

# **SYSTEMY INFORMATYCZNE WSPOMAGAJĄCE HODOWLĘ**

---

**Struktura efektywnej bazy danych**

**Zastosowanie pakietu MS Excel do tworzenia baz danych**

## 1. Dane

- Przykłady
- Edycja
- Zarządzanie

## 2. Bazy danych

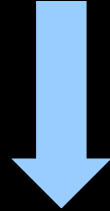
- Definicje
- Przykłady

## 3. MS Excel

- Przykład funkcji bazy danych

# DANE: struktura danych

## PROSTE

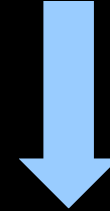


- Brak powiązań pomiędzy poszczególnymi wierszami danych

np.

1	7	9	5697	4.1
2	7	8	4890	3.8
3	6	5	7321	3.5

## ZŁOŻONE



- Poszczególne wiersze danych są od siebie zależne

np.

1	7	9	2001	10.05.2005	23
1	7	9	2001	17.06.2005	34
1	7	9	2001	14.07.2005	30

# Skwantyfikowane wartości fluorescencji dla 14 świń o różnej zawartości tłuszczu mięśniowego

Row	016704_082506	016711_082506	016728_082506	016735_082506	016742_082506	016759_082506	016766_082506
1	13.1106479673798	13.1132795209229	12.9834604132767	12.884528922872	12.6115511630794	12.1749361072167	12.3896615012937
2	13.0324151002927	13.1716643523601	12.8792415430502	13.0844544762107	12.3192414619149	12.1574148265494	12.2614450999793
3	6.63134107118108	5.75488750216347	4.53827227890132	4.48214848663448	9.4938068805185	4.82552584558946	5.57169160699491
4	5.14385618977472	4.68157487234083	4.40442564574610	7.14107364197284	8.76369268427829	5.2980948780722	6.35712275883306
5	14.5847520671715	14.4785435462449	14.3385970120200	14.3092131471956	14.1831376483079	14.0829234232925	14.1683397635055
6	14.6498444849318	14.4659092134881	14.3425498908588	14.4235227214130	13.9846148632020	13.9825645359342	14.1591574202364
7	13.313446251816	13.0179147323452	12.6151343348291	13.0263135319689	12.3137688068651	12.2990869203456	12.7468044247061
8	13.3803059523063	13.0859502657448	12.760905264364	13.0489112072987	12.4024685653143	12.3280045938058	12.6967835730374
9	11.5531031815049	11.4521949327792	11.0055402447147	10.9312384694443	11.1969978915091	10.9391858279843	11.2306437933922
10	10.9755969943301	11.1782203431133	11.1813085053879	11.3775272151661	11.0544143945519	10.8128187527548	11.3868307024195
11	9.39742684781266	9.90445067357883	9.66350436206034	9.75941972897389	9.95968773350612	9.28258451678451	9.39434164470718
12	10.0796055401111	10.0023077513825	9.83007369793772	9.91305465429455	9.0545480318536	9.28212313297198	9.6738193343128
13	9.10361442738257	10.9504037446487	7.71050642788961	10.1778035192598	8.75561766384003	9.46358804822658	9.76357430569706
14	9.68727991128155	10.3725770011988	8.20087309004986	10.8097175300163	8.67506817431194	7.82385568530072	10.3977313142243
15	11.9123006166177	6.37493471369842	7.35798099512757	7.86621622335256	6.44093935380022	6.0659284803044	4.47393934010705
16	7.27145547298457	10.3122339792481	6.90689059560852	8.21461345826091	7.50770752619334	6.52373238665921	7.51808680627674
17	10.7148816300111	10.5682562191293	10.2590797171674	11.0965797791442	10.2189867472938	10.3759330033045	10.1876257983754
18	10.6427620680904	10.4099951859368	10.5877762981028	10.6394725453857	10.1209078049171	10.1209398015248	10.5667711823325
19	11.0677174988467	10.3420352771953	10.4943253561753	10.6171455983475	10.2595613532726	10.1950771292004	10.6636236313279

# DANE: złożone

## INSEMIK

Nazwa buhaja	Nr buhaja	Data prod	1 Obj	1 Rm	1 Rpl	1 Konc	1 Po roz	1 Dyskw	1 Morf	2 Obj	2 Rm	2 Rpl	2 Konc	2 Po roz	2 Dyskw	2 Morf
LEGINS	PL00506243 0529	02/01/2003	7	2	70	1463	50		0	0	0	0	0	0		0
LEGINS	PL00506243 0529	03/01/2003	7.5	3	80	1306	60		0	0	0	0	0	0		0
LEGINS	PL00506243 0529	03/02/2003	5	3	80	1330	50		0	5	3	80	1144	0		0
LEGINS	PL00506243 0529	06/02/2003	5	3	80	1257	50		0	3.5	2	70	695	0		0
LEGINS	PL00506243 0529	10/02/2003	6	3	80	1647	60		0	6	3	80	841	0		0
LEGINS	PL00506243 0529	13/02/2003	3	0	0	0		NAS.B. ORZADKIE	0	2	0	0	0		NAS.B. ORZADKIE	0
LEGINS	PL00506243 0529	17/02/2003	6	3	80	1365	50		0	4.5	3	80	726	0		0
LEGINS	PL00506243 0529	21/02/2003	5	3	80	1483	60		0	5	2	70	668	0		0
LEGINS	PL00506243 0529	24/02/2003	4	3	80	1600	50		0	4	3	80	1119	0		0
LEGINS	PL00506243 0529	27/02/2003	5	3	80	1543	60		0	4	3	80	766	0		0
LEGINS	PL00506243 0529	03/03/2003	5.5	3	80	983	50		0	4	2	70	1077	0		0
LEGINS	PL00506243 0529	06/03/2003	9	3	80	1237	60		0	6	2	70	872	0		0

# EDYCJA I ZAZĄDZANIE DANYMI

---

1. Edycja danych (czyszczenie danych etap I) → usuwanie błędnych informacji

1	7	8	34.0	11	12	12	22
2	7	5	31.7	11	22	12	12
<del>3</del>	<del>6</del>	<del>8</del>	<del>371</del>	<del>11</del>	<del>11</del>	<del>22</del>	<del>11</del>

2. Preprocessing (czyszczenie danych etap II) → usuwanie niepotrzebnych informacji, usuwanie "szumu" z danych

1	7	8	34.0	<del>11</del>	12	12	22
2	7	5	31.7	<del>11</del>	22	12	12
<del>3</del>	<del>6</del>	<del>8</del>	<del>371</del>	<del>11</del>	<del>11</del>	<del>22</del>	<del>11</del>

3. Analiza statystyczna danych

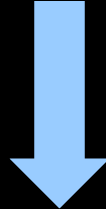
- Istniejące metody
- Nowe metody
- Data mining

4. Wnioskowanie !!!

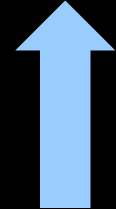
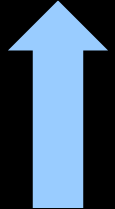
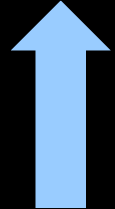
# EDYCJA I ZARZĄDZANIE DANYMI

**DUŻE ZBIORY**

**DYNAMICZNE ZMIANY**



**ZARZĄDZANIE DANYMI → BAZY DANYCH**  
**MANIPULOWANIE DANYMI → PROGRAMY EDYCJI DANYCH**



**KWANTYFIKACJA**

**EDYCJA**

**PREPROCESSING**

## QTL Cartographer

The screenshot displays the WinQTLCart software interface for a file named 'maize.mcd'. The window title is 'WinQTLCart - maize.mcd'. The menu bar includes File, Edit, View, Method, Tools, and Help. The toolbar contains icons for New, Open, Import, Save To, View, Print, IM, CIM, MIM, VO, Result, Simu, DIR, Note, DrawChr, About, and Help.

The interface is divided into several sections:

- Left Panel:** A tree view showing the file structure: Message, Source Files (maize.mcd), Text Files (wqcart.sam), and Result Files (wqcart.sam).
- Source Data File Information:**
  - Summary Information:**

File name:	maize.mcd
File ID number:	901540162
Chromosome numbers:	1
Sample size:	171
Cross type:	SF2
Trait numbers:	1
Other trait numbers:	0
  - Source data view and modify:**
    - Marker values:** A dropdown menu shows 'Chromosome1 - C1' and a 'Markers...' button.
    - Trait Values:** 'Traits...' and 'Other Traits...' buttons.
    - Modify:** 'Map information...' and 'Cross information...' buttons.
  - Analysis:** A dropdown menu shows 'Composite Interval Mapping' and a 'GO' button.
- Bottom Panel:** A text area displaying the source data file content:

```
#FileID 901540162
#bychromosome
/* One way to make comment
  on data source file */
-type position //default is interval
-function 1 //default is 1
-Units cM //default is cM
-chromosomes 1
-maximum 12
-named yes
-start
-Chromosome C1
// Another way to comment
Mk01_01      0.0000
Mk01_02      37.8000
Mk01_03      49.1000
Mk01_04      59.8000
Mk01_05      62.8000
```

The status bar at the bottom indicates 'For Help: press F1' on the left and the date and time '2004-05-25 (2) 03:14 PM' on the right.

# EDYCJA I ZAZĄDZANIE DANYMI

## QTL Express

### Genotype file

```
14
MK1 MK2 MK3 MK4 MK5 MK6 MK7 MK8 MK9 MK10 MK11 MK12 MK13 MK14
LW MS F1 F2
1 2
0
1039 1529 1530 1 F1 4 1 4 7 1 2 1 2 5 2 1 1 2 5 4 2 1 2 1 3 2 4 6 3 1 2 5 2
1040 1529 1530 2 F1 4 1 4 7 1 2 1 2 7 2 1 1 2 5 2 5 1 2 1 3 2 4 1 3 4 2 0 0
```

### Map file

```
2
1
SSC1 7 1
MK2 42 MK1 28 MK7 8 MK10 28 MK3 3 MK4 22 MK5
SSC7 7 1
MK6 35 MK8 25 MK9 17 MK11 22 MK12 25 MK13 21 MK14
```

### Phenotype file

```
2 1 1
Sex Group Weight Fat_depth
-999
1117 1 1 67800 19
1118 2 1 75400 24
1119 2 1 69800 29
1124 1 1 64800 18
1125 1 1 65800 19
```

## PODSUMOWANIE

- 1. Duże rozmiary danych wymagają zarządzania nimi w bazach danych**
- 2. Dynamika uaktualniania zbiorów danych wymaga wymaga zarządzania i bieżącej kontroli poprawności napływających danych**
- 3. Przed analizą z danych należy wsunąć błędy**
- 4. Przeglądanie i wizualizacja danych (Excel, Notatnik) są często niemożliwe**
- 5. Różne pakiety statystycznej analizy danych wymagają różnych formatów**
- 6. Dobrze stworzone i udokumentowane bazy danych i programy do ich edycji można wykorzystać do innych analiz**

## NIE MODYFIKOWAĆ DANYCH "NA PIECHOTĘ"

---

1. Np. usuwanie genotypów, przygotowywanie formatu danych dla różnych pakietów statystycznych
2. Modyfikacja danych "na piechotę" generuje dużo błędów
3. Strata czasu
4. Niemożliwa dla większości zbiorów danych z powodu ich dużych rozmiarów

## TWORZYĆ DOKUMENTACJĘ

---

1. Informatywne nazwy tabel w bazie danych
2. Umieszczanie **OBSZERNYCH** komentarzy w programach do edycji danych
3. Tworzenie dokumentacji programów:
  - Cel programu
  - Data modyfikacji
  - Pliki wejściowe i wyjściowe - nazwy i format danych
  - j. angielski
4. Automatyzacja działania poszczególnych programów

# ASPEKTY ZARZĄDZANIA I MANIPULOWANIA DANYMI

---

## PRZYKŁAD ROBOCZEJ DOKUMENTACJI

----- running programmes

```
ASRem1 -NS9 /home/szyda/MICROARRAY/PROGRAMS/fixed.as  
nohup R --vanilla < directslve.R > directslve.log &
```

----- file extensions

try:

```
*.as -> giv variance=0.01  
*.as2 -> more sparse covariance matrix  
*.as3 -> giv variance=0.03 noiter FINISHED  
*.as4 -> using diagonal covariance matrix  
*.as5 -> giv variance=0.03 noiter FINISHED  
*.as6 -> giv variance=0.03 noiter FINISHED  
*.as7 -> giv variance=0.03 noiter
```

----- sequence of analysing data

```
1. readmatrix-1.f <- macierzN.csv -> fort.macierzN  
2. readmatrix.sas <- kody.txt + fort.macierzN -> all_col.macierzN  
3. readgenematrix1.f <- macierzN.csv + all_col.macierzN -> genecovN.txt  
4. readgenematrix0.f <- genecovN.txt -> outputs on a terminal min/max value of  
the matrix  
5. readgenematrix3.f <- genecovN.txt -> genecovN_100.full  
6. directslve.R <- genecovN_100.full -> genecovN_100.inv1  
7. readinversematrix1.f <- genecovN_100.inv1 -> genecovN_100.giv  
8. readmatrix1.sas <- all_col.macierzN + betweenarrayM.out1 ->  
betweenarrays.out1.GID
```

# **BAZY DANYCH – najważniejsze cechy**

---

## **CZYM JEST BAZA DANYCH ?**

- 1. Zbiorem danych**
- 2. Komputerowa baza danych – plik(i) elektroniczne**

## **JAKIE SĄ PODSTAWOWE CECHY ELEKTRONICZNEJ BAZY DANYCH ?**

- 1. Sprawdzanie błędów**
- 2. Proste manipulacje na danych**
- 3. Udostępnianie danych**

# BAZY DANYCH – najważniejsze pojęcia

---

## 1. Pole

- **Pojedyncze źródło informacji, jednostka danych, np. nr osobnika, rok urodzenia**

## 2. Rekord danych

- **Wiersz danych, grupa pól zawierających informacje o tym samym osobniku, oborze, itp.**

## 3. Tabela

- **Zbiór danych w formie tabeli o zdefiniowanych kolumnach=polach i wierszach=rekordach**

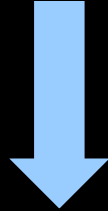
## 4. Relacja

- **Pole w danym rekordzie danych wskazujące na nr rekordu danych zawierającego powiązane z nim, informacje np. rekord z pochodzeniem osobnika i rekord z jego wydajnościami**

# BAZY DANYCH – proste

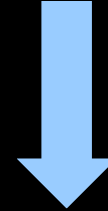
---

## PROSTE



- **Pojedyncza tabela danych lub kilka niezależnych tabel**

## RELACYJNE



- **Wiele plików danych powiązanych między sobą za pomocą indeksów**

# BAZY DANYCH – rodzaje

<b>nr lab</b>	<b>nr osobnika</b>	<b>abcg2</b>	<b>lepR</b>	<b>btn3</b>	<b>btn1</b>	<b>btn2</b>	<b>dgat1</b>	<b>lep2a</b>	<b>lep3</b>	<b>lept1</b>	<b>lep7</b>
942	PL005006200324	1	3	2	1	2	2	2	1	1	2
943	PL005006200355	1	3	2	1	3	2	1	1	1	3
944	PL005006200416	1	3	2	1	3	2	3	3	1	1
945	PL005006200423	1	3	2	2	2	2	2	2	1	2
947	PL005006800463	1	3	2	1	2	1	3	2	1	1
948	PL005001502973	1	3	3	2	3	1	3	3	1	1
949	PL005001503178	1	2	1	1	3	1	2	1	1	2

# BAZY DANYCH – relacyjne

## TABELA: GENOTYPY

nr lab	nr osobnika	abcg2	lepR	btn3	btn1	btn2	dgat1	lep2a	lep3	lept1	lep7
942	PL005006200324	1	3	2	1	2	2	2	1	1	2
943	PL005006200355	1	3	2	1	3	2	1	1	1	3

## TABELA: KROWY

nr lab	dzień urodzenia	miesiąc urodzenia	rok urodzenia	hodowca
942	15	03	2001	2
943	23	10	2003	7

## TABELA: WYDAJNOŚCI

nr lab	obora	dzień próbnego udoju	miesiąc próbnego udoju	rok próbnego udoju	wydajność mleka	% tłuszczu
942	1	10	3	2004	13	5.1
942	1	15	4	2004	19	4.6
...						
943	2	01	3	2006	31	4.0

# BAZY DANYCH – rodzaje relacji

---

## 1. Jeden-do-jednego

- Tabela genotypów i tabela pochodzenia krów

## 2. Jeden-do-wielu

- Tabela genotypów i tabela wydajności w próbnym udojach

## 3. Wiele-do-wielu

- Tabela wydajności próbnym udojów i tabela informacji o oborach

# BAZY DANYCH – przykładowe narzędzia baz danych

## 1. MS Excel

- Znane narzędzie, najprostsze bazy danych, Windows

## 2. MS Access

- Przyjazny dla użytkownika, komponent MS office Professional, Windows

## 3. MySql

- Stosunkowo przyjazny dla użytkownika, darmowy <http://dev.mysql.com/>, Windows + Linux



The screenshot shows the MySQL Downloads page. At the top, there is a navigation bar with links: MySQL.com, Downloads (highlighted), Developer Zone, Partners & Solutions, and Customer Login. Below this is a secondary navigation bar with links: Downloads, Archives, Snapshots, and Mirrors. The main content area is titled "MySQL Downloads" and features the MySQL logo (two fish in a circle). To the left of the main content is a sidebar with a dark blue background and white text, listing various MySQL products: MySQL Community Server, MySQL Proxy, MySQL Cluster, MySQL Workbench, GUI Tools, and Connectors. To the right of the logo, there is a section titled "MySQL Downloads" with a list of three bullet points: "MySQL software is provided under the GPL License", "OEMs, ISVs and VARs can purchase Commercial Licenses", and "Learn about MySQL Products and MySQL Services".

## 4. SAS

- **Profesjonalny pakiet zarządzania danymi, baaardzo drogi, różne systemy operacyjne**

## 5. Oracle

- **Szeroko stosowany profesjonalny pakiet, wszystkie systemy operacyjne**

# BAZY DANYCH – przykład Excel

---

1. Otworzyć dane produkcja.txt w notatniku
2. Otworzyć dane w TextPad
3. Otworzyć w Excelu (rozdzielane spacjami)
4. Utworzyć prostą bazę danych
  - Nadać nazwy kolumn
  - W kolejnym arkuszu opisać nazwy kolumn = utworzyć dokumentację
  - filtry
    - Ustawić filtr dla kolumny z nazwą buhaja
    - Dane-filtruj
    - Przykładowe filtry tekstu np. tylko buhaje polskie „PL”
    - Ustawić filtr dla kolumny z wartością hodowlaną wydajności mleka
    - Przykładowe filtry liczbowe np. powyżej średniej
    - Kombinacje filtrów
  - Poprawność danych
    - Dodać kolumny z dniem, miesiącem i rokiem urodzenia osobnika
    - Opisać je w dokumentacji
    - Zdefiniować kontrolę poprawności: Dane-poprawność danych-pełna liczba-między
    - Wprowadzić próbne dane: poprawne i niepoprawne

## **1. Dane**

- **Przykłady**
- **Edycja**
- **Zarządzanie**

## **2. Bazy danych**

- **Definicje**
- **Przykłady**

## **3. MS Excel**

- **Przykład funkcji bazy danych**