



VŠĮ „Informatikos mokslų centras“



MOKSLAS • EKONOMIKA • SANGLAUDA



EUROPOS SAJUNGA  
EUROPOS REGIONINIS  
PLĖTROS FONDAS

*Kuriame Lietuvos ateitį*

# Duomenų analizės įrankis DAMIS

**dr. Jolita Bernatavičienė**

**Kaunas, 2015**



# DAMIS – nacionalinio mokslo informacijos archyvo MIDAS duomenų analizės įrankis

The screenshot shows the MIDAS web application interface. The browser address bar displays the URL `http://test.m...vate-app.html`. The page header includes the MIDAS logo and navigation links: "Duomenų analizės įrankis" (highlighted with an orange circle), "Užduotys", "Pranešimai +8", and a user profile icon. The left sidebar contains navigation options: "+ Naujas tyrimas", "PUBLIKUOTI TYRIMAI", "NEPUBLIKUOTI TYRIMAI" (with sub-items "123", "ee", "SKLanketa"), "ŠIUKŠLIADĖŽĖ", and "FTP". The main content area shows a notification for the last login, a search bar with the text "Paieška", and a list of unpublished research items. The table below lists three items with columns for name, type, modification date, status, and request type.

PAVADINIMAS	TIPAS	MODIFIKUOTAS	BŪSENA	PRIEIGOS TIPAS
123		2015-04-03 08:41:15	Nepublikuotas	Tyrimo dalyviai
ee		2015-04-09 16:19:52	Nepublikuotas	Tyrimo dalyviai
SKLanketa		2015-04-07 16:40:47	Nepublikuotas	Tyrimo dalyviai

# DAMIS funkcionalumas



- DAMIS – **internetinė sistema** (web application):  
nereikia jokio papildomo diegimo, užtenka interneto naršyklės.

# DAMIS funkcionalumas

- DAMIS – **internetinė sistema** (web application): nereikia jokio papildomo diegimo, užtenka interneto naršyklės.
- Galima rinktis **lygiagrečių ir paskirstytųjų skaičiavimų resursus** (VU MII klasteris – VU MIF superkompiuteris), todėl gali būti analizuojami įvairios apimties duomenis.

# DAMIS funkcionalumas

- DAMIS – **internetinė sistema** (web application): nereikia jokio papildomo diegimo, užtenka interneto naršyklės.
- Galima rinktis **lygiagrečiųjų ir paskirstytųjų skaičiavimų resursus** (VU MII klasteris – VU MIF superkompiuteris), todėl gali būti analizuojami įvairios apimties duomenis.
- Įgyvendintas **mokslinių užduočių sekų** principas.

# DAMIS funkcionalumas



- DAMIS – **internetinė sistema** (web application): nereikia jokio papildomo diegimo, užtenka interneto naršyklės.
- Galima rinktis **lygiagrečiųjų ir paskirstytųjų skaičiavimų resursus** (VU MII klasteris – VU MIF superkompiuteris), todėl gali būti analizuojami įvairios apimties duomenis.
- Įgyvendintas **mokslinių užduočių sekų** principas.
- Gautus rezultatus galima **išsaugoti** naudotojo kompiuteryje arba **MIDAS saugykloje**.

# DAMIS funkcionalumas



- DAMIS – **internetinė sistema** (web application): nereikia jokio papildomo diegimo, užtenka interneto naršyklės.
- Galima rinktis **lygiagrečiųjų ir paskirstytųjų skaičiavimų resursus** (VU MII klasteris – VU MIF superkompiuteris), todėl gali būti analizuojami įvairios apimties duomenis.
- Įgyvendintas **mokslinių užduočių sekų** principas.
- Gautus rezultatus galima **išsaugoti** naudotojo kompiuteryje arba **MIDAS saugykloje**.
- Įgyvendinti **įvairūs duomenų analizės** algoritmai, gali būti analizuojami **skaitiniai daugiamačiai duomenys**, kurie pateikti lentelės pavidalu.

# Daugiamačių duomenų pavyzdys (krūties vėžio duomenys)

$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$	$x_9$	C
5	1	1	1	2	1	3	1	1	b
5	4	4	5	7	10	3	2	1	b
3	1	1	1	2	2	3	1	1	b
6	8	8	1	3	4	3	7	1	b
4	1	1	3	2	1	3	1	1	b
1	1	1	1	2	10	3	1	1	b
2	1	2	1	2	1	3	1	1	b
2	1	1	1	2	1	1	1	5	b
4	2	1	1	2	1	2	1	1	b
...	...	...	...	...	...	...	...	...	...
8	10	10	8	7	10	9	7	1	m
5	3	3	3	2	3	4	4	1	m
8	7	5	10	7	9	5	5	4	m
7	4	6	4	6	1	4	3	1	m
10	7	7	6	4	10	4	1	2	m
7	3	2	10	5	10	5	4	4	m
10	5	5	3	6	7	7	10	1	m
...	...	...	...	...	...	...	...	...	...
4	8	8	5	4	5	10	4	1	m

## Duomenų požymiai:

$x_1$  – clump thickness,  
 $x_2$  – uniformity of cell size,  
 $x_3$  – uniformity of cell shape,  
 $x_4$  – marginal adhesion,  
 $x_5$  – single epithelial cell size,  
 $x_6$  – bare nuclei,  
 $x_7$  – bland chromatin,  
 $x_8$  – normal nucleoli,  
 $x_9$  – mitoses.

## Klasės:

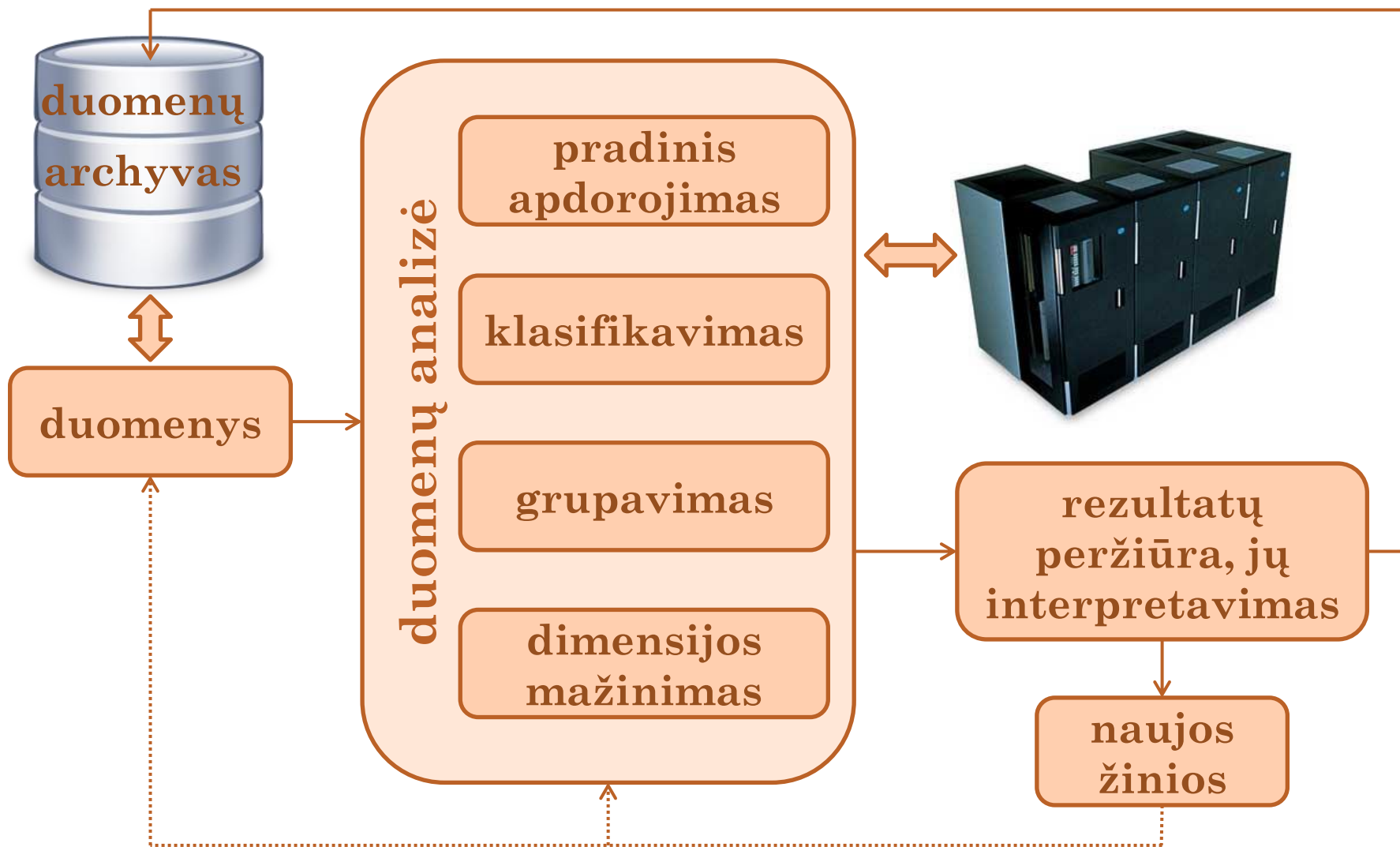
C – class (**b**enign, **m**alignant)

dimensija (matmenų skaičius)  
 $n = 9$

viena eilutė – vienos pacientės  
 duomenys



# Duomenų analizė






# DAMIS grafinė naudotojo sąsaja

DAMIS [Kurti eksperimentą +](#) [Eksperimentai ☰](#) [Failų sąrašas 📄](#)    [Lietuvių ▾](#)

MII klasteris ⓘ MIF VU SK2 ⓘ

▾ Duomenų įkėlimas

 ⓘ  ⓘ  ⓘ

[kelti naują failą] Pasirinkti [kelti failą iš MIDAS]

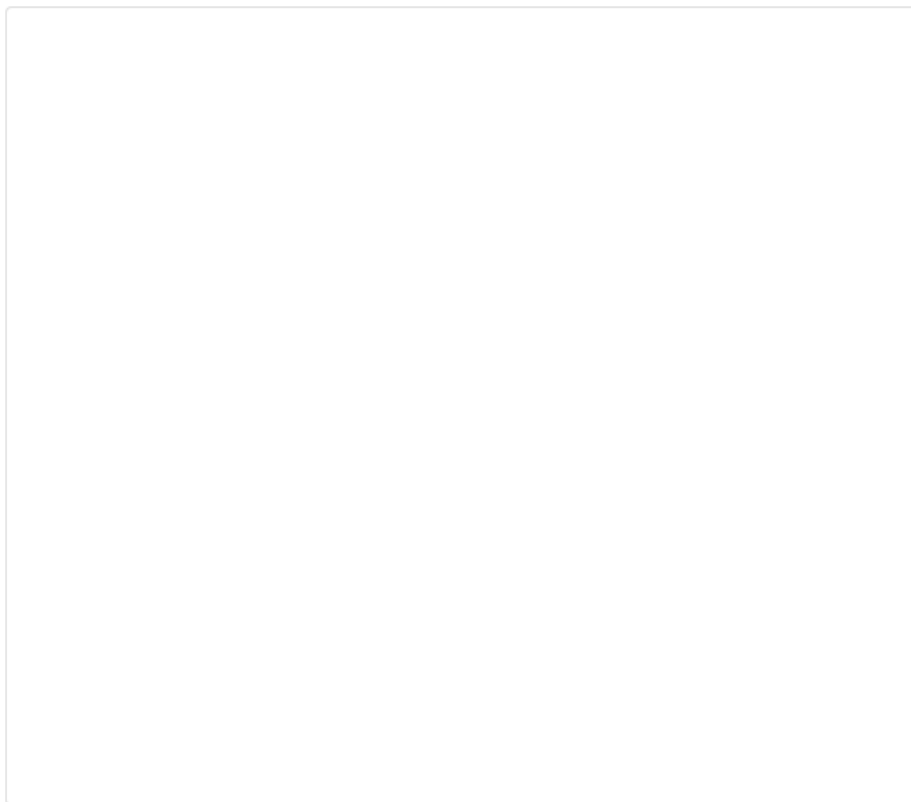
▶ Pirminis apdorojimas

▶ Statistiniai primityvai

▶ Dimensijos mažinimas

▶ Klasifikavimas, grupavimas

▶ Rezultatų peržiūra



[Naujas eksperimentas](#) [Išsaugoti](#) [Vykdyti](#)

# DAMIS grafinė naudotojo sąsaja



eksperimentų vykdymo valdymas

Darbalaukis, kuriame  
bus formuojama  
eksperimento  
užduočių seka  
(scientific workflow)

kompo-  
nentės

Duomenų įkėlimas



Įkelti naują failą   Pasirinkti įkeltą failą iš MIDAS

▶ Pirminis apdorojimas

▶ Statistiniai primityvai

▶ Dimensijos mažinimas

▶ Klasifikavimas, grupavimas

▶ Rezultatų peržiūra

Naujas eksperimentas

Išsaugoti

Vykdyti

# DAMIS duomenų analizei




## ○ Įgyvendinta:


- **pradinio apdorojimo** algoritmai (valymas, filtravimas, skaidymas, transponavimas, normavimas, požymių atrinkimas),
- **pagrindinės statistinės charakteristikos** (minimumas, maksimumas, vidurkis, standartinis nuokrypis, mediana),
- **dimensijos mažinimo** (vizualizavimo) algoritmai,
- **klasifikavimo** ir **grupavimo** (klasterizavimo) algoritmai,
- gautų **rezultatų** peržiūra.


# Skaičiavimų resursų parinkimas

MII klasteris <sup>i</sup> MIF VU SK2 <sup>i</sup>

▼ Duomenų įkėlimas

 <sup>i</sup> Įkelti naują failą

 <sup>i</sup> Pasirinkti įkeltą failą

 <sup>i</sup> Įkelti failą iš MIDAS

► Pirminis apdorojimas

► Statistiniai primityvai

► Dimensijos mažinimas

► Klasifikavimas, grupavimas

► Rezultatų peržiūra

Naujas eksperimentas

Išsaugoti

Vykdyti

# Skaičiavimų resursų parinkimas

MII klasteris <sup>i</sup> MIF VU SK2 <sup>i</sup>

## Skaičiavimo telkinio informacija

Vilniaus universiteto Matematikos ir informatikos instituto paskirstytųjų skaičiavimų klasteris

Klasterio svetainė:

<http://hpc.mii.vu.lt/>

Klasterio apkrovimas:

<http://hpc.mii.vu.lt/ganglia/>

▸ Dimensijos mažinimas

▸ Klasifikavimas,  
grupavimas

▸ Rezultatų peržiūra

Naujas eksperimentas

Išsaugoti

Vykdyti

# Skaičiavimų resursų parinkimas

MII klasteris MIF VU SK2

Skaičiavimo telkinio informacija

Vilniaus universiteto Matematikos ir informatikos fakulteto paskirstytųjų skaičiavimų superkompiuteris

Klasterio svetainė:  
<http://mif.vu.lt/cluster/>

Klasterio apkrovimas:  
<http://k007.mif.vu.lt/ganglia2/>

▼ Duomenys

Įkelti naują failą

► Pirmiausia

► Statistiniai primityvai

► Dimensijos mažinimas

► Klasifikavimas, grupavimas

► Rezultatų peržiūra

Naujas eksperimentas

Išsaugoti

Vykdyti

# Duomenų įkėlimas

MII klasteris **MIF VU SK2**

## ▼ Duomenų įkėlimas



Įkelti naują  
failą

Pasirinkti  
įkeltą failą

Įkelti failą iš  
MIDAS



Įkelti failą iš  
MIDAS

▶ Pirminis apdorojimas

▶ Statistiniai primityvai

▶ Dimensijos mažinimas

▶ Klasifikavimas,  
grupavimas

▶ Rezultatų peržiūra

Norima komponentė **plyte tempiama** į darbalaukį;

Prieš tai duomenys **turi būti tinkamai paruošti.**

Naujas eksperimentas

Išsaugoti

Vykdyti



# Duomenų įkėlimas iš MIDAS

MII klasteris MIF VU SK2

▼ Duomenų įkėlimas

Įkelti naują failą Pasirinkti įkeltą failą Įkelti failą iš MIDAS

Įkelti failą iš MIDAS

Pavadinimas	Modifikuota
<a href="#">Publikuoti tyrimai</a>	2015-04-09
Nepublikuoti tyrimai	2015-04-09

Patvirtinti Atšaukti

Naujas eksperimentas Išsaugoti Vykdyti

# Duomenų pirminis apdorojimas

MII klasteris

MIF VU SK2

▶ Duomenų įkėlimas

▼ Pirminis apdorojimas



Valymas



Filtravimas



Skaidymas



Transponavimas



Normavimas



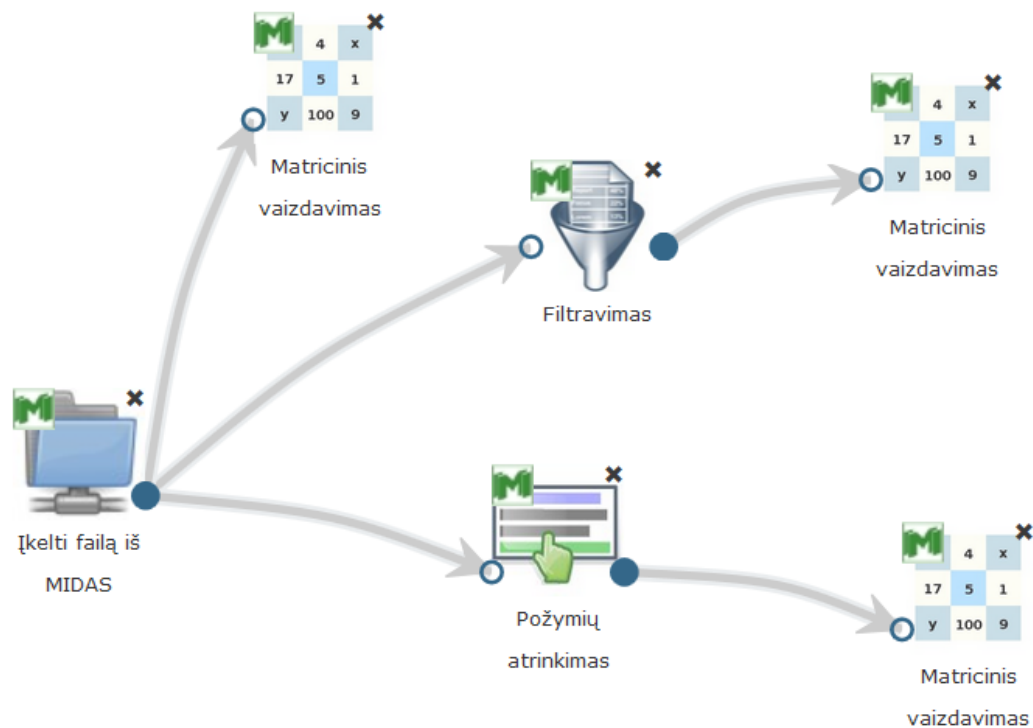
Požymių atrinkimas

▶ Statistiniai primityvai

▶ Dimensijos mažinimas

▶ Klasifikavimas, grupavimas

▶ Rezultatų peržiūra



Įkeltos komponentės sujungiamos į vadinamąsias **mokslinių užduočių sekas** (scientific workflows)

Naujas eksperimentas

Išsaugoti

Vykdyti

# Komponenčių parametų parinkimas

MII klasteris MIF VU SK2

Duomenų įkėlimas

Pirminis apdorojimas

Požymių atrinkimas

**Požymių atrinkimas**

**Požymiai**

- Uniformity\_of\_C
- Single\_Epithelia
- Bare\_Nuclei
- Bland\_Chromat
- Normal\_Nucleol
- Mitoses
- class\_attr

**Pasirinkti požymiai**

- Clump\_Thickne:
- Uniformity\_of\_C
- Marginal\_Adhes

**Klasės požymis**

class\_attr

Patvirtinti Atšaukti

Naujas eksperimentas Išsaugoti Vykdėti

Du kartus spustelėjus pelyte įkeltą komponentę atsiranda langas, kuriame **galima pasirinkti** komponenčių parametrus

# Rezultatų peržiūra

MII klasteris

MIF VU SK2

▸ Duomenų įkėlimas

▸ Pirminis apdorojimas

▸ Statistiniai primityvai

▸ Dimensijos mažinimas

▸ Klasifikavimas, grupavimas

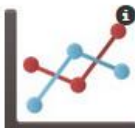
▾ Rezultatų peržiūra



Techninė  
informacija



Matricinis  
vaizdavimas



Grafinis  
vaizdavimas



Įkelti failą iš  
MIDAS



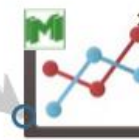
Matricinis  
vaizdavimas



Filtravimas



Matricinis  
vaizdavimas



Grafinis  
vaizdavimas



Požymių  
atrinkimas



Matricinis  
vaizdavimas

Naujas eksperimentas

Išsaugoti

Vykdyti

# Rezultatų matricinis vaizdavimas

**Matricinis vaizdavimas**

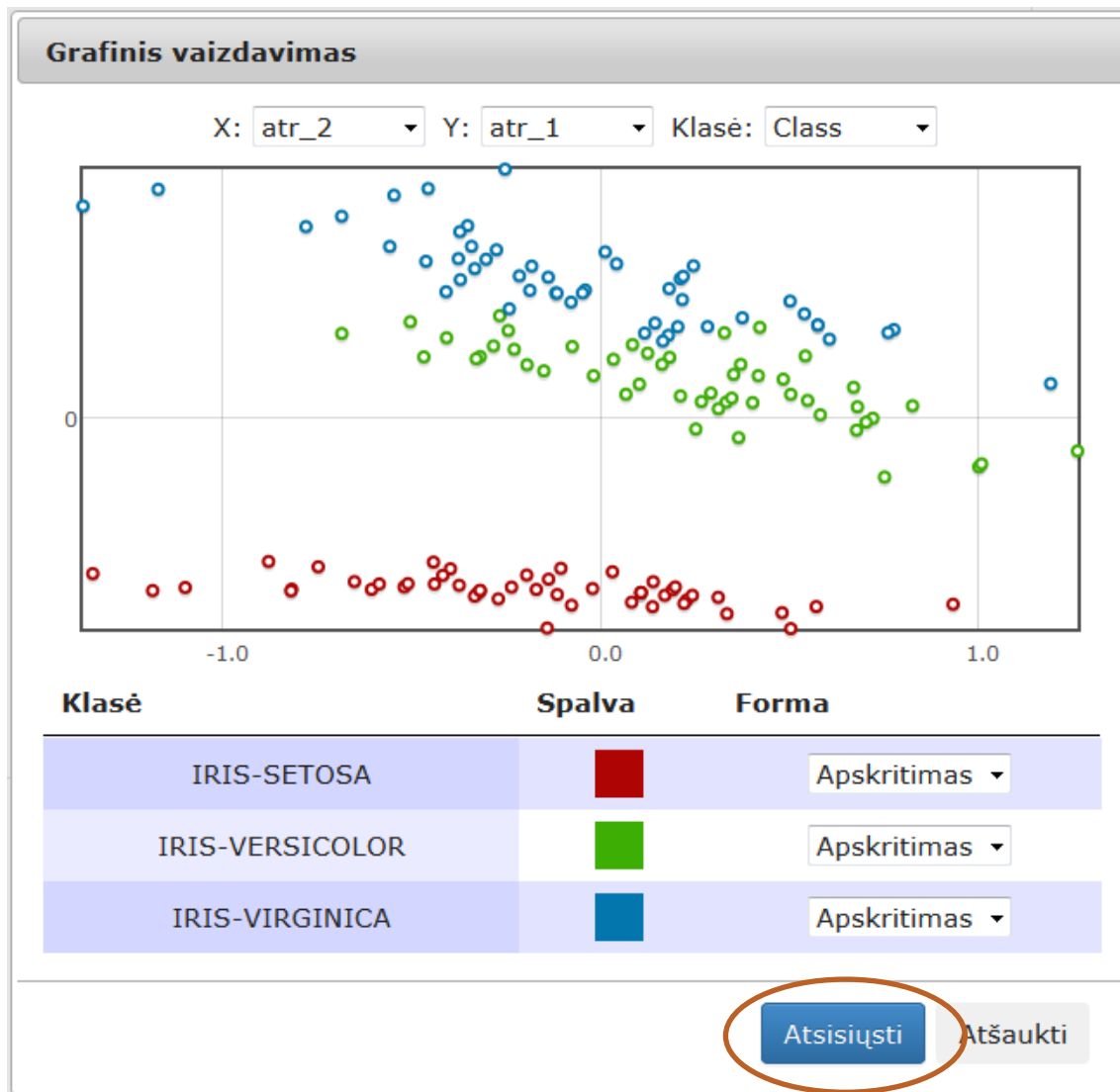
Size	Bare_Nuclei (REAL)	Bland_Chromatin (REAL)	Normal_Nucleoli (REAL)	Mitoses (REAL)	CL
1	3	1	1	2	
10	3	2	1	2	
2	3	1	1	2	
4	3	7	1	2	
1	3	1	1	2	
10	9	7	1	4	
10	3	1	1	2	
1	3	1	1	2	

Atsisiųsti Atšaukti

# Rezultatų grafinis vaizdavimas

- Svarbu duomenis **pateikti vizualia forma.**
- Žmogui **lengviau suvokti** vizualizuotus duomenis, nei jų skaitinius įverčius.
- Įrankyje įgyvendinta **dvimatė taškinė diagrama** (*scatter plot*).
- Jei vizualizuojami duomenys, kuriuos apibūdina daugiau nei du požymiai, galima **peržiūrėti visų galimų porų vaizdus.**

# Rezultatų grafinis vaizdavimas



# Eksperimento vykdymas

MII klasteris

MIF VU SK2

▶ Duomenų įkėlimas

▶ Pirminis apdorojimas

▶ Statistiniai primityvai

▶ Dimensijos mažinimas

▶ Klasifikavimas, grupavimas

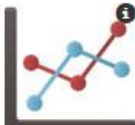
▼ Rezultatų peržiūra



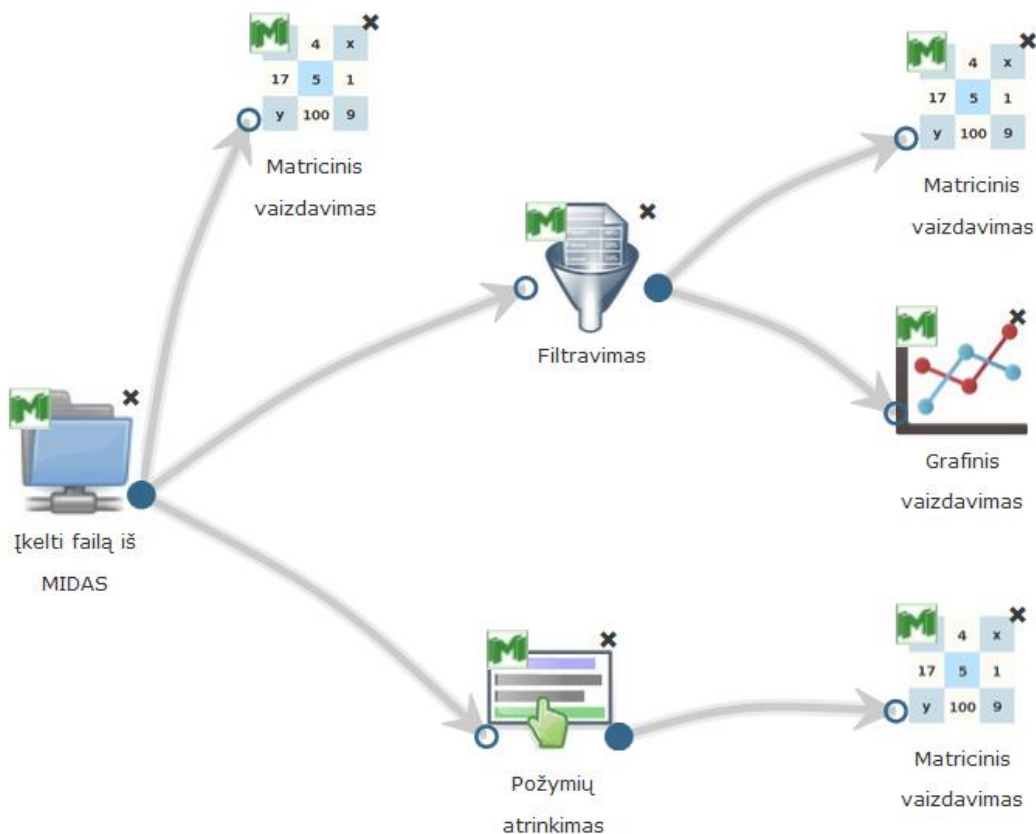
Techninė  
informacija



Matricinis  
vaizdavimas



Grafinis  
vaizdavimas



Naujas eksperimentas

Išsaugoti

Vykdyti



# Eksperimentų vykdymo istorija

DAMIS

Kurti eksperimentą +

Eksperimentai

Failų sąrašas



Lietuvių

## Eksperimentų istorija



<input type="checkbox"/>	Pavadinimas	Statusas	Veiksmai
<input type="checkbox"/>	exp139	Vykdoma	
<input type="checkbox"/>	exp110	Įvykdytas	
<input type="checkbox"/>	exp109	Įvykdytas	
<input type="checkbox"/>	exp108	Įvykdytas	
<input type="checkbox"/>	exp107	Įvykdytas	
<input type="checkbox"/>	exp000	Įvykdytas	

# Duomenų statistinės charakteristikos

MII klasteris

MIF VU SK2

▶ Duomenų įkėlimas

▶ Pirminis apdorojimas

▼ Statistiniai primityvai

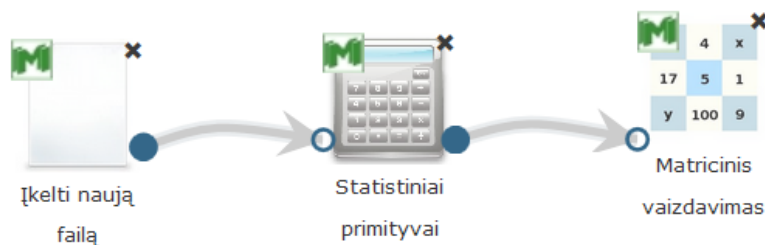


Statistiniai  
primityvai

▶ Dimensijos mažinimas

▶ Klasifikavimas,  
grupavimas

▶ Rezultatų peržiūra



Statistiniai primityvai – min, max, vidurkis, standartinis nuokrypis, mediana

Naujas eksperimentas

Išsaugoti

Vykdyti

# Duomenų dimensijos mažinimas

MII klasteris

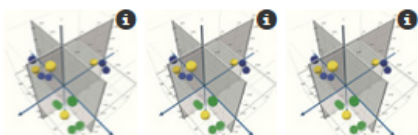
MIF VU SK2

▶ Duomenų įkėlimas

▶ Pirminis apdorojimas

▶ Statistiniai primityvai

▼ Dimensijos mažinimas

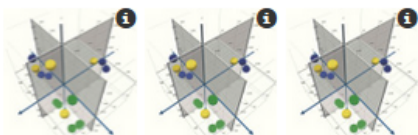


PCA

SMACOF

DMA

(MDS)



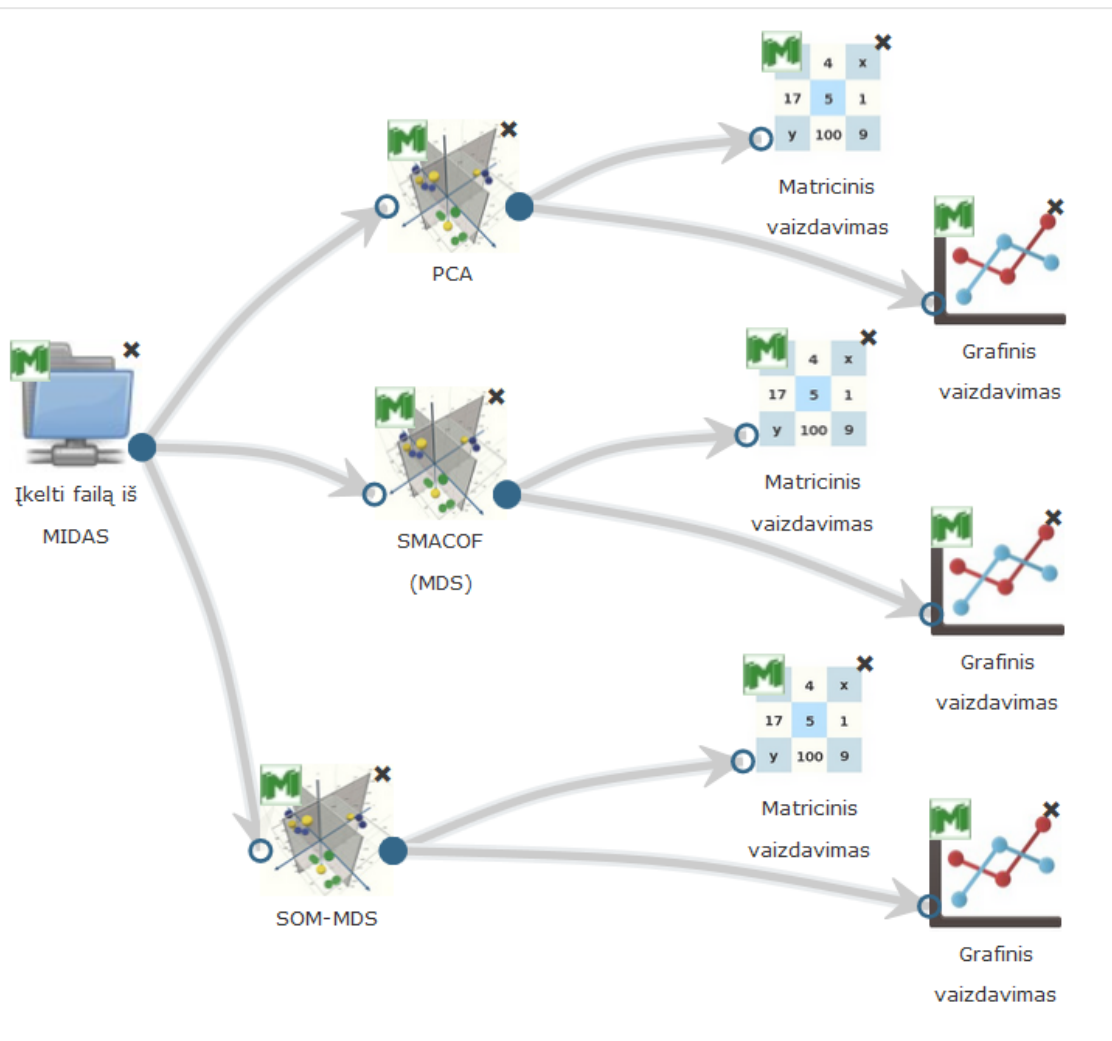
Relative  
MDS

SAMANN

SOM-MDS

▶ Klasifikavimas,  
grupavimas

▶ Rezultatų peržiūra



Naujas eksperimentas

Išsaugoti

Vykdyti

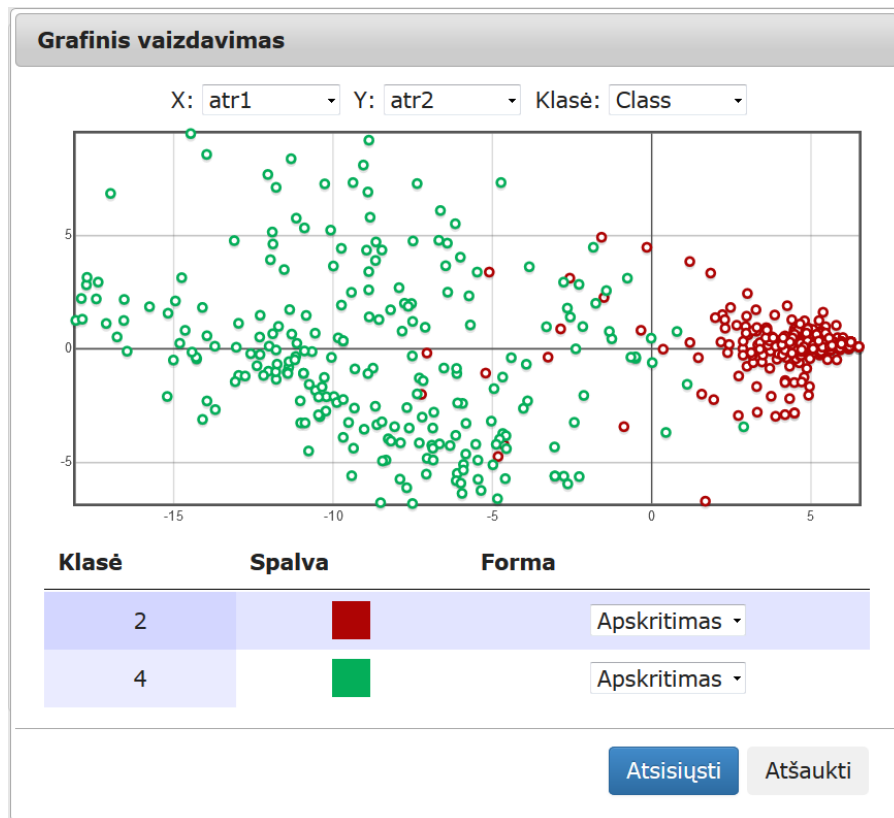
# Duomenų dimensijos mažinimas

- Dažnai duomenis **apibūdina daug požymių**, todėl jie yra daugiamačiai.
- Būtina **sumažinti dimensiją**, siekiant palengvinti tolesnę duomenų analizę.
- Įgyvendinti šie duomenų **dimensijos mažinimo metodai**:
  - PCA – pagrindinių komponentų analizė,
  - SMACOF(MDS), DMA, Relative MDS – daugiamatės skalės ir jų variantai,
  - SAMANN – dirbtiniai neuroniniai tinklai daugiamatėms skalėms,
  - SOM-MDS – saviorganizuojančių neuroninių tinklų ir daugiamačių skalių junginys.

# Krūties vėžio duomenų dimensijos mažinimas (vizualizavimas) MDS metodu

Daugiamačių duomenų transformavimas į mažesnės dimensijos erdvę, pavyzdžiui, dvimatę

$$(x_{i1}, x_{i2}, \dots, x_{i9}) \rightarrow (y_{i1}, y_{i2})$$



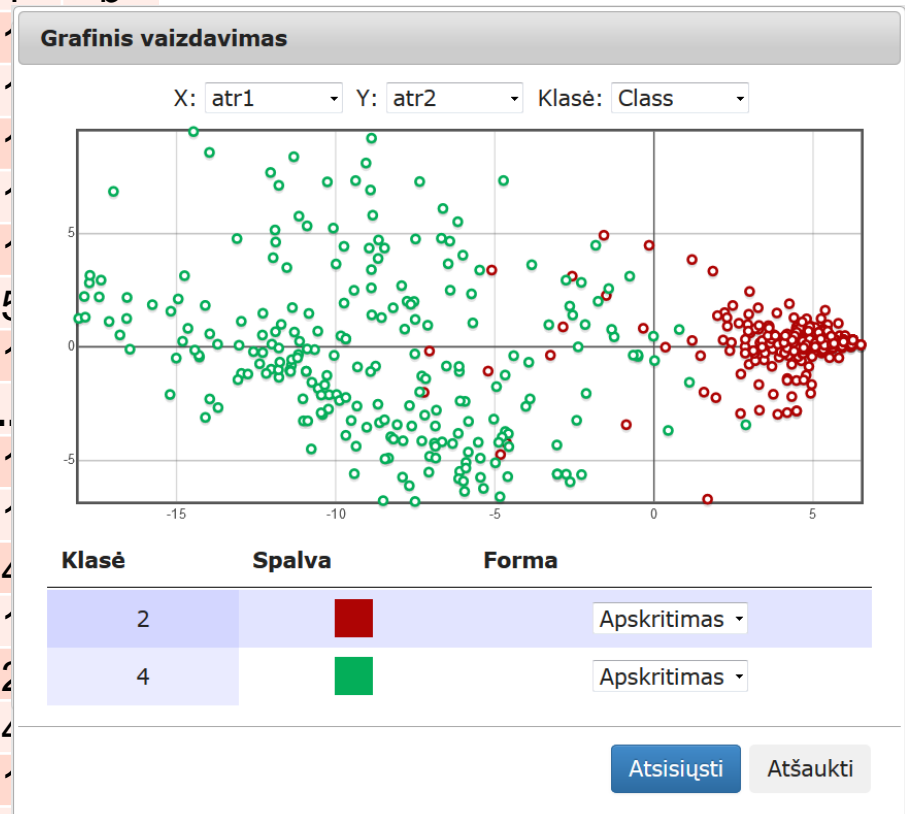
- vienas **taškas** atitinka vieną **pacientę**
- vienas **taškas** apjungia visus **devynis požymius**
- vizualus pateikimas padeda lengviau suvokti **informacijos visumą**

2 – **benign**, 4 – **malignant**

# Krūties vėžio duomenų dimensijos mažinimas (vizualizavimas) MDS metodu

$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$	$x_9$	C
5	1	1	1	2	1	3	1	1	b
5	4	4	5	7	10	3	2	1	b
3	1	1	1	2	2	3	1	1	b
6	8	8	1	3	4	3	7	1	b
4	1	1	3	2	1	3	1	1	b
1	1	1	1	2	10	3	1	1	b
2	1	2	1	2	1	3	1	1	b
2	1	1	1	2	1	1	1	1	b
4	2	1	1	2	1	2	1	1	b
...	...	...	...	...	...	...	...	...	...
8	10	10	8	7	10	9	7	1	b
5	3	3	3	2	3	4	4	1	b
8	7	5	10	7	9	5	5	1	b
7	4	6	4	6	1	4	3	1	b
10	7	7	6	4	10	4	1	1	b
7	3	2	10	5	10	5	4	1	b
10	5	5	3	6	7	7	10	1	b
...	...	...	...	...	...	...	...	...	...
4	8	8	5	4	5	10	4	1	m

Class (C):  
 2 – **benign**  
 4 – **malignant**



# Duomenų klasifikavimas

MII klasteris

MIF VU SK2

▶ Duomenų įkėlimas

▶ Pirminis apdorojimas

▶ Statistiniai primityvai

▶ Dimensijos mažinimas

▼ Klasifikavimas, grupavimas



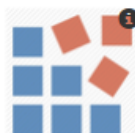
SOM



MLP

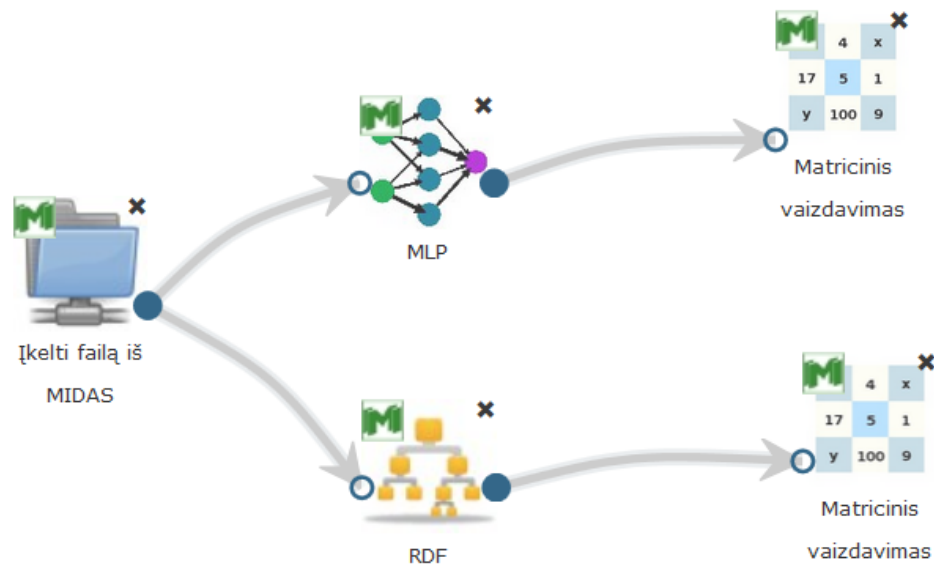


RDF



K-MEANS

▶ Rezultatų peržiūra



Naujas eksperimentas

Išsaugoti

Vykdyti

# Duomenų klasifikavimas

- **Klasifikavimo tikslas** – turint duomenis, kurių klasės yra žinomas, rasti klases duomenims, kurių klasės nėra žinomos.
- Klasifikavimo **uždaviniai yra sprendžiami** medicinoje (preliminari diagnostika), ekonomikoje, finansuose ir kt.
- Įgyvendinti šie **duomenų klasifikavimo metodai**:
  - MLP – daugiasluoksnis neuroninis tinklas (perceptronas),
  - RDF – sprendimų medžiais pagrįstas klasifikatorius.



# Krūties vėžio duomenys

$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$	$x_9$	C
5	1	1	1	2	1	3	1	1	b
5	4	4	5	7	10	3	2	1	b
3	1	1	1	2	2	3	1	1	b
6	8	8	1	3	4	3	7	1	b
4	1	1	3	2	1	3	1	1	b
1	1	1	1	2	10	3	1	1	b
2	1	2	1	2	1	3	1	1	b
2	1	1	1	2	1	1	1	5	b
4	2	1	1	2	1	2	1	1	b
...	...	...	...	...	...	...	...	...	...
8	10	10	8	7	10	9	7	1	m
5	3	3	3	2	3	4	4	1	m
8	7	5	10	7	9	5	5	4	m
7	4	6	4	6	1	4	3	1	m
10	7	7	6	4	10	4	1	2	m
7	3	2	10	5	10	5	4	4	m
10	5	5	3	6	7	7	10	1	m
...	...	...	...	...	...	...	...	...	...
4	8	8	5	4	5	10	4	1	m

## Duomenų požymiai:

$x_1$  – clump thickness,  
 $x_2$  – uniformity of cell size,  
 $x_3$  – uniformity of cell shape,  
 $x_4$  – marginal adhesion,  
 $x_5$  – single epithelial cell size,  
 $x_6$  – bare nuclei,  
 $x_7$  – bland chromatin,  
 $x_8$  – normal nucleoli,  
 $x_9$  – mitoses.

## Klasės:

C – class (**b**enign, **m**alignant)

dimensija (matmenų skaičius)

$n = 9$

viena eilutė – vienos pacientės duomenys

# Krūties vėžio duomenų klasifikavimas

Matricinis vaizdavimas

i L)	atr7 (REAL)	atr8 (REAL)	atr9 (REAL)	probability_2 (REAL)	probability_4 (REAL)	CLASS (2, 4)
	3	1	1	1	0	2
	3	2	1	0.64	0.36	2
	3	1	1	1	0	2
	3	7	1	0.786667	0.213333	2
	3	1	1	1	0	2
	9	7	1	0	1	4
	3	1	1	0.933333	0.0666667	2
	3	1	1	1	0	2

Class:  
2 – **benign**  
4 – **malignant**

Atsisiųsti

Atšaukti

# Duomenų grupavimas (klasterizavimas)

MII klasteris

MIF VU SK2

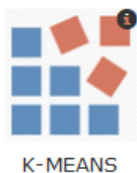
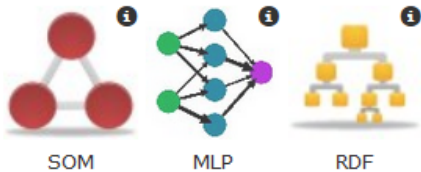
▶ Duomenų įkėlimas

▶ Pirminis apdorojimas

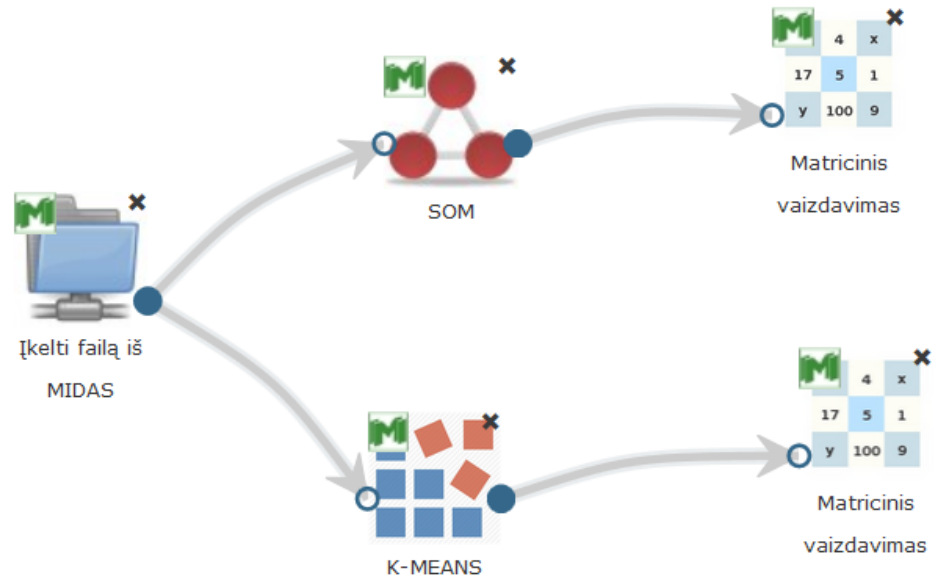
▶ Statistiniai primityvai

▶ Dimensijos mažinimas

▼ Klasifikavimas, grupavimas



▶ Rezultatų peržiūra



Naujas eksperimentas

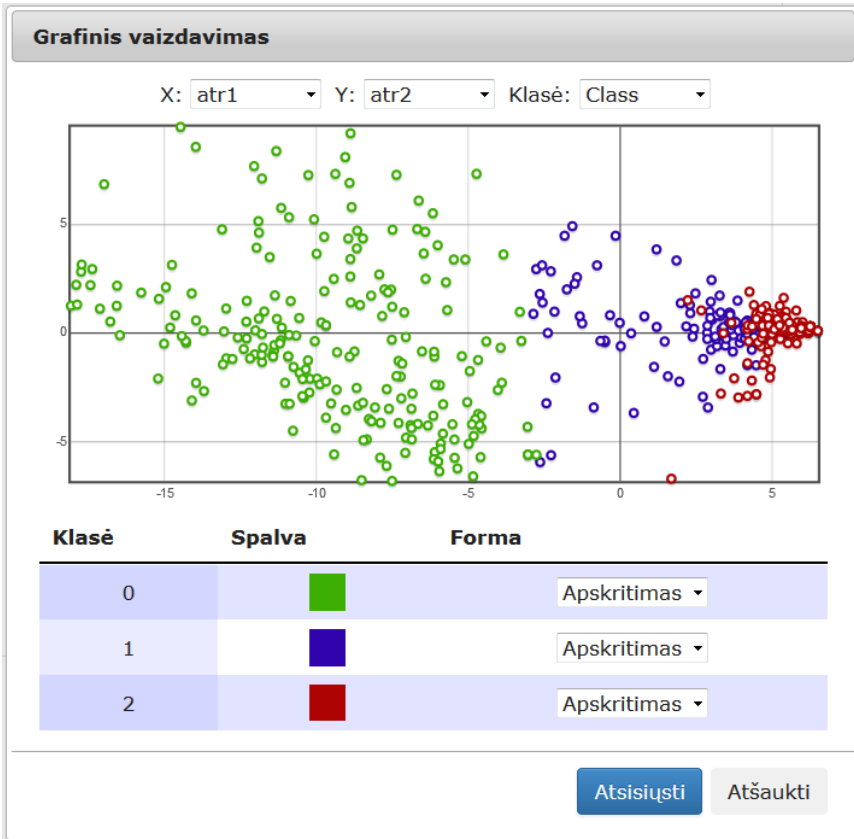
Išsaugoti

Vykdyti

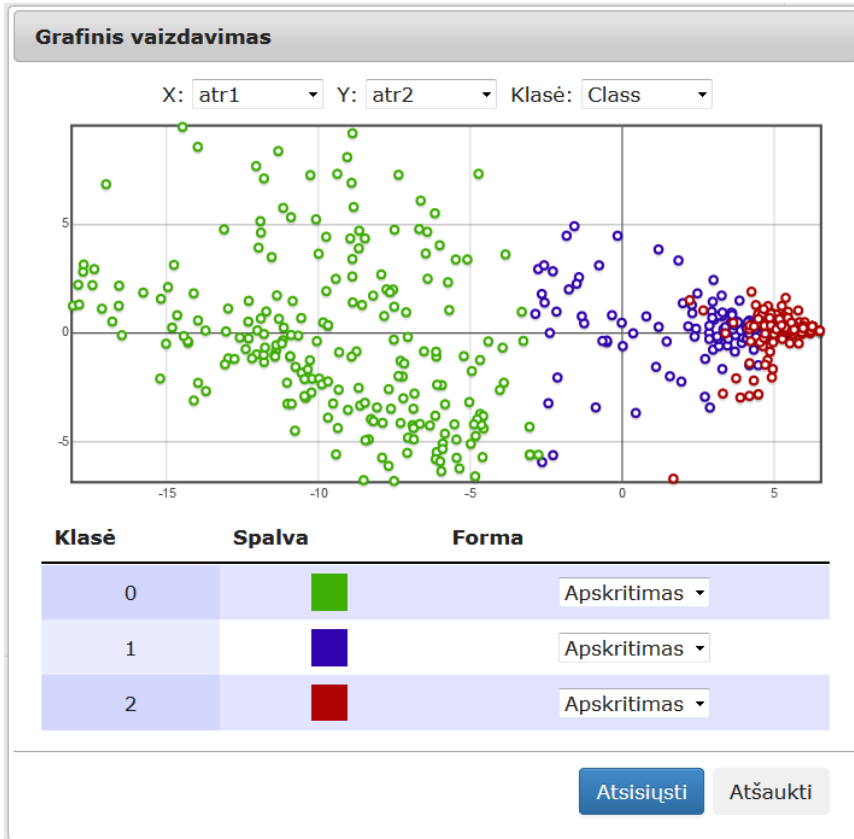
# Duomenų grupavimas (klasterizavimas)

- **Klasterizavimo tikslas** – suskirstyti duomenis į grupes pagal jų panašumą, taip, kad toje pačioje grupėje duomenys būtų panašūs, o duomenys iš skirtingų grupių būtų nepanašūs.
- Dažniausiai klasterizavimo uždaviniai yra sprendžiami tada, **kai nežinomos duomenų klasės**.
- Įgyvendinti **šie duomenų grupavimo (klasterizavimo) metodai**:
  - k-means – k-vidurkių klasterizavimo metodas,
  - SOM – saviorganizuojantis neuroninis tinklas.

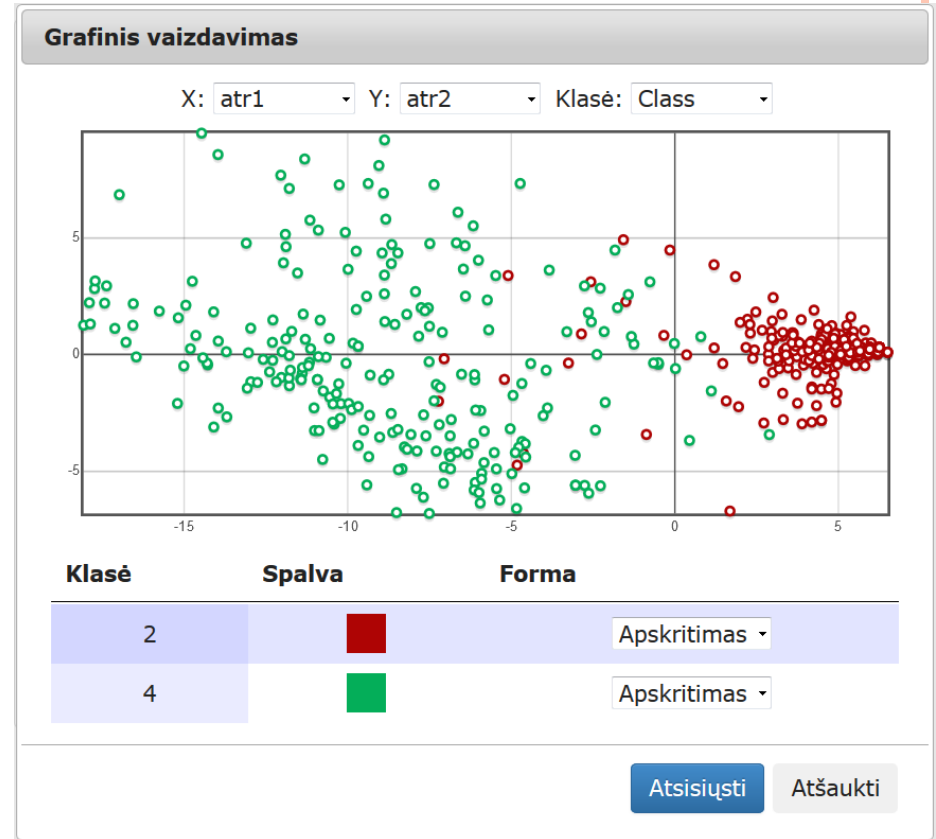
# Krūties vēžio duomenų grupavimas ir vizualizavimas



# Krūties vėžio duomenų grupavimas, klasifikavimas ir vizualizavimas



Grupavimo į tris grupes rezultatas

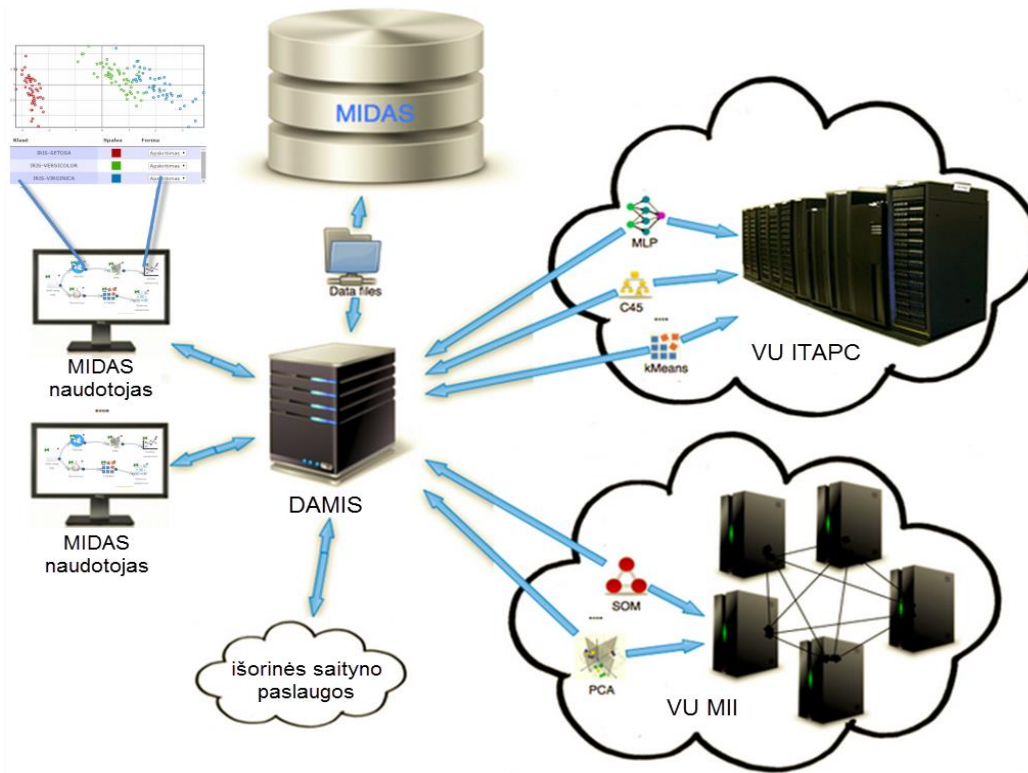


Klasių vizualizavimas

# Kodėl verta rinktis DAMIS?



- **Praplečia MIDAS galimybes**: tai duomenų archyvas, turintis duomenų analizės galimybę visų sričių mokslininkams.
- **Nereikia** jokios **papildomos** programinės **įrangos**.
- Galimybė **rinktis** lygiagrečiųjų ir paskirstytųjų skaičiavimų **resursus** iš siūlomų alternatyvų.
- Tyrėjas norėdamas spręsti tam tikrą duomenų analizės uždavinį, tačiau eksperimentą vykdyti su kitais duomenimis ar kitais algoritmų valdymo parametrais, gali panaudoti jau sukurtas **mokslinių užduočių sekas**.
- Numatytos **DAMIS plėtros galimybės** pagal mokslininkų poreikį.



**Ačiū už dėmesį**

Pastabų, komentarų lauksime ir el. paštu [info@inscience.lt](mailto:info@inscience.lt)