet in the World



Source: Internet World Stats - www.internetworldstats.com/stats.htm
3,079,339,857 Internet users estimated for Dec 31, 2014
Copyright © 2015, Miniwatts Marketing Group

Geographic Penetration of Internet Users



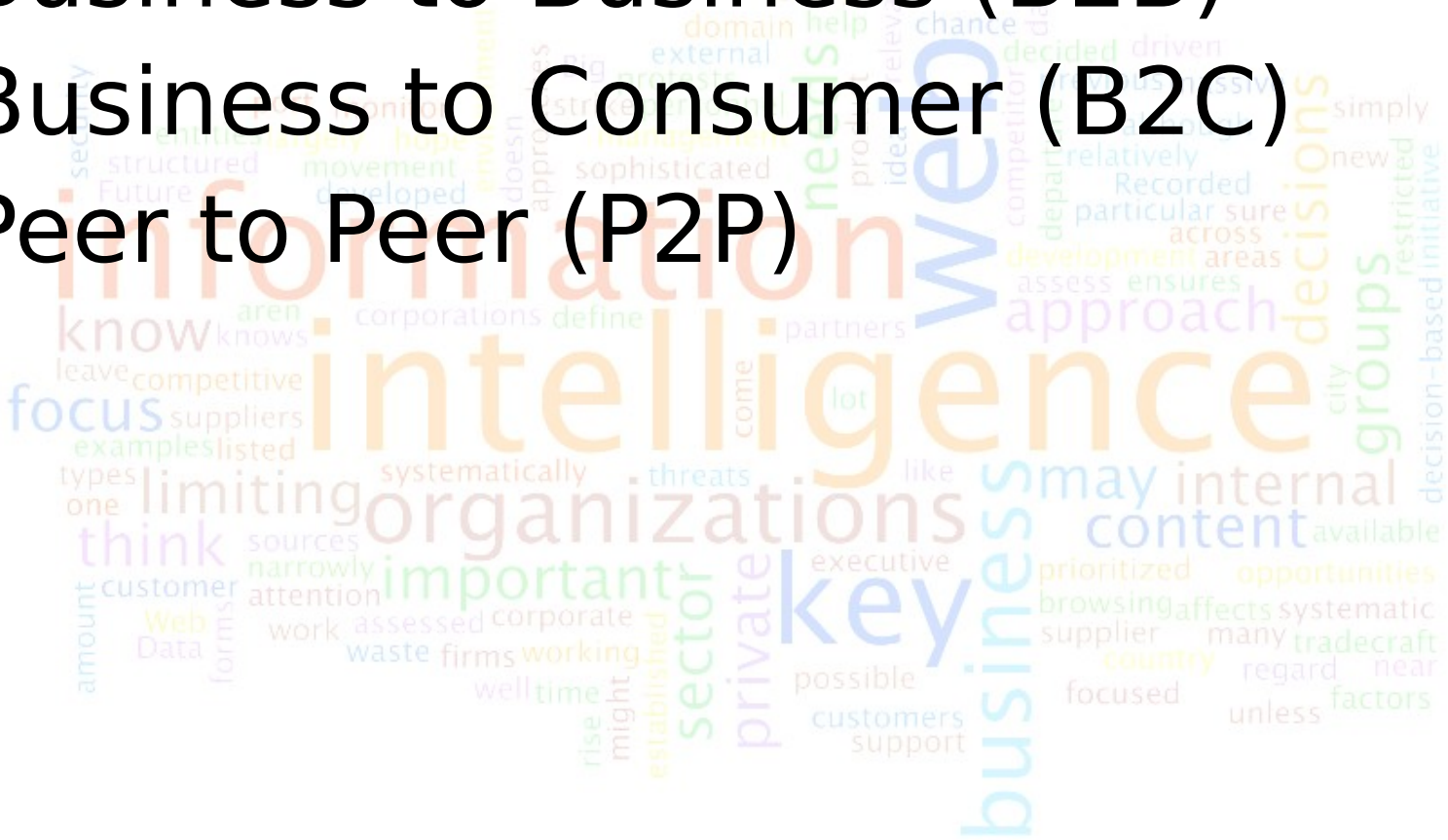
Source: Internet World Stats - www.internetworldststs.com/stats.htm
Penetration Rates are based on a world population of 7,264,623,793 and 3,079,339,857 estimated Internet users on Dec 31, 2014.

The computer is the network, the network is the computer

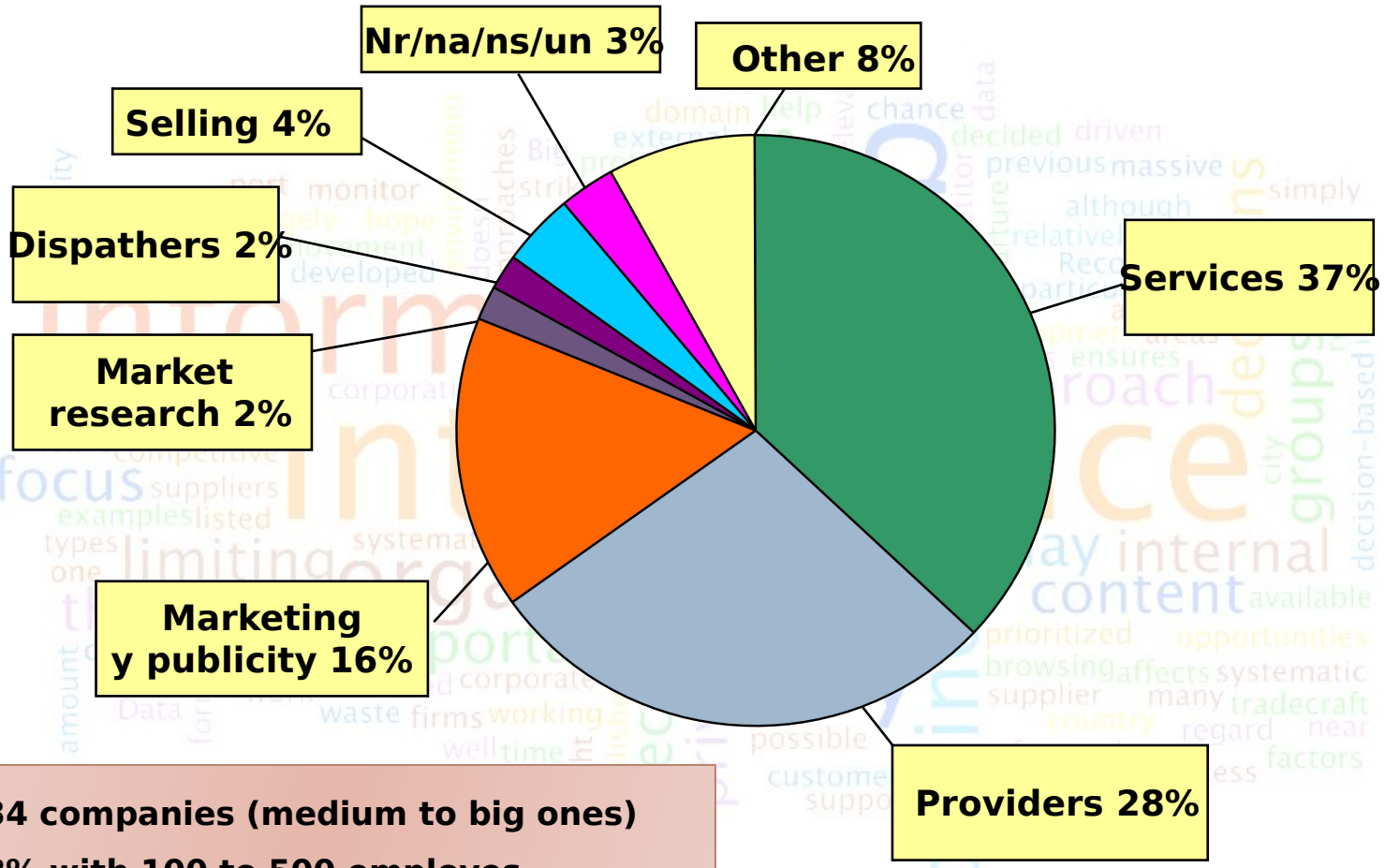
- The web is changing everything.
- The new economy: Google Model, Amazon, e-Banking.
- Change in the Supply Chain.
- Consumer directly perform OnLine product request.
- Reducing Information asymmetry gap.
- New kind of problem: Retain consumer in a web environment, The new web consumer profile, the web site as the e-commerce “service”.

E - Business

- Business to Business (B2B)
- Business to Consumer (B2C)
- Peer to Peer (P2P)



E-business distribution



- 134 companies (medium to big ones)
- 48% with 100 to 500 employes
- 52% over 500 employes

Source: IDC

E-business

- Something changed abruptly.
- Supply chain was altered.
- If intermediaries don’t add value to products, then they need to be erased.
- The new business models based on the new internet channel, lowers the chain costs.

E-business (2)

- Transactions Cost

Business	Transaction type	Cost (US$)
Cost by transaction	Cashier	1,07
	Phone	0,52
	ACM	0,27
	Internet	0,13
Cost of plane ticket	Travel Agency	8,0
	Internet	1,0
Insurance	Agent	550
	Internet	275
Software	Reseller	15
	Internet	0,35

But … there are some warnings

NASDAQ index evolution dotcom

The web portal: Our Point of Sale

What is the **ideal structure and content** of a web site?

- **Different users** have **distinct goals**
- The **behaviour** of users **changes over time**.
- **Sites must be restructured** as they **grow** to meet current needs, typically by accumulating pages and links.

The Adaptive Web Site (AWS)

- Based on **user behaviour**
- Web Site **recommendation**
- Use of Web Intelligence (WI)
  - Understanding **user preferences**
  - **Applications**
    - Web Usage Mining
    - Web Structure Mining
    - Web Content Mining
- Use of **Information Retrieval**

Data Mining techniques on the Web

- Web *Intelligence (WI)*
- Web Data: Very large amount of
  - Logs
  - Text and multimedia content
  - Structure of links
- Several tools of data mining apply to this field:
  - Clustering, regressions, association rules
- Important benefits returns from the mining process:
  - Google growth, e-business, e-market campaigns, CRM applications

Applications of web mining

- Recommendation System
- System for personalization
- Web Personalization
- Adaptive Web-based system
- Opinion Mining
- Community Discovery

Section 1.2

The KDD Process

KDD: Extracting knowledge from data

- The data resume for **decision making** is a **traditional report of the statistics**.
- Today **Information** is a **valuable resource** that have important implication in the **productivity of the company**. It administration is called **knowledge discovery**.
- This process usually imply a **large amount of data**.

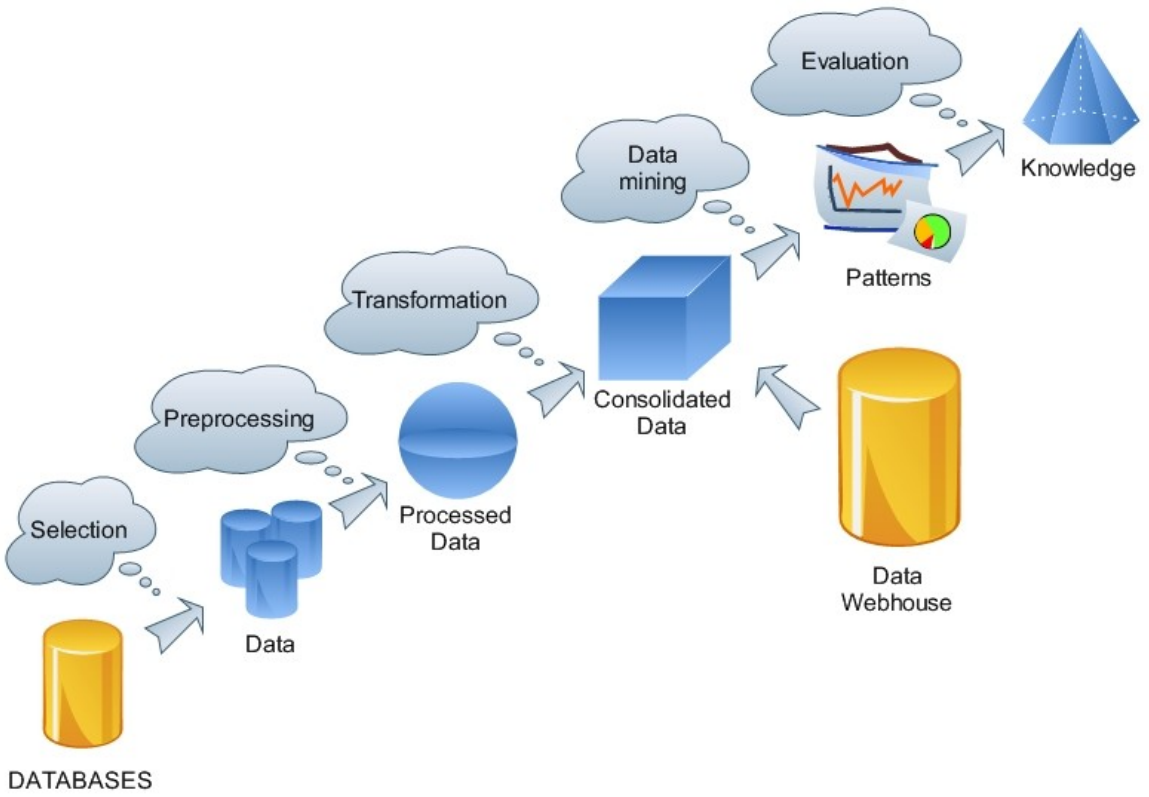
KDD: Extracting knowledge from data

- **The Goal**
  - Find **knowledge valid, useful, relevant** and **new** over a phenomena or activity.
- A **visual representation** of the result in order for an **easier interpretation**.
- The **usability** must allow a flexible, dynamic and collaborative process.
- **Scalability** and **efficiency** are important requirements.

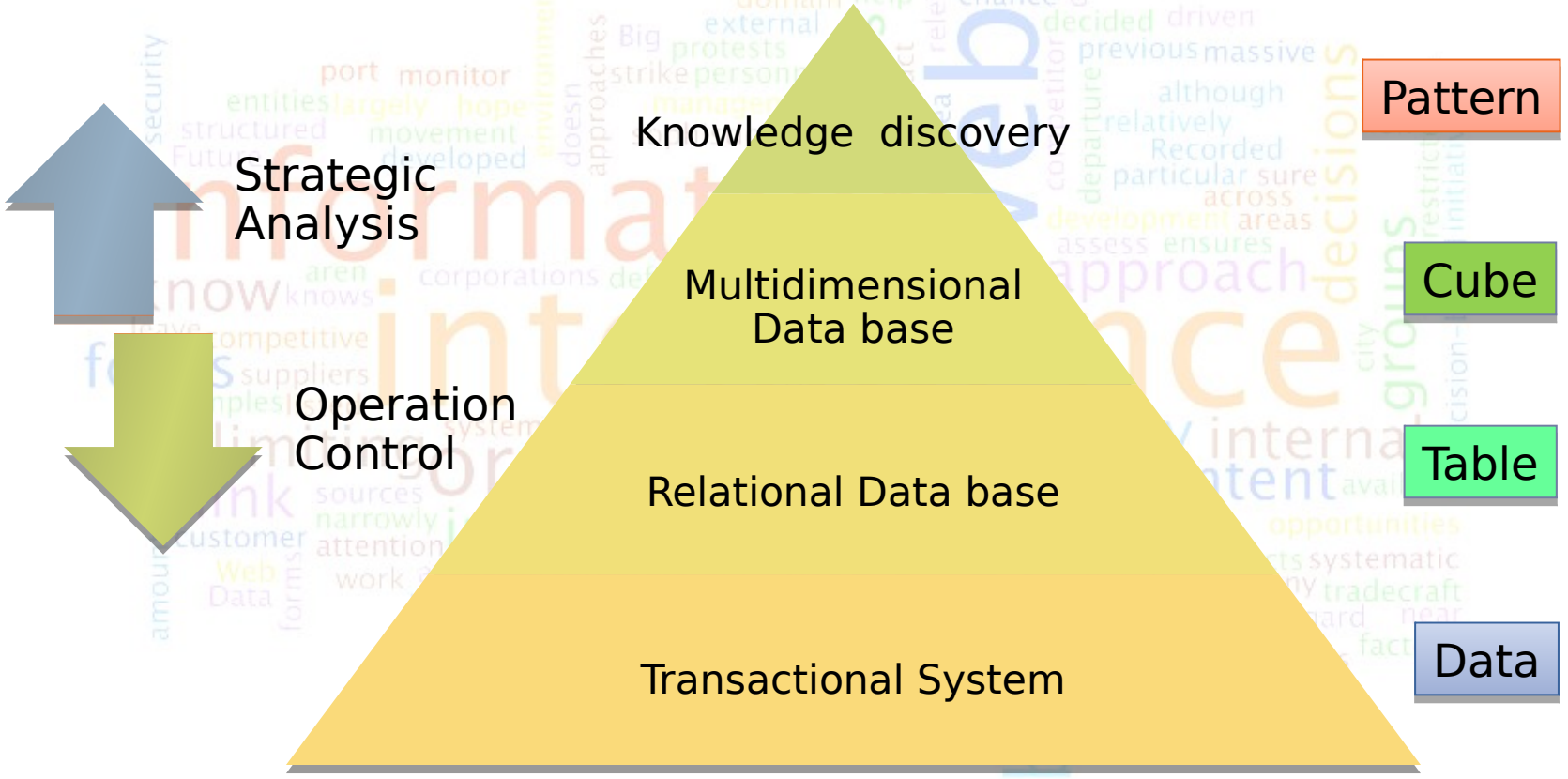
The KDD process

The **Knowledge Discovery in Databases (KDD) process** is commonly defined with the stages:

1) Selection
2) Preprocessing
3) Transformation
4) Data mining
5) Evaluation



The Knowledge Pyramid



Considerations

- Scientific Method:
  - Hypothesis-Experiment-Knowledge
- Knowledge Discovery:
  - Data-Hypothesis-Knowledge
- Expert support
- Space-Time dimension.
- Data Quality, Homogeneity

Asking the right question

Who is the right person?

Query. 1 Query. 2 Query. 3 Query. 4 Query. 5

Asking the right question (2)

- **First, have** a **clear objective**
- **Understand the goals** of the process.
- **Be Question oriented** to generate knowledge.
- We don't want another statistical report.

Asking the right question (3)

- Evaluating Point Of Sale
  - Which **POS** has **better/worst production level**?
  - Which **products** are **best selling** on some POS?
  - Which **promotions** has been with the **best impact** in sales.
- Evaluating promotion
  - Which **effect over the total sales** had the last year **marketing campaign**?
  - Has the **promotion** an **effective cost**?

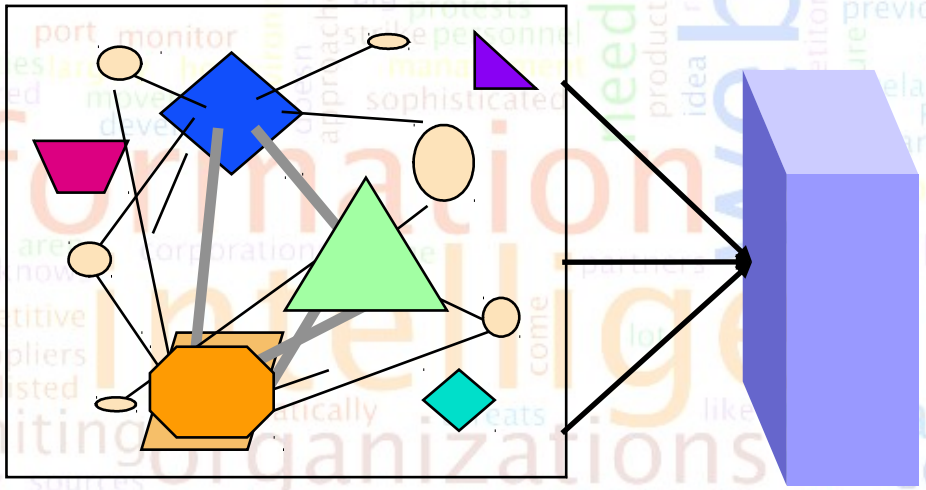
Asking the right question (4)

- Sales and tendencies
  - Which **product** has **top selling score** and which ones doesn't moves?
  - How much **utility** are perceived by **each product**?
  - How affected are the **pattern of purchases** by a **change in the price**?
  - How affected are the **popularity of a product** in relation to **the time**?
  - Which **special characteristics** has the **client** that **buy this product**?
  - Which are the **best selling brand**?

Asking the right question (5)

- The client preferences and sales
  - How many **man bought** this product?
  - How many **married women bought** in this store?
  - What is the **impact of the education** on the **family sales pattern**?
  - What is the impact the **change on the utility** by product on the **sales of men v/s women**?
- Inventory
  - Which warehouse had the **best usage ranking on special period**?
  - What is the **behaviour of the inventory** level of the warehouses?
  - What are the **most valuable warehouse**?

Step 1: Data Selection



Operational Data

Data Selection

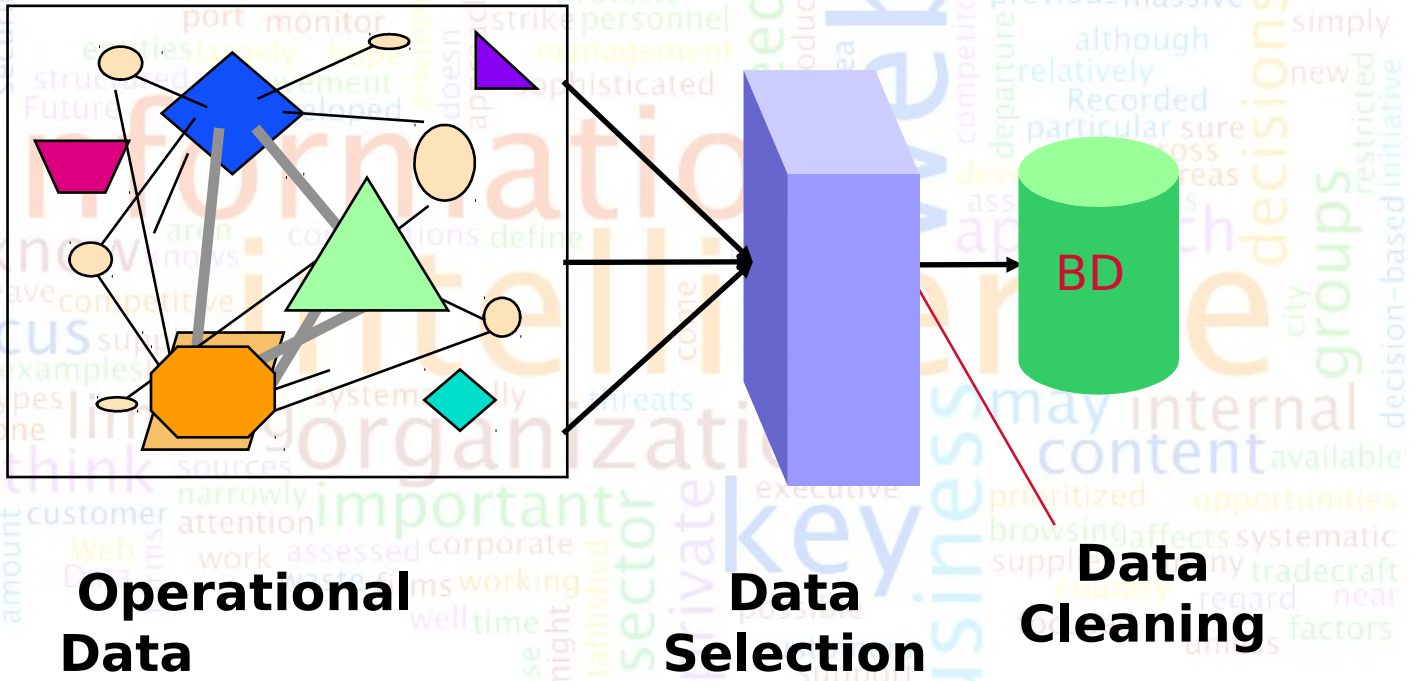
Step 1: Data Selection (2)

- All data in a Data Warehouse.
  - Huge amount of data, more time to process.
- Data Marts, segmenting the study to an operational sector.
  - More conventional amount of data, but less time to process.
  - Oriented to a more global strategy for the business.

Step 1: Data Selection (3)

- The **identification of real data sources** is an **important step** in the KDD process
- **Irrelevant data** (noise) often leads to **analytic errors**
- Different Data Formats introduce a **cost of interpretation/transformation**
  - **Metadata** allows us to **standardize** the data

Step 2: Preprocessing and Cleaning



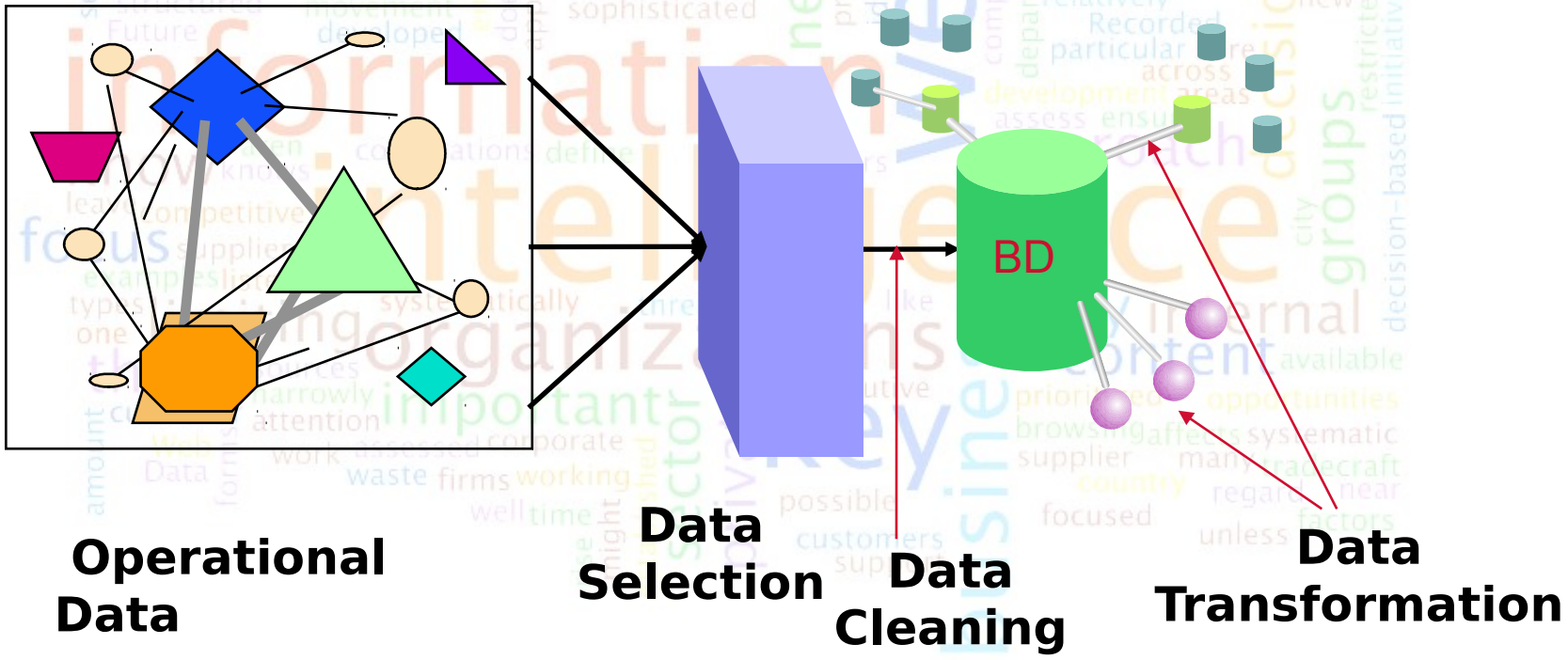
Step 2: Preprocessing and Cleaning (2)

- **Missing** Values
- **Dynamical** data
- **Distributed** and **big data** bases.
- **Noise**

Step 2: Preprocessing and Cleaning (3)

- **Data consistency**: Operational system are constructed on base of the **direct business requirements**. That means any other requirement on them (like KDD) have been never implemented and tested.
  - That imply **inconsistency**.
- **Data Manipulation Errors**: usually occur when **testing is avoided**. Example: Client with name "Batman" that remains from the development process.
- **Irrelevant Data**: Some data **need to be filtered** because is not part of the analysis

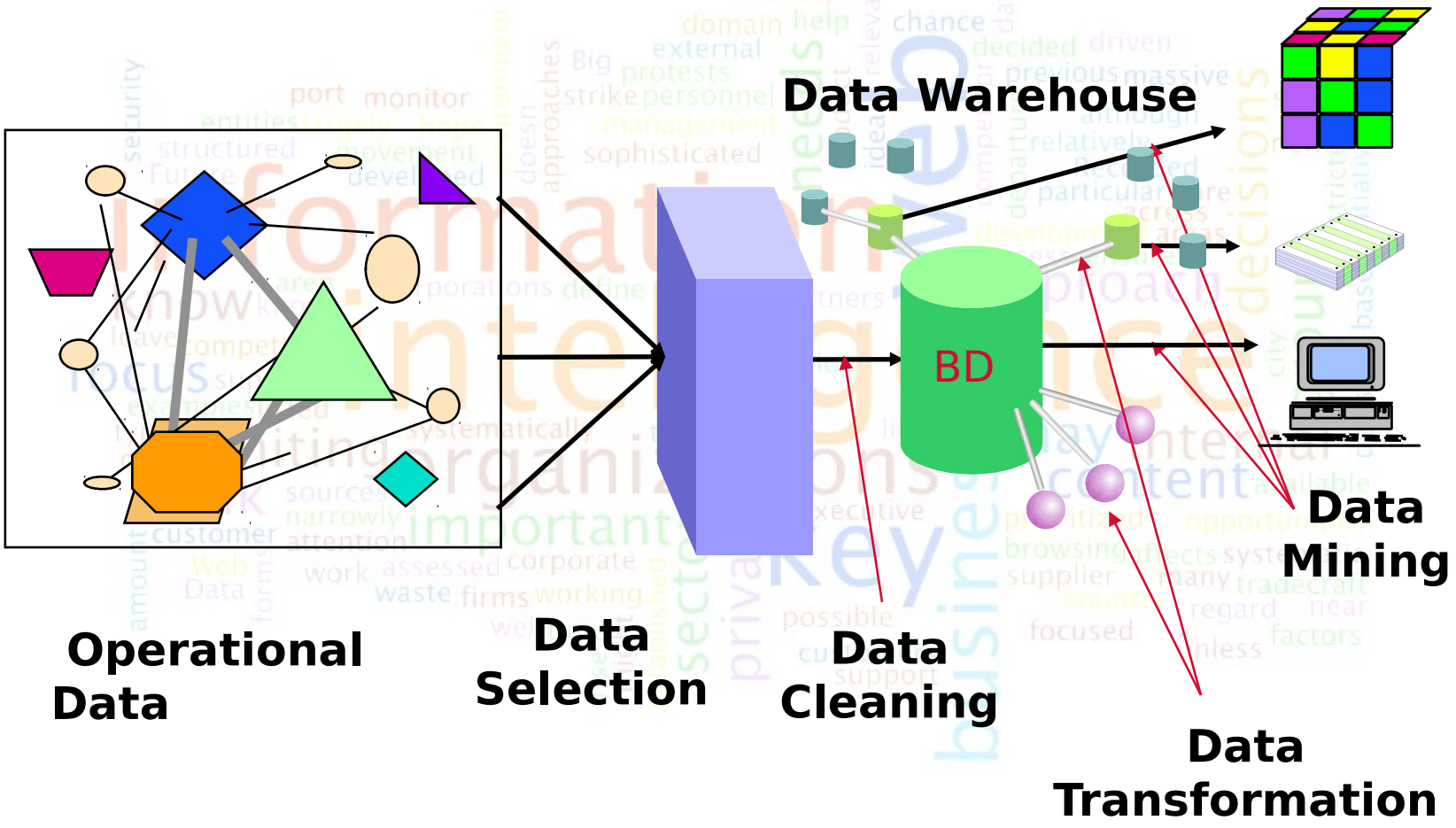
Step 3: Data Transformation



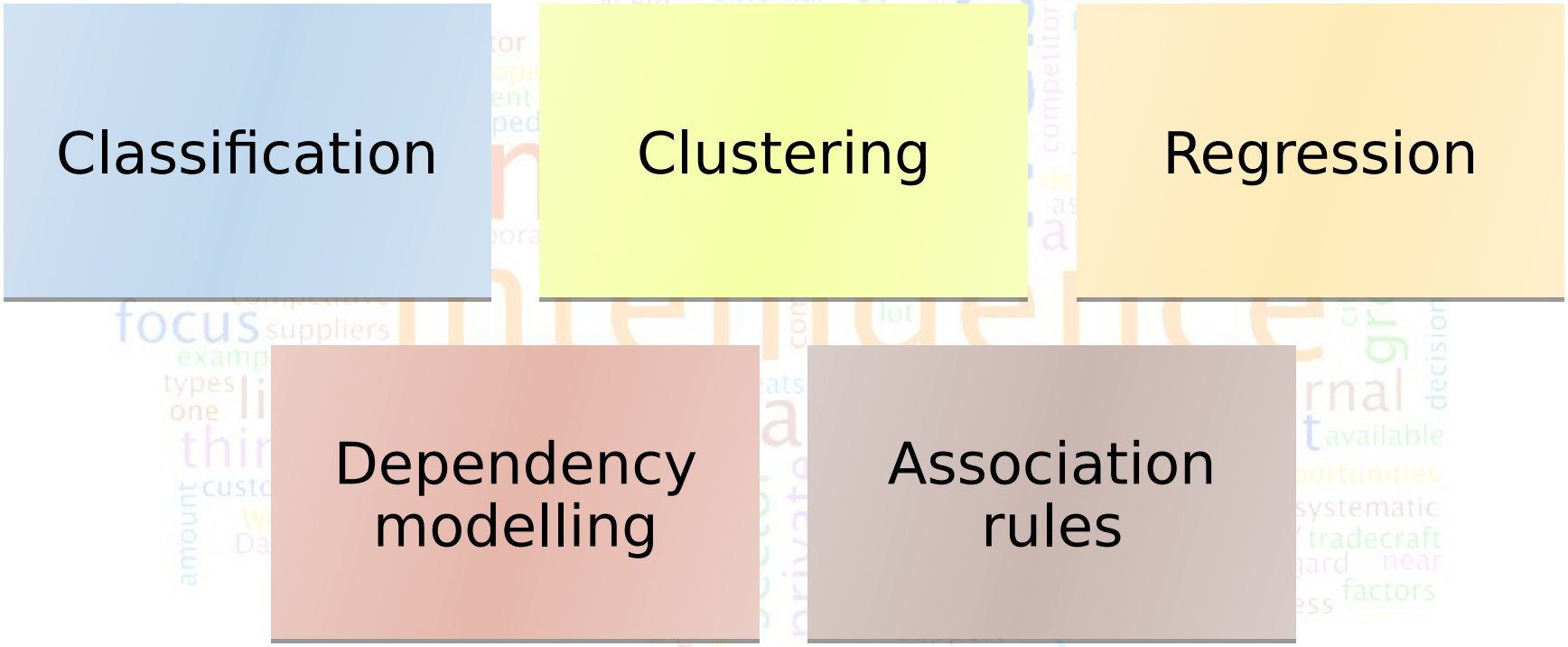
Step 3: Data Transformation (2)

- **Data to vectors**: Data mining algorithms require vectors in R^n.
  - Vector Space Model for text
- **Dimensionality reduction:** When data has too many dimensions, preprocessing can be done to reduce data dimensionality
  - Principal Component Analysis

Step 4: Data Mining



Step 4: Data Mining (2)



Step 4: Data Mining (3)
USING DATA MINING TO EXTRACT KNOWLEDGE

- ¿How to use the **prediction** for knowledge?
- ¿How to find the **rules implied** by the **generated models**?
- ¿How to **model human behavior**?

Step 4: Data Mining (3)
RULES: KNOWLEDGE FOR HUMANS

- Human usually understand better **knowledge expressed in the form of "Rules".**
- **EEUU legislation** doesn't **allow credit assignations** based in **black box predictors**.
- **Decision Trees** and **Bayesian Networks throw directly a set of rules**.
- **ANN are black box predictors** but recent discovery allows us to extract rules from them
  - V. Palade, S. Bumbaru, M.G. Negoita (1998). "A method for compiling neural networks into fuzzy rules using genetic algorithms and hierarchical approach", Proceedings of the 2nd IEEE International Conference on Knowledge-Based Intelligent Electronic Systems- KES1998, vol. 2 pp. 353-358, Adelaide - Australia, 1998.

Step 5: Interpreting and verifying results

- **Summarized critical factors**.
  - Observing its **impact** on the business and **try to explain** them.
- The **Expert** can identify the **knowledge**.
- **Store the knowledge generated**, **reuse** it on a future KDD process.

Conclusion

- **Tactical Value** of large amount of data.
- **Analysis capacities**,
  - finding useful knowledge v/s Cost and time.
- The **new knowledge MUST BE validated** by the **new data** in order to plan the business.
- The **expert must validate** each step in the process.
- **The correct interpretation** of the result will generate the **knowledge**
- **The new discovery must be stored** in a structure that allows **reuse in others KDD**.

Section 1.3

Web Data and Web Mining

The Web Data

- **The problem**: **Garbage-in, Garbage-out**
- **Web Data**
  - Highly Variable in **Type**
  - Highly Variable in **Format**
- HTML includes:
  - Tags
  - Text
  - Multimedia
- **Logs**: have an standard but the information that we want (sessions) is **not explicit**.
- **Web Sites** Change over time and **usually nobody track these changes**.

Web Data in the KDD process

- We require:
  - Pages transformed to **Feature Vectors** (*to be explained in future chapters*)
  - **Clean individual Session** from users.
  - The **Web site graph**
- The **web data cleaning** and **pre-processing** activities should store the result in an **information repository** for further **data mining process**.

What happens with web data?

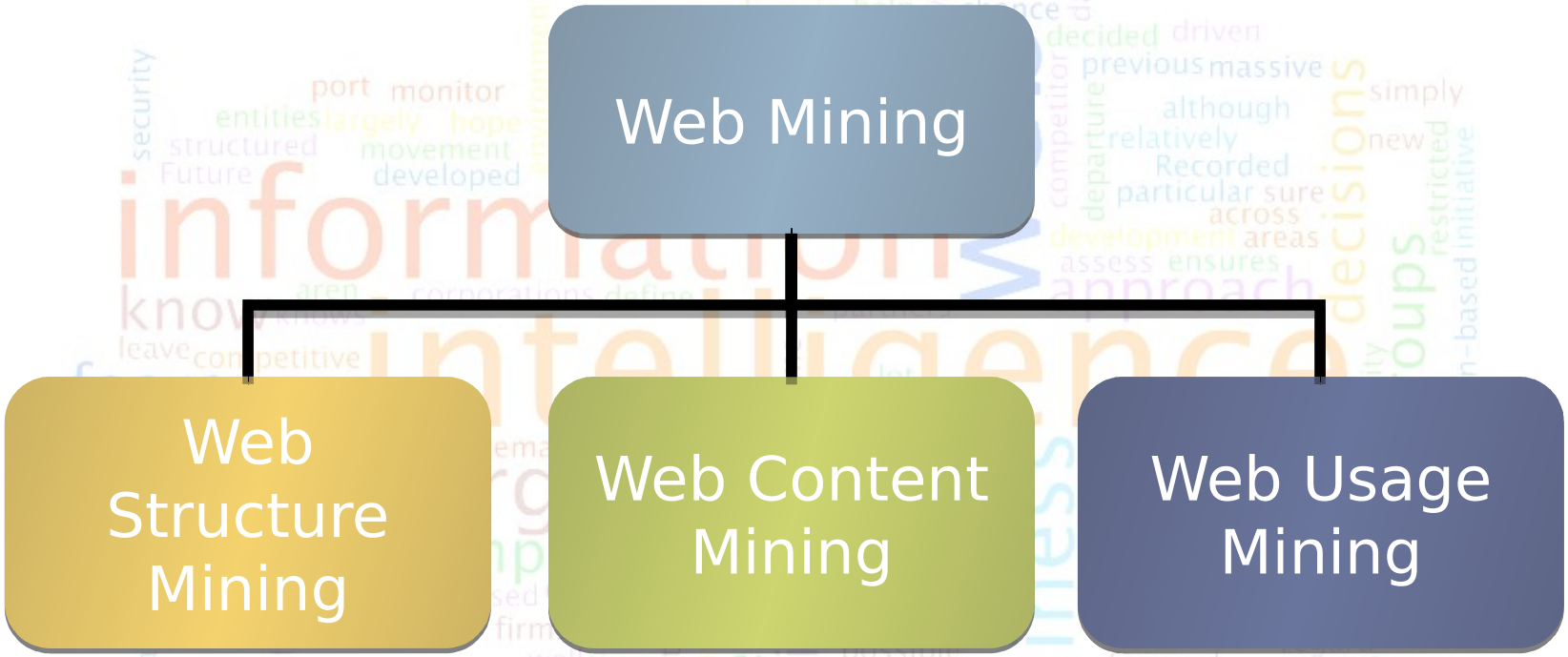
"The web is a huge collection of heterogeneous, unlabeled, distributed, time variant, semi-structured and high dimensional data".

S.K. Pal 2002

Mining the Web: Web Mining

- Web mining techniques are the **application of data mining** theory in order to **discovery patterns from web data**.
- Web mining usually considers three important steps:
  - Pre-processing
  - Pattern discovery
  - Pattern analysis

Web Mining Taxonomy [Jooshi00,Lu03]



Web Structure Mining (WSM)

- It deals with the **mining of the web hyperlink structure** (*inter document structure*).
- A **website is represented** by a **graph of its links**, within the site or between sites.
- Facts like the **popularity of a web page** can be studied, for instance, if a page is referred by a lot of other pages in the web.
- The web link structure allows to develop a **notion of hyperlinked communities**.
- It can be used by search engines, like GOOGLE or YAHOO, in order to get the **set of pages more cited for a particular subject**.

Web Structure Mining (2)

- To discover the **link structure of the hyperlinks** at the **inter-document level** to generate **structural summary about the Website and Web page**.
  - ***Direction 1***: Based on the hyperlinks, categorizing the Web pages and generated information.
  - ***Direction 2***: Discovering the structure of Web document itself.
  - ***Direction 3***: Discovering the nature of the hierarchy or network of hyperlinks in the Website of a particular domain.

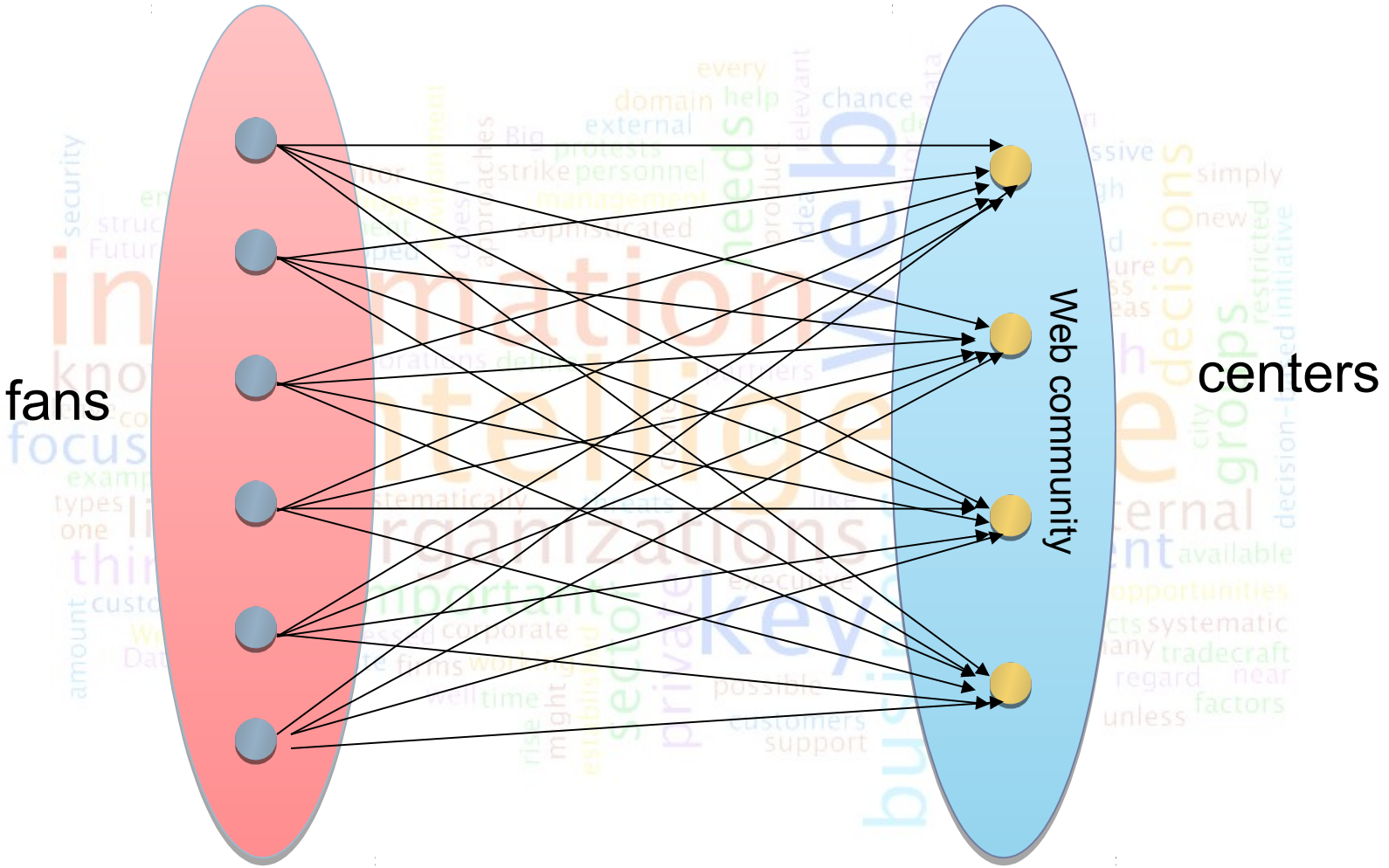
Web Structure Mining (3)

- Finding **authoritative web pages**
  - Retrieving pages that are not only relevant, but also of **high quality**, or **authoritative on the topic**
- Hyperlinks can infer the **notion of authority**
  - The Web consists not only of pages, but also of hyperlinks pointing from one page to another
  - These hyperlinks contain an enormous amount of **latent human annotation**
  - A hyperlink pointing to another web page, this can be considered as the **author's endorsement of the other page**

Web Structure Mining (4)

- Web pages **categorization**
  - (Chakrabarti, et al., 1998)
- Discovering **micro communities** on the Web
  - Example:
    - Clever system (Chakrabarti, et al., 1999)
    - Google (Brin and Page, 1998)
- **Schema Discovery** in Semi-structured Environment

Web Structure Mining: Example



Web Community Centers: many web pages go there

Web Content Mining (WCM)

- The goal is to find **useful information from the web content**.
- In this sense, *WCM is similar to Information Retrieval (IR)*.
- However, *web content is not only free text*, other objects like *pictures*, *sound* and *movies* belong also to the content.
- There are two main areas in WCM :
  - the **mining of document contents** (web page content mining)
  - **the improvement of content search** in tools like search engines (search result mining)

Web Content Mining (2)

- Web content
  - Text, image, audio, video, metadata and hyperlinks
- Information Retrieval View (Structured + Semi-Structured)
  - Assist / Improve information finding
  - Filtering Information to users on user profiles
- Database View
  - Model Data on the Web
  - Integrate them for more sophisticated queries

Web Content Mining (3)

- Mining **multimedia data**
  - Mining **image** from satellite
    - (Fayyad, et al. 1996)
  - Mining **image to identify small volcanoes on Venus**
    - (Smyth, et al 1996)

Issues in Web Content Mining

- Developing intelligent tools for IR
  - Finding **keywords** and **key phrases**
  - Discovering **grammatical rules** and collocations
  - **Hypertext classification/categorization**
  - Extracting **key phrases** from text documents
  - Learning **extraction models/rules**
  - **Hierarchical clustering**
  - Predicting **(words) relationship**

Web Usage Mining (WUM)

- Also known as **Web log mining**
  - Mining techniques to **discover interesting usage patterns** from the secondary data derived from the **interactions of the users while surfing the Web.**

Web Usage Mining: Considerations

- WUM applies traditional data mining methods in order to **analyze usage data**.
- The **sessionization process** is necessary to **correct the problems** detected in the data.
- The goal is to discover **patterns in usage data** applying different kinds of data mining techniques.
- Applications of WUM can be grouped in two main categories:
  - **User modelling in adaptive interfaces**, known as personalization.
  - **User navigation patterns**, in order to improve the *web site structure*.

Web Usage Mining: Applications

- Target potential customers for electronic commerce
- Enhance the quality and delivery of Internet information services to the end user
- Improve Web server system performance
- Identify potential prime advertisement locations
- Facilitates personalization/adaptive sites
- Improve site design
- Fraud/intrusion detection
- Predict user's actions (allows pre-fetching)

Web Usage Mining in the economy

- **Association rules and ANNs**:
  - **predicting next web page to be visited** in a session path.
- **Virtual Shopping**
  - prediction are useful for
    - **Displaying desired information**
    - Direct **hyperlink to similar product** preferred by others
    - Etc.

Web Usage Mining in the economy (2)

- **Clustering techniques**
  - Segment into *group of web users*.
- **Marketing**
  - *opinion poll* or *market sample survey*
    - market segmentation
- We could **predict the behavior of users** in a **same cluster**.

Section 1.4

Web Intelligence

What is Web Intelligence?

- **Web intelligence** is the area of study and research of the application of artificial intelligence and information technology on the web in order to create the next generation of products, services and frameworks based on the Internet.
- These include systems, services, amongst other activities, all of which are carried out by the Web Intelligence Consortium.

The 2013 IEEE/WIC/ACM International Conference on Web Intelligence

A little history on Web Intelligence…

- Since the late 1999's, many new algorithms, methods and techniques were developed and used extracting both knowledge and wisdom from the data originating from the Web.
- In this context, the term "Web Intelligence" was born in a paper written by Ning Zhong, Jiming Liu Yao and Y.Y.Ohsuga in the Computer Software and Applications Conference in 2000.

Research Fields

- Research about Web Intelligence covers many fields as data mining, information retrieval, semantic web and web data warehousing and Adaptive web sites. Different techniques and technologies have been used by researchers and practitioners over the years.
  - Web information repositories
  - Web user behavior analysis
  - Web content and structure mining
  - Social Network Analysis
  - The Semantic Web
  - Knowledge Discovery from Databases
  - Knowledge Representation