Systemy ekspertowe

Część druga

Od bazy danych do bazy wiedzy

Krótkie wprowadzenie do zagadnień eksploracji danych

Autor
Roman Simiński

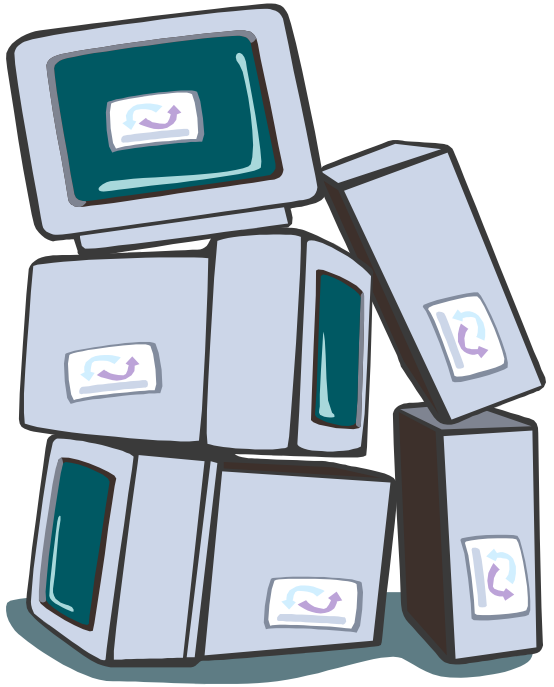
Kontakt
siminski@us.edu.pl
www.us.edu.pl/~siminski

Korporacyjne bazy danych końca XX i początku XXI wieku

Fakty

- Rozmiar baz danych współczesnych systemów informatycznych osiąga wielkości rzędu *terabajtów*.
- Średniej wielkości hipermarket rejestruje dziennie sprzedaż przynajmniej *kilkunastu tysięcy produktów*.
- Puchną bazy danych systemów *e-commerce*, dostępnych na bieżąco, 24 godziny na dobę – wzrasta liczba ich *klientów* oraz liczba *zawieranych transakcji*.

Aktualne możliwości analizowania i rozumienia dużych zbiorów danych są dużo mniejsze od możliwości ich zbierania i przechowywania.



Korporacyjne bazy danych końca XX i początku XXI wieku

Jednocześnie

- Konkurencja pomiędzy firmami zaostrza się.
- Coraz trudniej znaleźć nowe obszary ekspansji, nisze rynkowe.
- Coraz trudniej utrzymać dotychczasowych klientów.
- Bazy danych zawierają ogromne ilości użytecznych informacji, pozwalających firmom utrzymać lub wzmocnić ich pozycje rynkową.

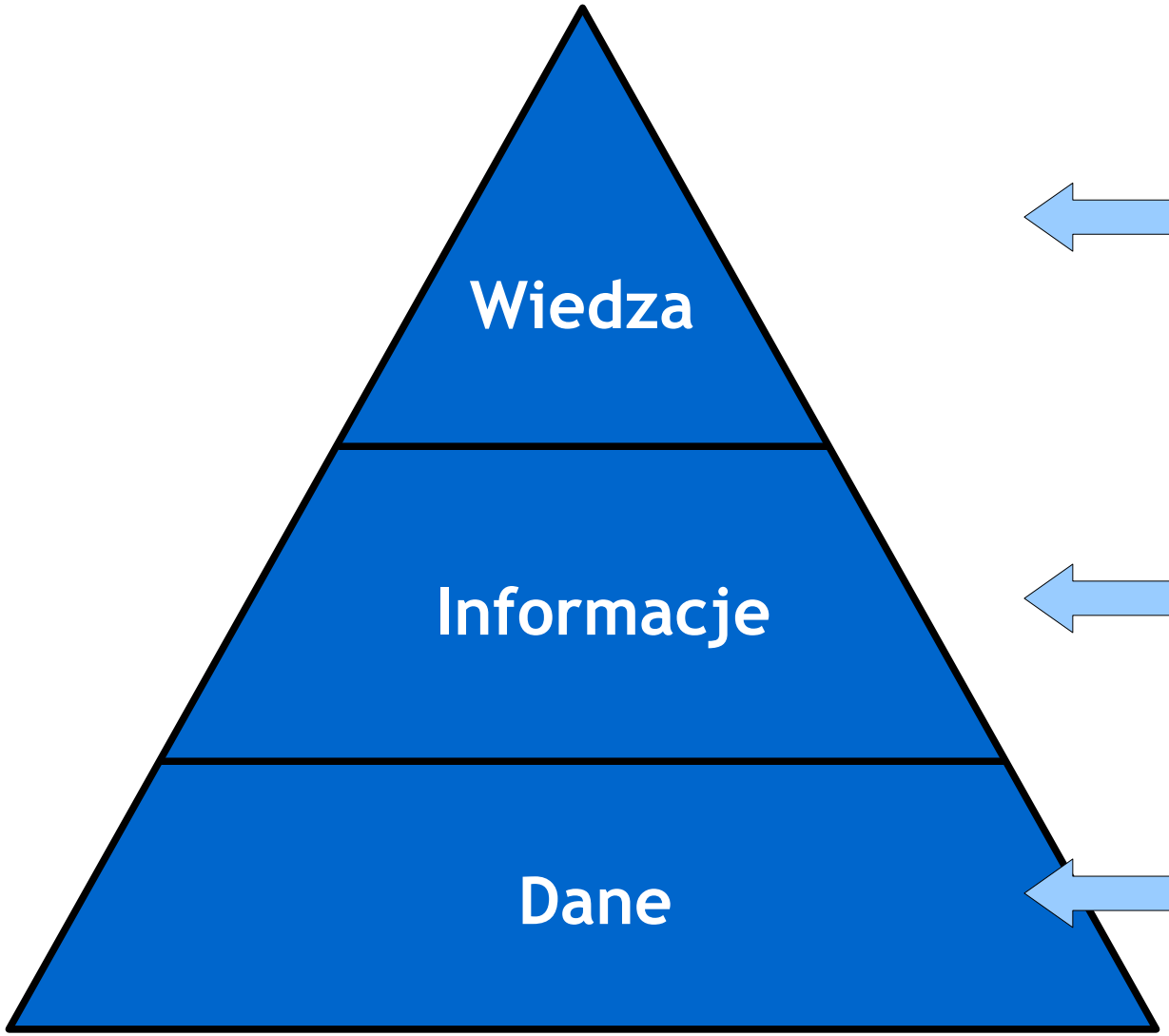
Korporacyjne bazy danych końca XX i początku XXI wieku

Korporacyjne bazy danych kopalnią użytecznych informacji

- Użyteczne informacje są wyrażone niejawnie, są *ukryte w danych*, należy je *odkryć*, *wydobyć*.
- Proces ten nazywa się potocznie *eksploracją danych*.
- Świadomość istnienia ukrytego potencjału informacyjnego baz danych jest znana od lat.
- Jednak dopiero w ciągu ostatnich kilkunastu lat intensywnie prowadzi się badania nad odkrywaniem metod eksploracji danych oraz wykorzystuje się te metody w praktyce.

Korporacyjne bazy danych kopalnią użytecznych informacji

Jakiego rodzaju informacje można wydobyć z danych?



Uogólnione informacje opisujące związki, relacje, prawidłowości występujące w danych. Np. zależność pomiędzy średnią sprzedażą w danym asortymencie a czasem przechowywania surowców na magazynie.

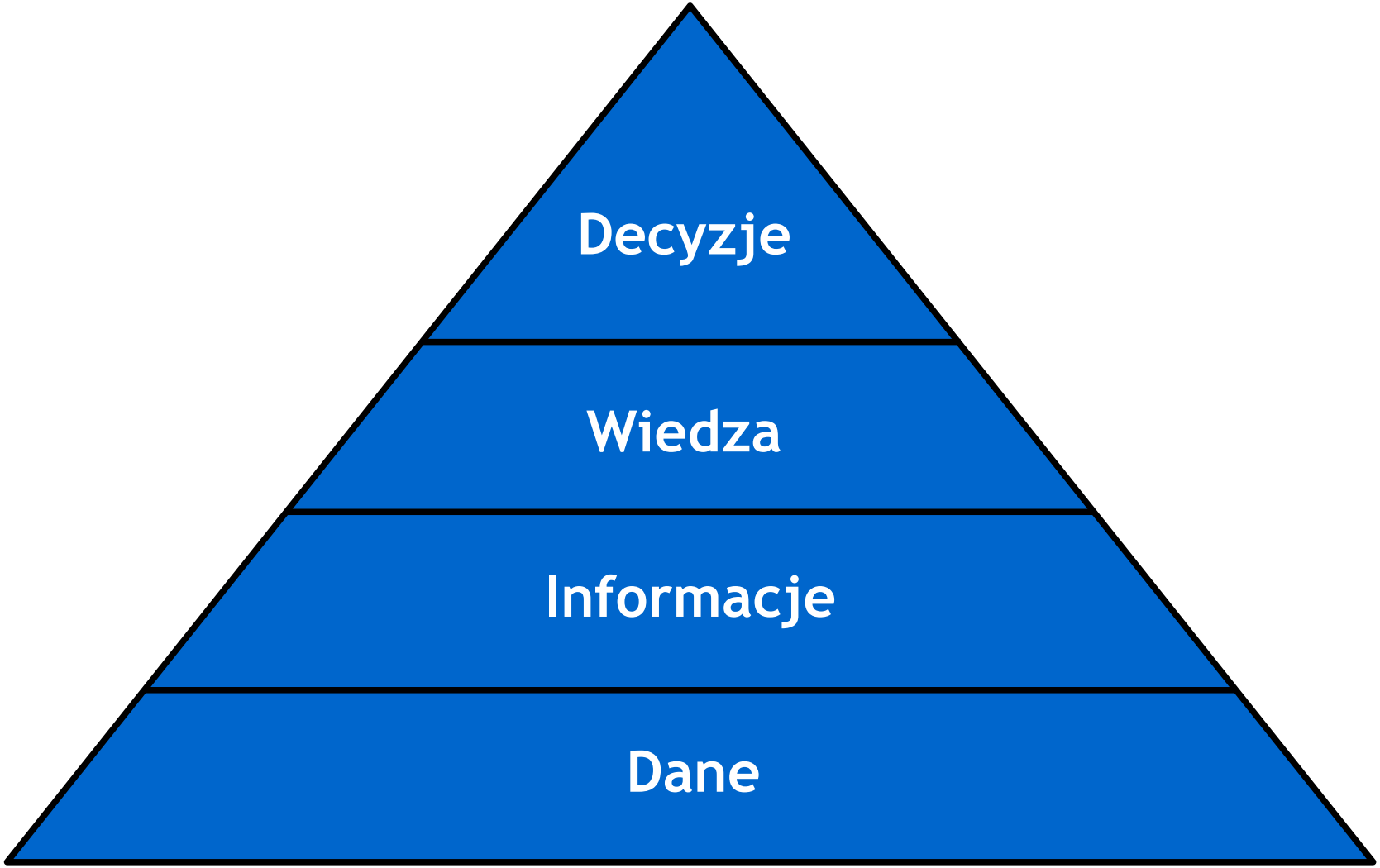
Dane zagregowane, wyznaczone (wyliczone) na podstawie zawartości baz danych – np. średnia sprzedaż w danym asortymencie towaru w zadanym okresie czasu.

Dane odczytywane bezpośrednio z baz danych – np. cena zakupu, stan magazynowy, dane kontrahenta, itp.

W ekonomii taki schemat nazywa się *Business Intelligence*

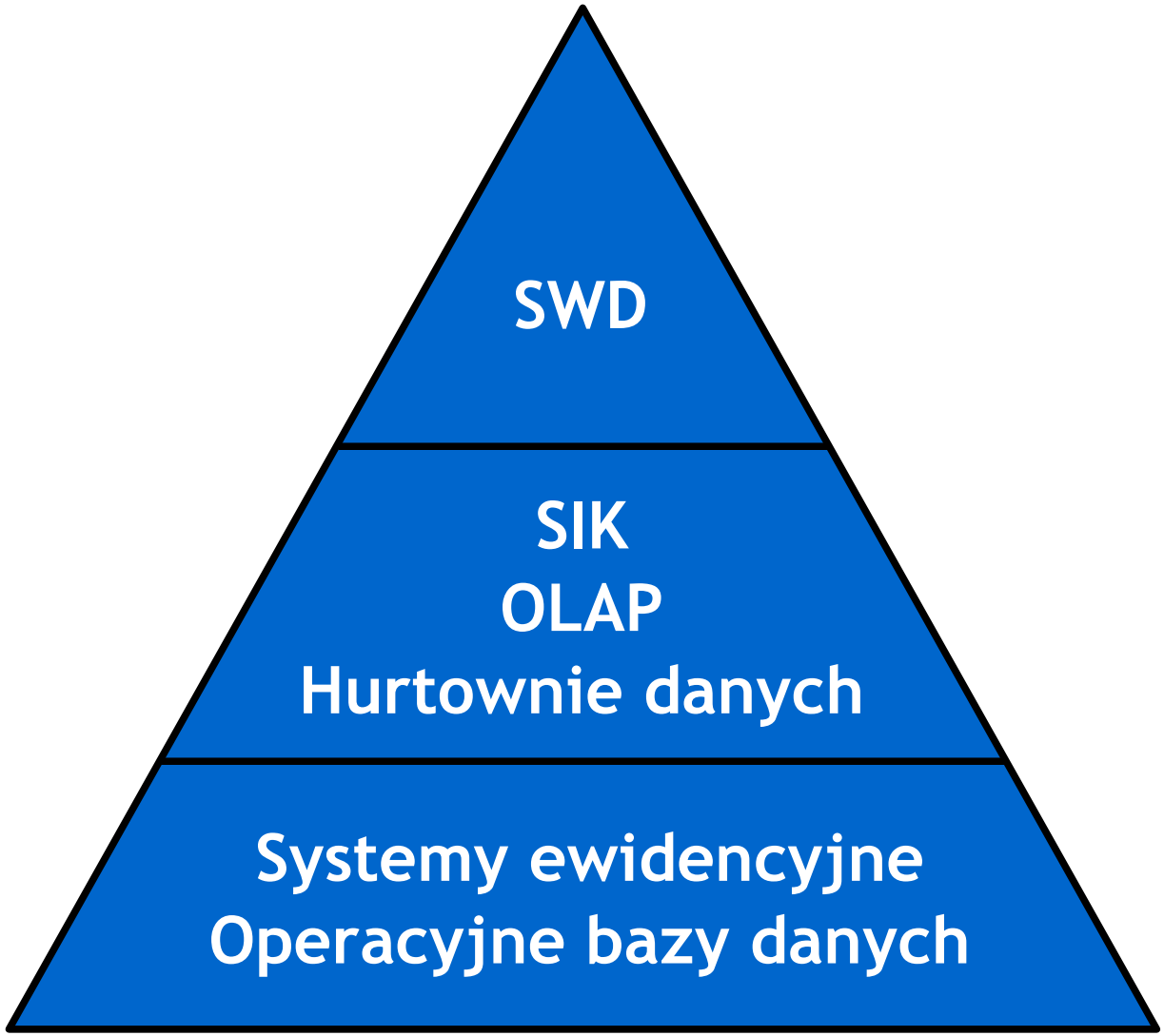
Korporacyjne bazy danych kopalnią użytecznych informacji

Z biznesowego punktu widzenia wiedza ma wspomagać podejmowanie decyzji



Korporacyjne bazy danych kopalnią użytecznych informacji

Informacje a systemy je wykorzystujące



SWD – Systemy Wspomagania Decyzji, ang. DSS – Decision Support Systems.

SIK – Systemy Informowania Kierownictwa, ang. EIS, MIS – Management Information Systems.

OLAP – Przetwarzanie Analityczne Online, od ang. Online Analytical Processing,

Hurtownie danych, magazyny Danych od ang. Data Warehouses.

Korporacyjne bazy danych kopalnią użytecznych informacji

Problemy na styku OLAP a wspomaganie decyzji

- Systemy OLAP działają zwykle obliczając zagregowane wielkości na podstawie danych pochodzących z magazynu danych.
- Systemu OLAP pozwalają na analizowanie tego co się wydarzyło na podstawie danych o przeszłości.
- Działanie OLAP jest sterowane hipotezą sformułowaną przez użytkownika (ang. *query-driven eksploration*), system OLAP jest pasywny.
- Używając systemów OLAP można wchodzić w głąb, dochodząc do danych bardziej szczegółowych, ale użytkownik nadal pozostaje odpowiedzialny za identyfikowanie interesujących trendów czy powiązań.
- Koncepcje postrzegania danych jako „wielowymiarowych kostek” powoduje problemy w percepcji przeprowadzanych analiz.

Do skutecznego podejmowania decyzji potrzebna jest wiedza o prawidłowościach rządzących daną dziedziną. Decydenci oczekują, iż systemy informatyczne prawidłowości te odkryją, potwierdzając to, co już wiemy lub dostarczą nam nowej wiedzy.

Eksploracja danych jako rozwijająca się dziedzina informatyki

Koncepcja Data Mining - Eksploracja Danych

Istnieje wiele definicji koncepcji Eksploracji danych (Data Mining):

- *Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.*

David Hand, Heikki Mannila, and Padhraic Smyth, Principles of Data Mining, MIT Press, Cambridge, MA, 2001.

- *The non-trivial extraction of novel, implicit, and actionable knowledge from large datasets.*

Graham Williams, Markus Hegland and Stephen Roberts, A Data Mining Tutorial, The Second IASTED International Conference on Parallel and Distributed Computing and Networks, PDCN'98.

Eksploracja danych jako rozwijająca się dziedzina informatyki

Koncepcja Data Mining - Eksploracja Danych, cd.

Inne definicje koncepcji Eksploracji danych (Data Mining):

- *Data mining is an interdisciplinary field bringing togther techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases.*

Peter Cabena, Pablo Hadjinian, Rolf Stadler, JaapVerhees, and Alessandro Zanasi, Discovering Data Mining: From Concept to Implementation, Prentice Hall, Upper Saddle River, NJ, 1998.

- *Technology to enable data exploration, data analysis, and data visualisation of very large databases at a high level of abstraction, without a specific hypothesis in mind.*

Graham Williams, Markus Hegland and Stephen Roberts, A Data Mining Tutorial, The Second IASTED International Conference on Parallel and Distributed Computing and Networks, PDCN'98.

Eksploracja danych jako rozwijająca się dziedzina informatyki

Koncepcja Data Mining - Eksploracja Danych, cd.

Dla potrzeb tych zajęć przyjmijmy:

- *Data mining – eksploracja danych – jest dziedziną informatyki zajmującą się odkrywaniem wiedzy zapisanej niejawnie w dużych zbiorach danych oraz przedstawieniem jej w zrozumiały dla użytkownika sposób.*
- Pod pojęciem *wiedzy* rozumieć będziemy *relacje*, *powiązania*, *związki* i *wzorce* odkrywane przez algorytmy eksploracji danych w sposób autonomiczny.

Eksploracja danych (DM – *Data Mining*) określana jest również pojęciem *odkrywania wiedzy w bazach danych* (KDD – *Knowledge Discovery in Databases*) .

Eksploracja danych jako rozwijająca się dziedzina informatyki

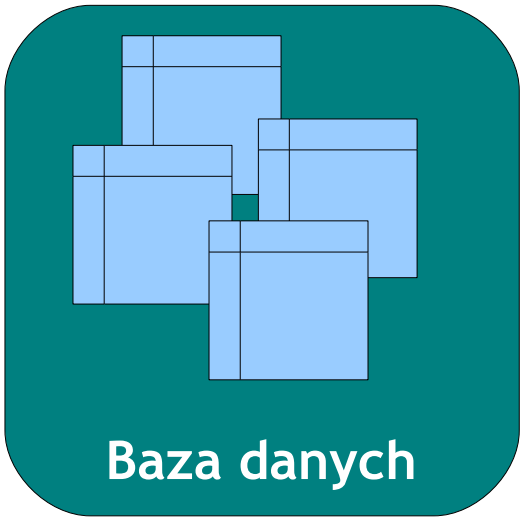
Różne definicje eksploracji danych - różne metody jej realizacji

Odkrywanie wiedzy w bazach danych realizowane jest najróżniejszymi metodami:

- Drzewa decyzyjne,
- Sztuczne sieci neuronowe,
- Analiza skupień,
- Techniki maszynowego uczenia,
- Metody statystyczne,
- . . .

Eksploracja danych jako rozwijająca się dziedzina informatyki

Różne metody realizacji - wspólny cel



Baza danych

Eksploracja danych

Relacje
Związki
Powiązania
Wzorce
Sekwencje

Baza wiedzy

Jak zapisać wiedzę ?

Dane o sprzedaży

Dane o klientach

Dane o pacjentach

itp.,itd.

Jak zapisać wiedzę odkrywaną w bazach danych?

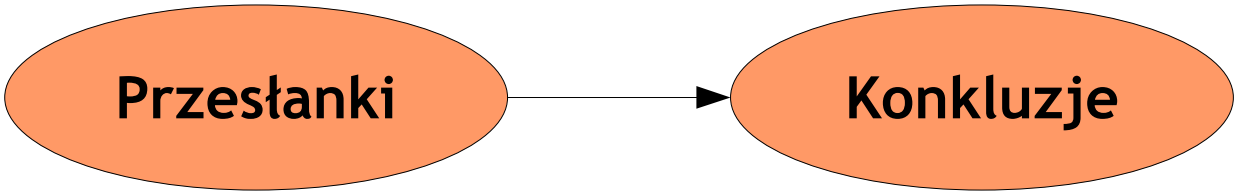
Inżynieria wiedzy dostarcza różnych metod reprezentacji wiedzy

- Drzewa decyzyjne,
- Tablice decyzyjne,
- Reguły,
- Ramy,
- Sieci semantyczne,
- Scenariusze,
- Sieci Bayesa,
- . . .

Regułowa reprezentacja wiedzy

Reguły są intuicyjnie najprostszą metodą reprezentacji wiedzy

Istnieje wiele formatów zapisu reguł. Koncepcja jest jednak zwykle ta sama:



„Działanie" reguły odbywa się według wywodzącego się z logiki schematu *modus ponens*:

$$\frac{p \rightarrow q \quad p}{q}$$

Jeżeli p implikuje logicznie q oraz p jest prawdziwe to q jest również prawdziwe.

Regułowa reprezentacja wiedzy

Reguły można zapisywać w różny sposób

- Z wykorzystaniem symbolu implikacji:

  przesłanki $\rightarrow$ *konkluzje*

- W postaci zbliżonej do instrukcji warunkowej:

  if *przesłanki* then *konkluzje*

- W postaci odwrotnie zapisanej instrukcji warunkowej:

  konkluzje if *przesłanki*

Regułowa reprezentacja wiedzy

Jak zapisywać przesłanki i konkluzje?

- Z wykorzystaniem zmiennych zdaniowych:

  p – procesor się przegrzewa

  q – sprawdź układ chłodzenia

  $$p \rightarrow q$$

- Z wykorzystaniem predykatów:

  $P(x)$ – procesor komputera x się przegrzewa

  $Q(x)$ – sprawdź układ chłodzenia komputera x

  $$P(x) \rightarrow Q(x)$$

- Z wykorzystaniem dwójek *atrybut-wartość*:

  if *stan_procesora = przegrzany*
  then *akcja_serwisowa = sprawdź_układ_chłodzenia*

Istnieją oczywiście inne formy zapisu literałów reguł.

Regułowa reprezentacja wiedzy

Reguły mogą przyjmować bardziej złożoną postać:

W inżynierii wiedzy wykorzystuje się standaryzowane formaty reguł wywodzące się z logiki:

- Koniunkcyjna postać normalna

$$p_1 \wedge p_2 \wedge \ldots \wedge p_n \rightarrow q_1 \vee q_2 \vee \ \ldots \vee \ q_m$$

- Klauzula Horna

$$p_1 \wedge p_2 \wedge \ldots \wedge p_n \rightarrow q$$

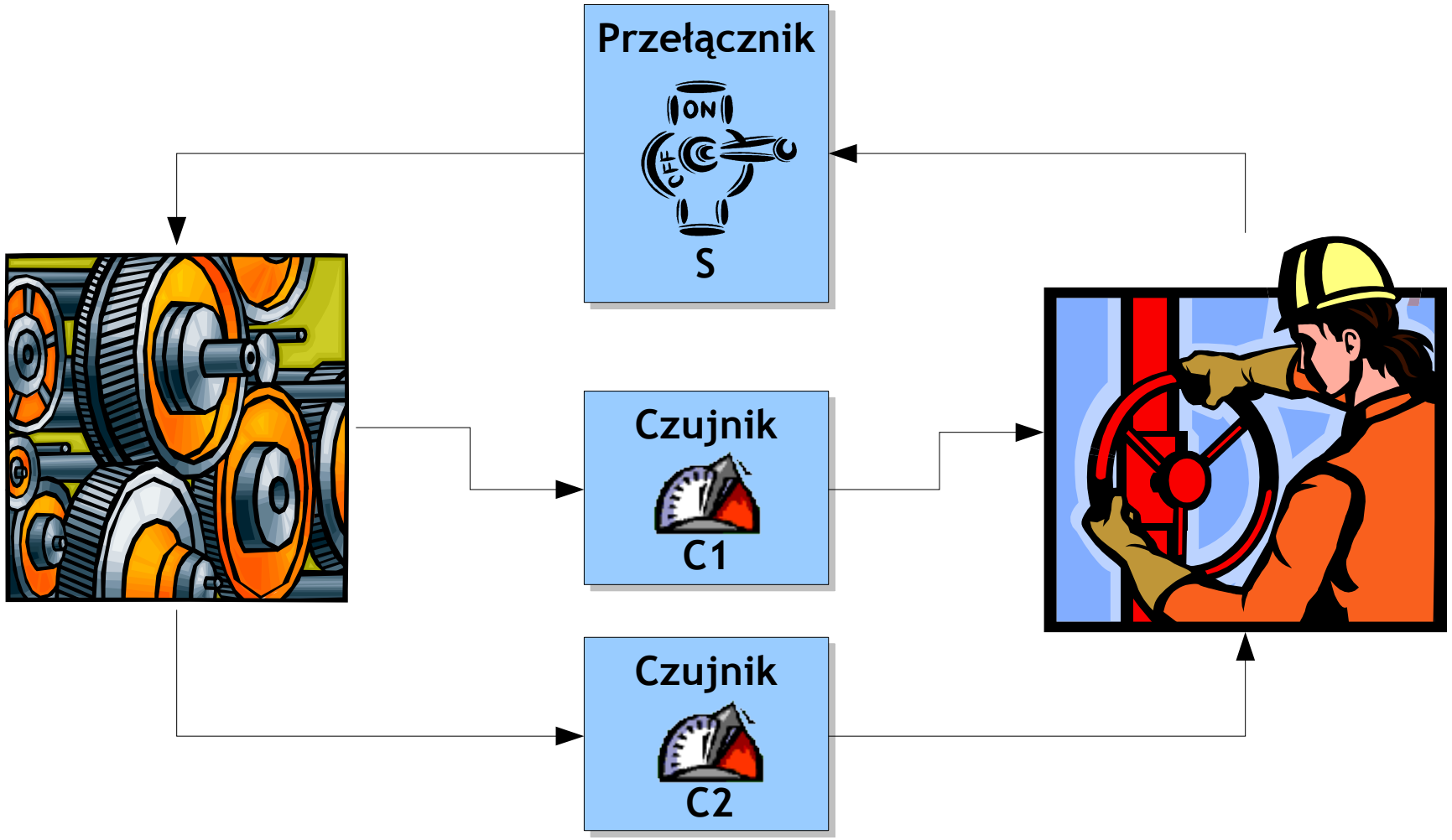
Najczęściej wykorzystuje się reguły posiadające postać klauzuli Horna

Istnieje wiele innych, czasem bardzo „dziwnych” postaci reguł – obowiązujących głównie w obrębie określonych systemów narzędziowych.

Jak odkrywać w danych wiedzę zapisaną w postaci reguł?

Prosty przykład

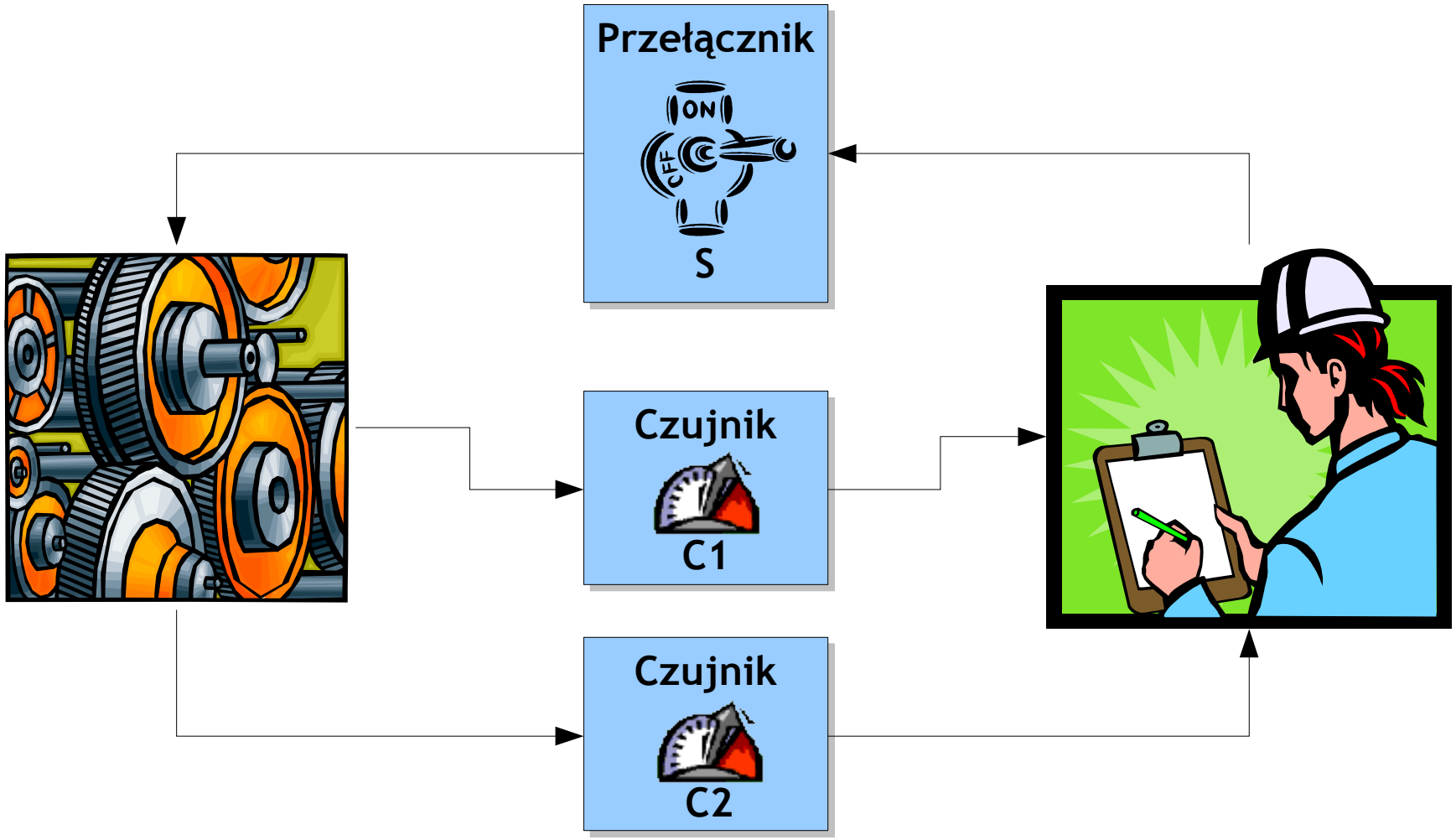
Pracę pewnego urządzenia monitorują czujniki C1 i C2. W zależności od ich wskazań pracownik obsługi ręcznie ustala stan przełącznika S, regulującego pewien parametr technologiczny. Chcemy przejść na komputerowe sterowanie włącznikiem S.



Jak odkrywać w danych wiedzę zapisaną w postaci reguł?

Prosty przykład, cd.

Należy wyznaczyć reguły sterowania przełącznikiem S. Przy jakich wartościach czujników C1 i C2 przełącznik ma być włączony a przy jakich wyłączony?

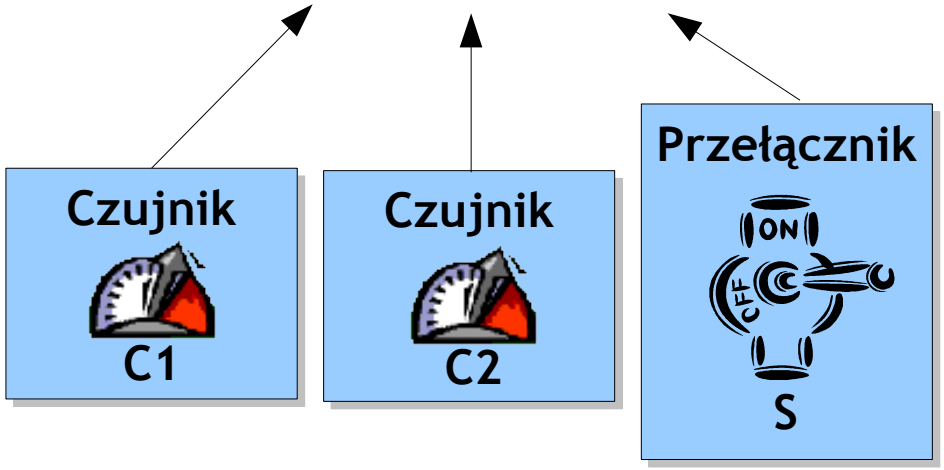


Jak odkrywać w danych wiedzę zapisaną w postaci reguł?

Prosty przykład,cd.

Tabela z danymi o tym, jaki stan przyjmuje przełącznik S przy określonych wartościach czujników C1 i C2.

	C1	C2	S
1	1	Niski	Off
2	1	Wysoki	On
3	2	Niski	Off
4	2	Wysoki	On
5	3	Niski	On
6	3	Wysoki	On
7	4	Niski	On
8	4	Wysoki	On

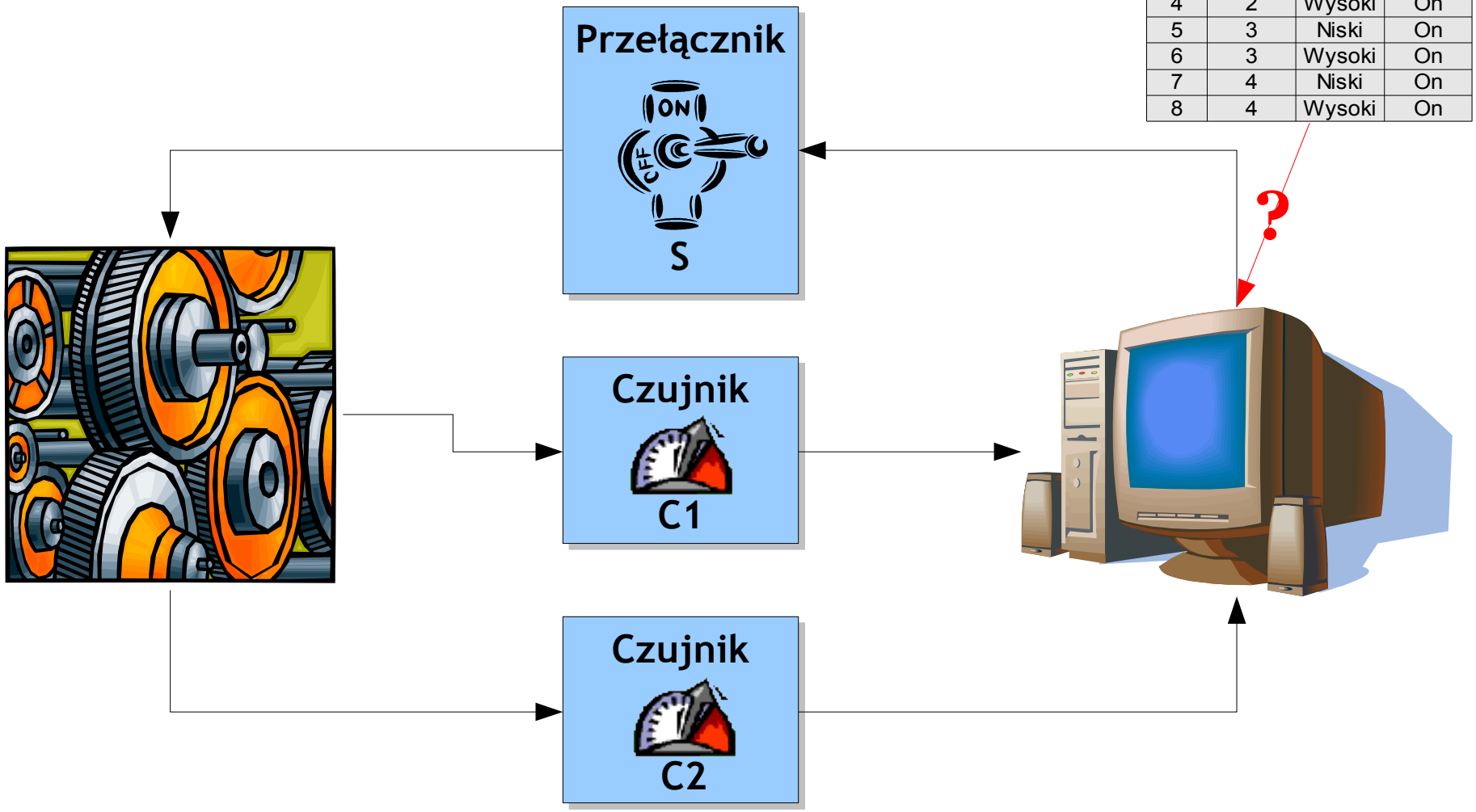


Jak odkrywać w danych wiedzę zapisaną w postaci reguł?

Prosty przykład, cd.

Przechodzimy na sterowanie komputerowe. W jaki sposób określić reguły działania programu sterującego?

	C1	C2	S
1	1	Niski	Off
2	1	Wysoki	On
3	2	Niski	Off
4	2	Wysoki	On
5	3	Niski	On
6	3	Wysoki	On
7	4	Niski	On
8	4	Wysoki	On



Jak odkrywać w danych wiedzę zapisaną w postaci reguł?

Prosty przykład, cd.

Pola C1 i C2 tabeli określają warunki a pole S określa decyzję odnośnie stanu przełącznika.

	C1	C2	S
1	1	Niski	Off
2	1	Wysoki	On
3	2	Niski	Off
4	2	Wysoki	On
5	3	Niski	On
6	3	Wysoki	On
7	4	Niski	On
8	4	Wysoki	On

Informacje wejściowe, warunki

Decyzja na temat stanu przełącznika

Jak odkrywać w danych wiedzę zapisaną w postaci reguł?

Prosty przykład, cd.

Po wyodrębnieniu atrybutów warunkowych i decyzyjnych taka tabela staje się *tablicą decyzyjną*.

Z tablicy decyzyjnej można próbować bezpośrednio odczytywać reguły.

	C1	C2	S
1	1	Niski	Off
2	1	Wysoki	On
3	2	Niski	Off
4	2	Wysoki	On
5	3	Niski	On
6	3	Wysoki	On
7	4	Niski	On
8	4	Wysoki	On

```
if C1=1 ∧ C2=Niski    then S=Off
if C1=1 ∧ C2=Wysoki   then S=On
if C1=2 ∧ C2=Niski    then S=Off
if C1=2 ∧ C2=Wysoki   then S=On
if C1=3 ∧ C2=Niski    then S=On
if C1=3 ∧ C2=Wysoki   then S=On
if C1=4 ∧ C2=Niski    then S=On
if C1=4 ∧ C2=Wysoki   then S=On
```

Jak odkrywać w danych wiedzę zapisaną w postaci reguł?

Prosty przykład, cd.

Osiem rekordów produkuje osiem reguł... . A jeżeli rekordów będzie kilkadziesiąt tysięcy? Kto potrzebuje wiedzy w postaci kilkudziesięciu tysięcy reguł!

	C1	C2	S
1	1	Niski	Off
2	1	Wysoki	On
3	2	Niski	Off
4	2	Wysoki	On
5	3	Niski	On
6	3	Wysoki	On
7	4	Niski	On
8	4	Wysoki	On

Można zauważyć, że:

`if C2=Wysoki then S=On`

niezależnie od wartości pola C1.

Można zauważyć, że:

`if C1>=3 then S=On`

niezależnie od wartości pola C2.

Można zauważyć, że:

`if C1<=2 ∧ C2=Niski then S=Off`

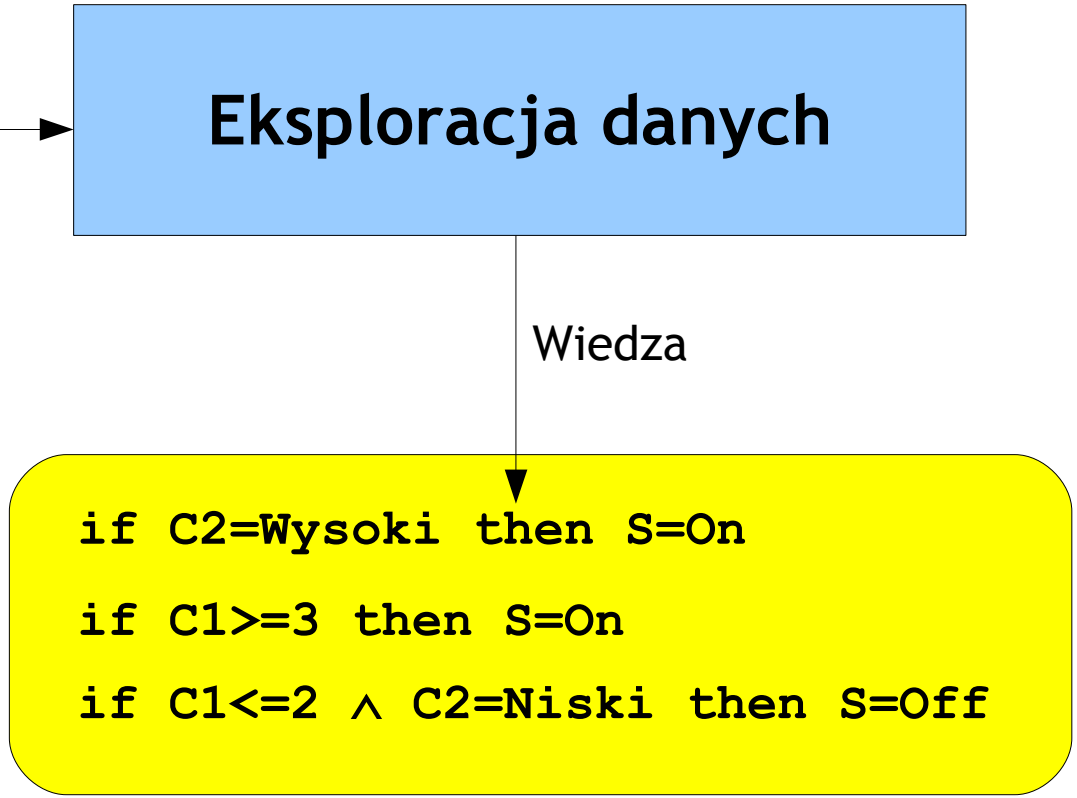
Jak odkrywać w danych wiedzę zapisaną w postaci reguł?

Prosty przykład, cd.

Stosując zdroworozsądkową analizę zbioru danych udało się odkryć zależności pomiędzy polami warunkującymi a polem decyzyjnym.

Czy to jest już Data Mining? Prawie tak, ale niech to robi komputer!

	C1	C2	S
1	1	Niski	Off
2	1	Wysoki	On
3	2	Niski	Off
4	2	Wysoki	On
5	3	Niski	On
6	3	Wysoki	On
7	4	Niski	On
8	4	Wysoki	On



Jak odkrywać w danych wiedzę zapisaną w postaci reguł?

Istnieje wiele metod eksploracji danych.

My skupimy się na jednej, wykorzystującej tablice decyzyjne i zbiory przybliżone.