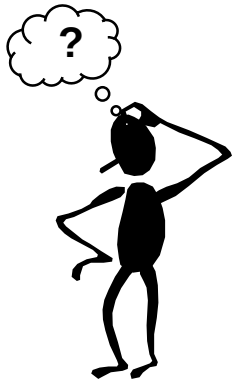
Internet jako wielka składnica danych

- **Olbrzymie zasoby informacji**
  - ***różna struktura***
  - ***różna jakość***
- **Łatwość dostępu do informacji**
  (*wystarczy mieć przeglądarke...*)
- **Łatwość wprowadzania własnych danych**
  (*wystarczy mieć serwer WWW...*)

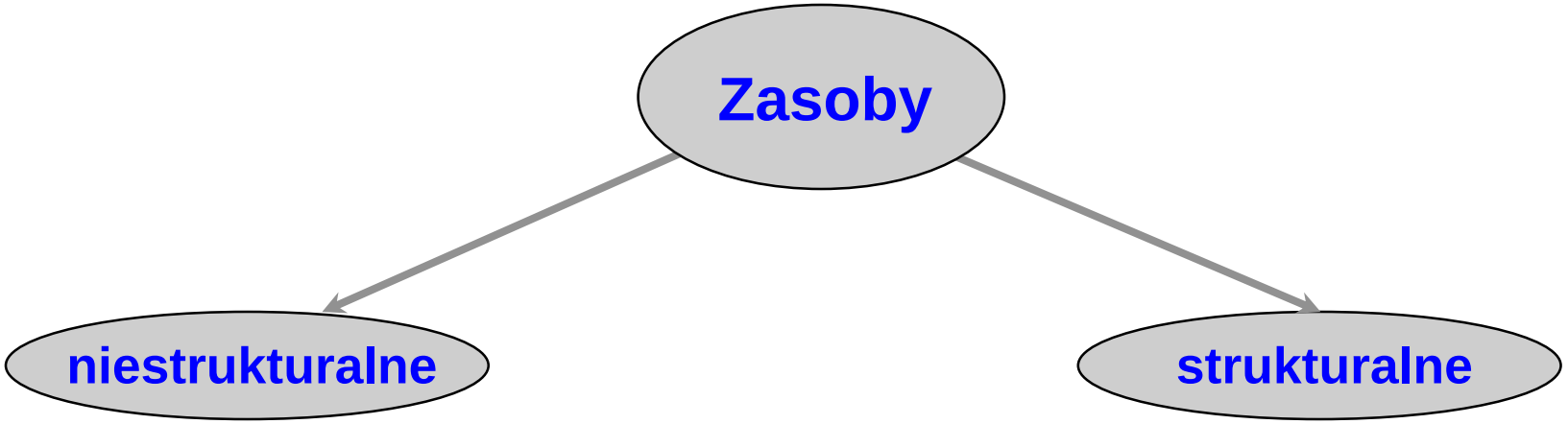
Zalety czy wady ?



Problemy

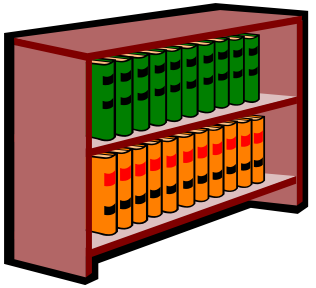
- **Jak znaleźć potrzebne informacje?**
  - ***odkrywanie źródeł informacji***
- **Jak odróżnić informacje użyteczne od zbędnych?**
  - ***filtrowanie informacji***
- **Gdzie gromadzić informacje?**
  - ***Web warehousing***
- **Jak wykorzystać pozyskane informacje?**
  - ***przetwarzanie analityczne (OLAP)***

Klasyfikacja zasobów danych



nie istnieje odpowiadający im *schemat danych*

istnieje ściśle zdefiniowany *schemat danych*



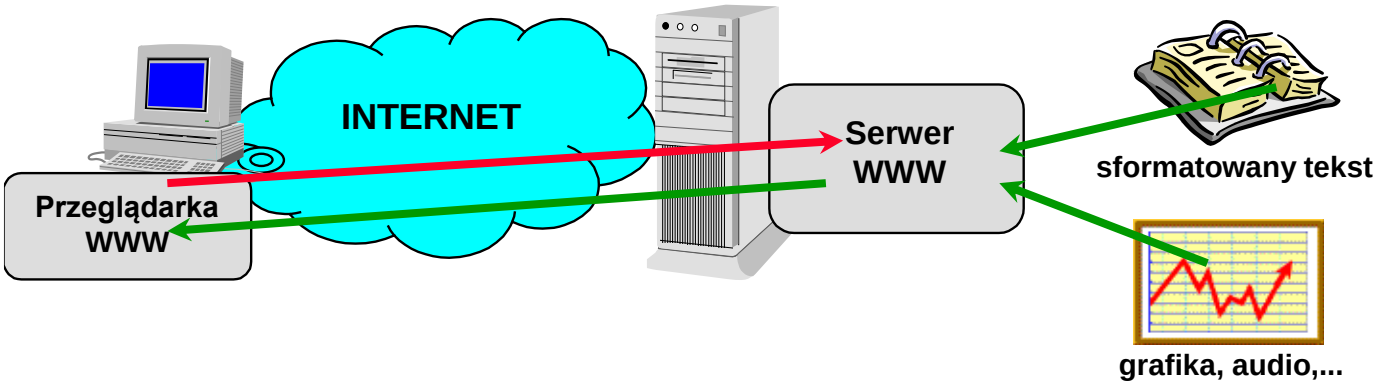
Zasoby niestrukturalne

Statyczne strony (pliki) HTML zawierające

- tekst
- grafikę
- dane audio  wideo  .
- odsyłacze do innych stron/plików

Identyfikowane adresem URL (*Uniform Resource Locator*):

protokół // *komputer*.*domena* / *plik*



Zasoby strukturalne

Dane zgromadzone w bazach danych dostępnych poprzez Internet

Schemat danych + Dane

Dostęp poprzez adres URL rozszerzony o parametry zapytania:

protokół // *komputer*.*domena* / *procedura* ? *p1*=*w1*&*p2*=*w2* ...

***Dynamiczna* strona HTML**

Zasoby strukturalne

Przykład:

http://quote.yahoo.com/q?s=intc&d=t

Views: Basic - DayWatch - Performance - Real-time Mkt - **Detailed** - [Create New View]

INTEL CORP (NasdaqNM:INTC) - Trade: Choose Brokerage

Last Trade	Change		Prev Cls	Open	Volume
Jun 5 · **28.18**	+0.68 (+2.47%)		27.50	27.60	41,657,800
Day's Range	Bid	Ask	P/E	Mkt Cap	Avg Vol
26.89 - 28.20	28.12	28.18	105.77	188.4B	43,997,045
52-wk Range	Bid Size	Ask Size	P/S	Div/Shr	Div Date
18.96 - 36.78	2,000	700	6.90	0.08	Jun 1
1y Target Est	EPS (ttm)	EPS Est	PEG	Yield	Ex-Div
41.38	0.26	0.70	2.18	0.29	May 3

Small: 1d 5d **1y** none
Big: 1d 5d 3m 6m 1y 2y 5y max

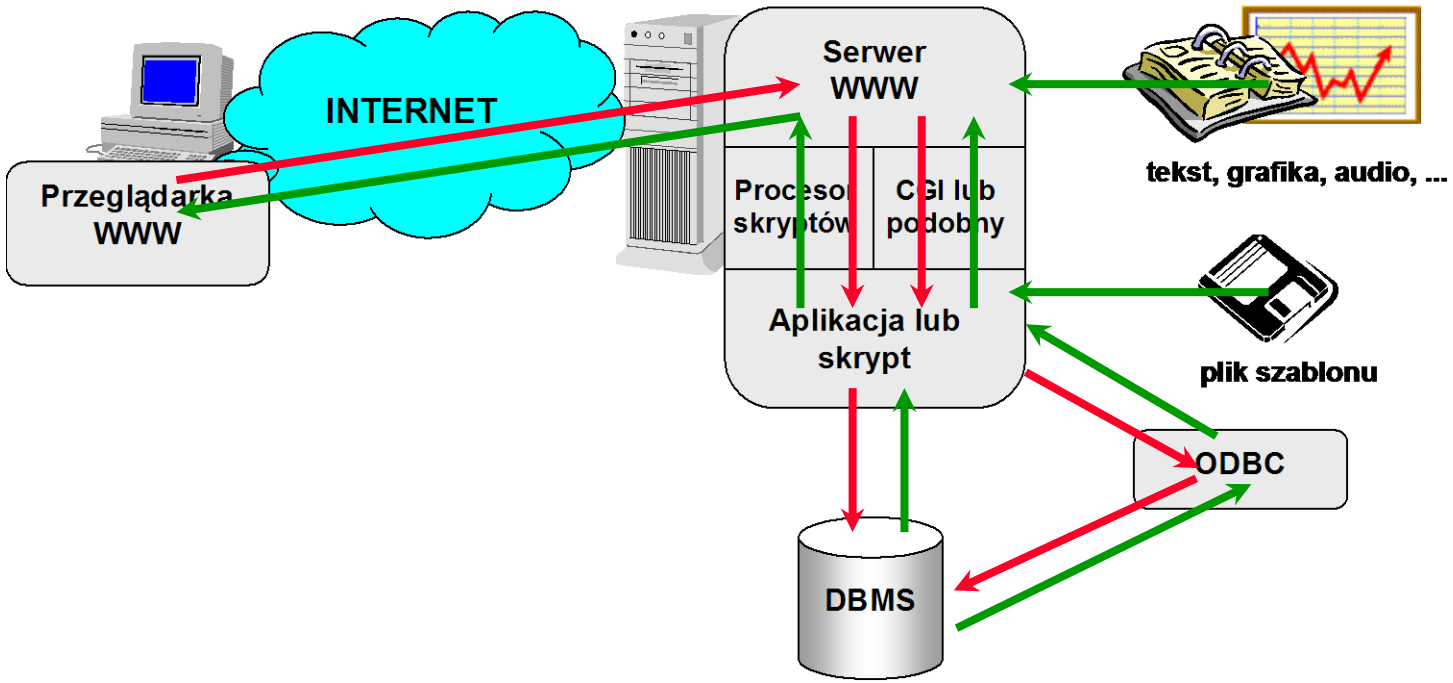
Chart, Financials, Historical Prices, Insider, Messages, News, Options
Profile, Reports, Research, SEC Filings, Upgrades, **more...**

Compare 22 long-distance plans at a glance. Yahoo! Savings Finder.

Add to My Portfolio - Set Alert

Download Spreadsheet

Zasoby strukturalne



Zasoby semistrukturalne

Dane semistrukturalne:

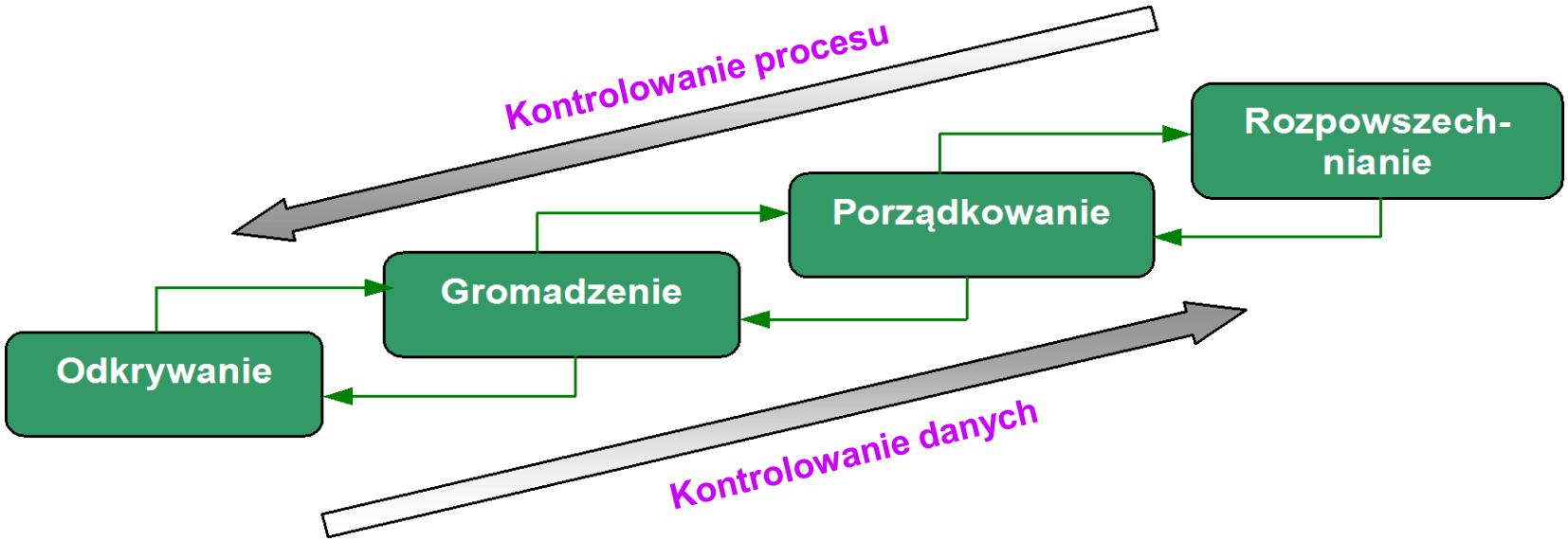
- o nieznanym (lub znanym częściowo) schemacie danych
- o strukturze nieregularnej
  - pola opcjonalne
  - alternatywne kombinacje pól
  - pola o nieokreślonej liczności
- niekompletne
- często zawierające błędy i niespójności

Język do opisu i zapisu tego typu danych:

eXtensible Markup Language (XML)

Web Farming

Metoda systematycznego pozyskiwania i wykorzystywania zasobów informacyjnych Internetu



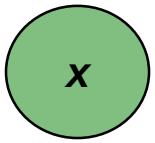
Odkrywanie

Poszukiwanie w Internecie miejsc istotnych z punktu widzenia prowadzonej działalności biznesowej

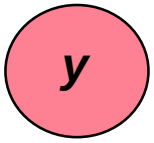
- Etap najtrudniejszy w całym procesie (trudny do zautomatyzowania)
- Narzędzia:
  - przeglądarki
  - wyszukiwarki
  - agenci

Przygotowanie roli pod uprawę

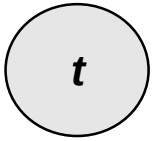
Trafność a precyzja odkrywania



miejsca istotne



miejsca odkryte

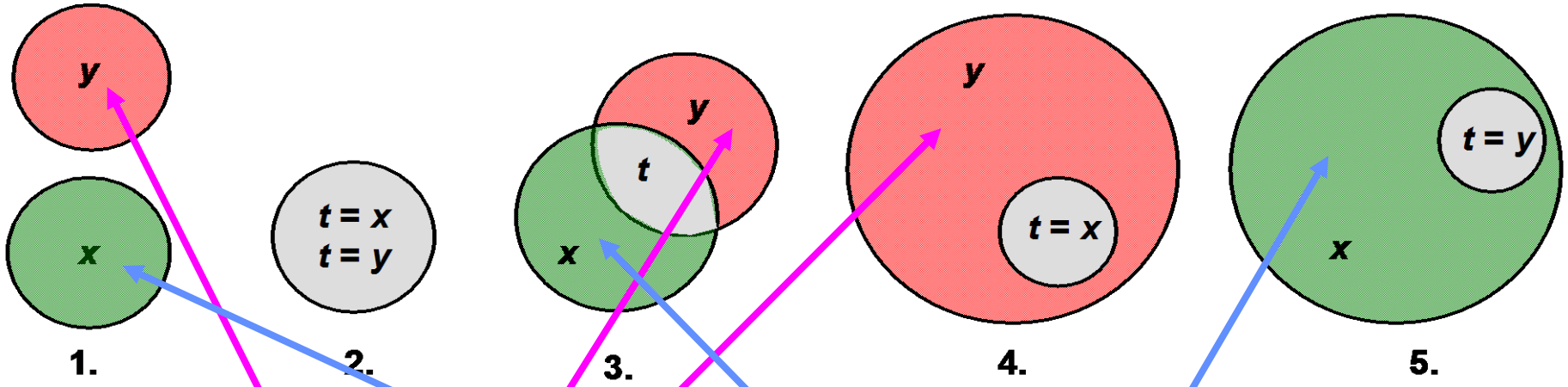


odkryte miejsca istotne

Precyzja $P = t / y$ [0..1] (*precision*)

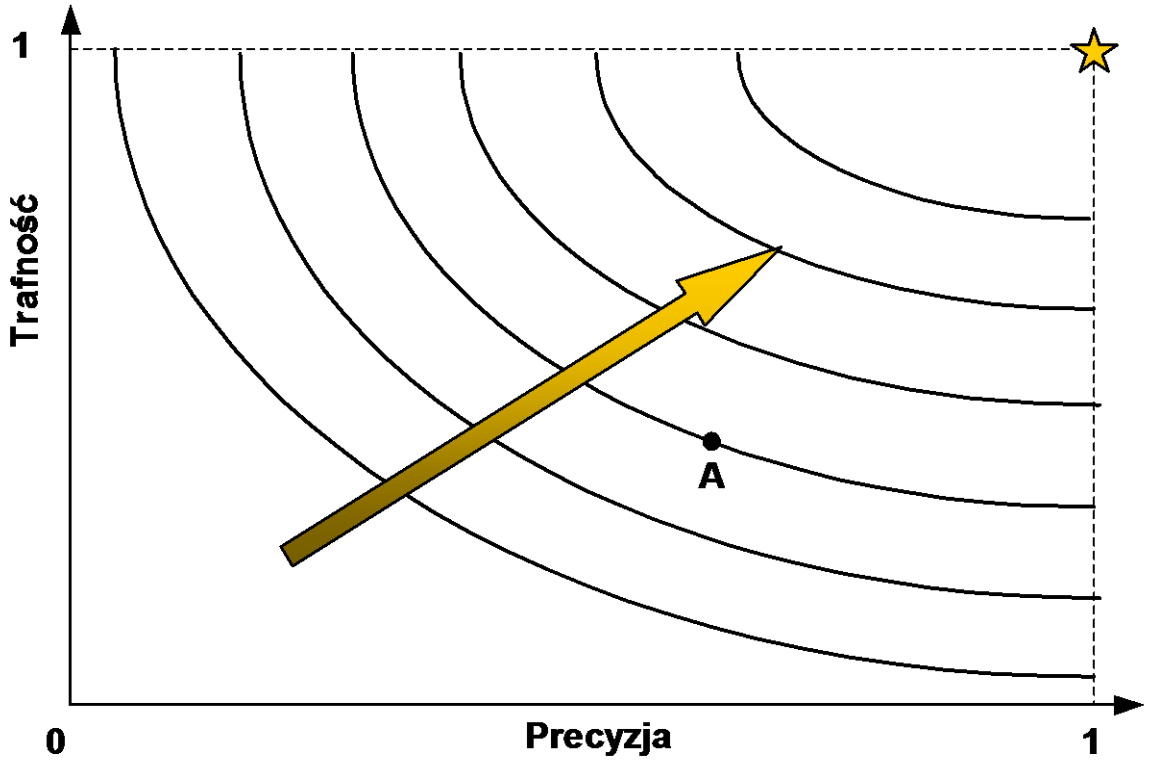
Trafność $T = t / x$ [0..1] (*recall*)

Trafność a precyzja odkrywania



Efektywność odkrywania

Efektywność = Trafność * Precyzja



Miara-F = 2 * Trafność * Precyzja / (Trafność + Precyzja)

- średnia harmoniczna trafności i precyzji

Gromadzenie

Zbieranie treści statycznych i dynamicznych stron WWW zidentyfikowanych na etapie odkrywania

- Podlega automatyzacji (głównie w odniesieniu do stron dynamicznych)
- Narzędzia:
  - agenty programowe
  - narzędzia systemu plików
  - bazy danych

Obsiewanie roli

Porządkowanie

Nadanie danym struktury stosownej do ich przechowywania i przetwarzania w hurtowni danych

- Operacje:
  - sprawdzanie danych
  - czyszczenie danych
  - transformowanie danych
- Narzędzia:
  - bazy danych i hurtownie danych
  - narzędzia do analizowania danych (*business intelligence*)
  - narzędzia do eksploracji danych

Pielęgnowanie zasiewów

Rozpowszechnianie

Dostarczanie uporządkowanych danych do miejsc, w których są potrzebne (czyli do ich konsumentów)

- Bezpośrednie przekazanie danych albo przekazanie wyników analiz

- Narzędzia:
  - bazy danych i hurtownie danych
  - narzędzia do analizowania danych (*business intelligence*)
  - narzędzia do eksploracji danych
  - narzędzia do prezentacji danych

Żniwa!