VšĮ „Informatikos mokslų centras“

MOKSLAS · EKONOMIKA · SANGLAUDA

EUROPOS SĄJUNGA

EUROPOS REGIONINĖS PLĖTROS FONDAS

Kuriame Lietuvos ateitį

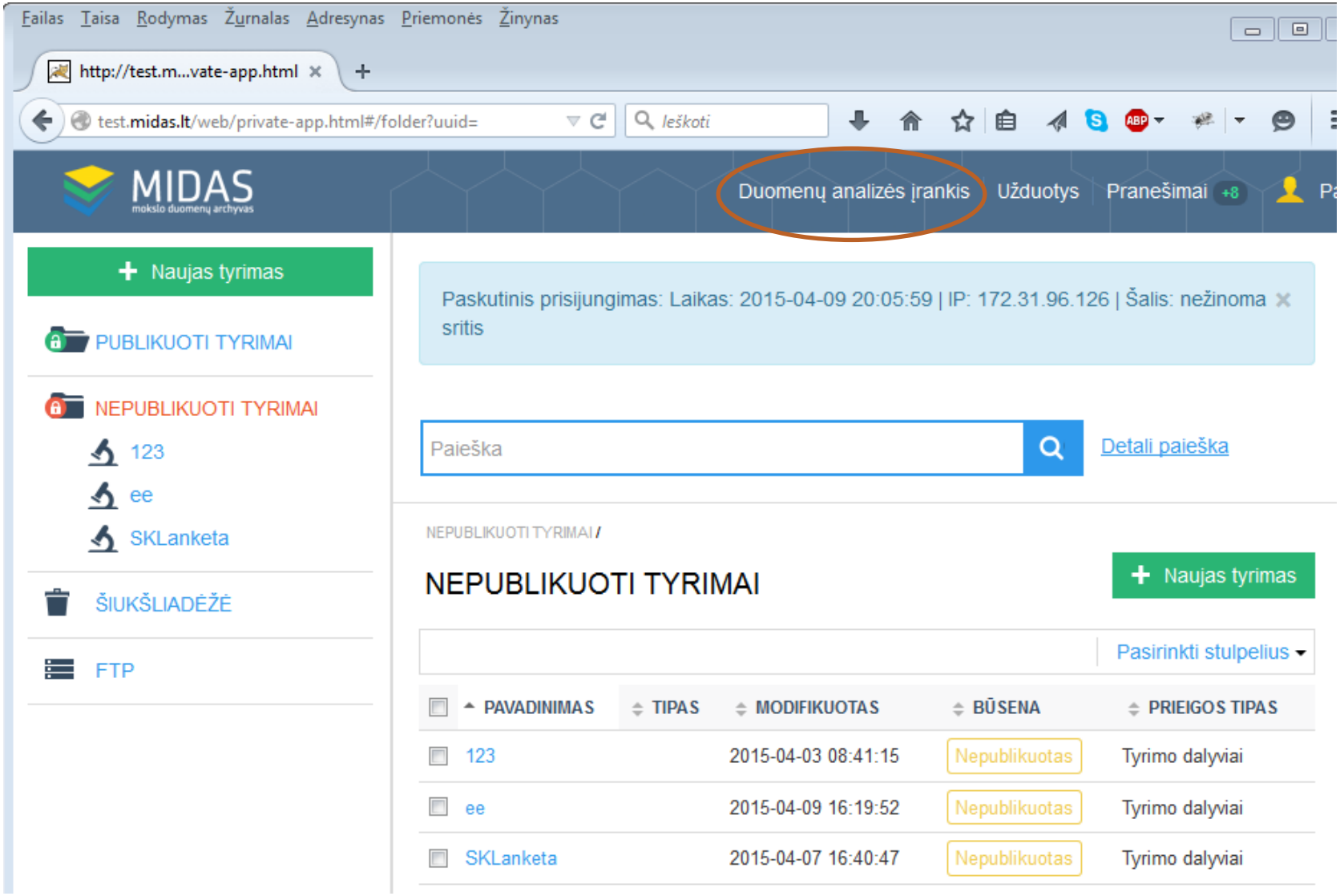
Duomenų analizės įrankis DAMIS

dr. Jolita Bernatavičienė

Kaunas, 2015

FOSTER

DAMIS – nacionalinio mokslo informacijos archyvo MIDAS duomenų analizės įrankis



DAMIS funkcionalumas

MIDAS
mokslo duomenų archyvas

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

DAMIS funkcionalumas

MIDAS
mokslo duomenų archyvas

- 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.

DAMIS funkcionalumas

MIDAS
mokslo duomenų archyvas

- 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

MIDAS
mokslo duomenų archyvas

- 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

MIDAS
mokslo duomenų archyvas

- 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:

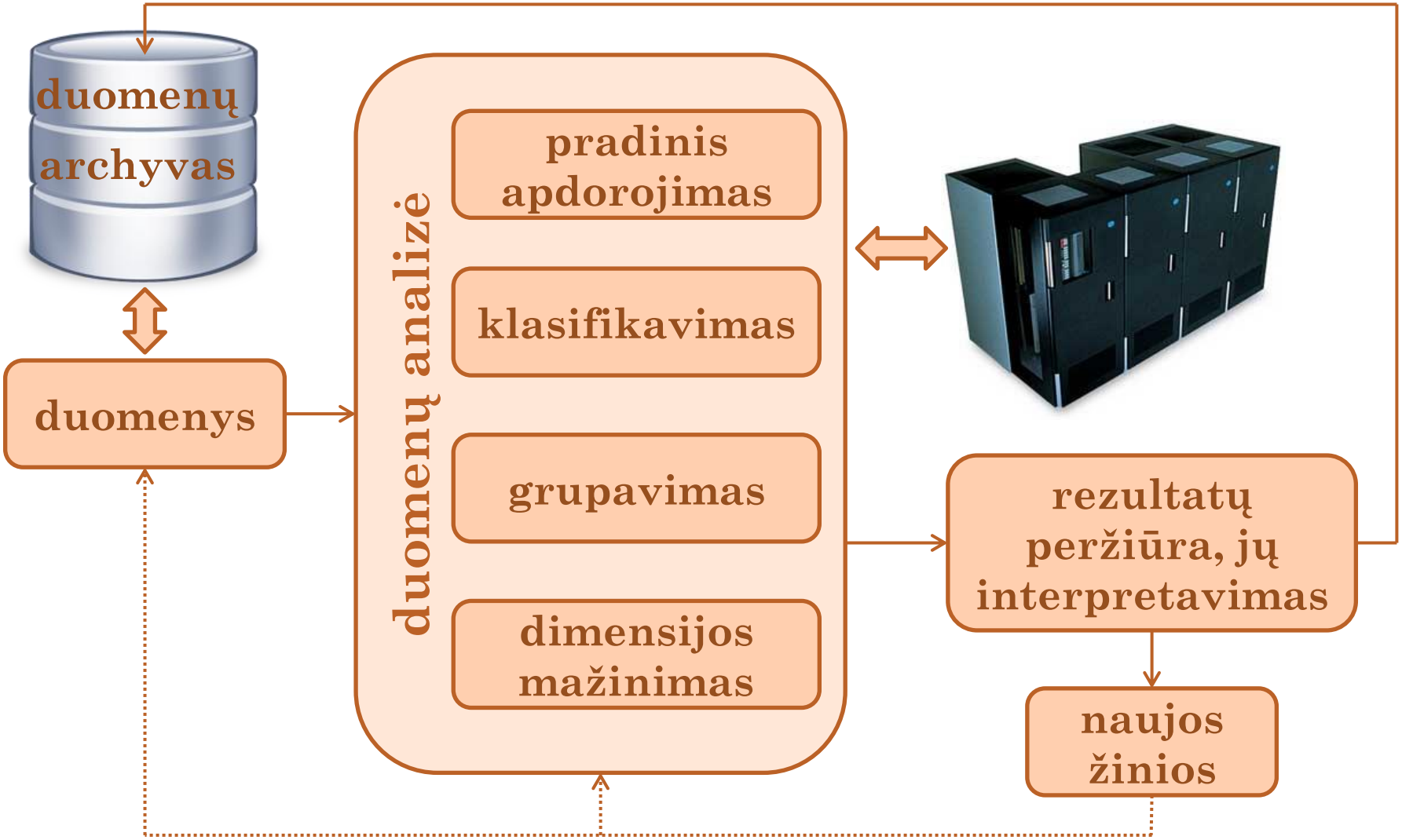
C – class (**b**enign, **m**alignant)

dimensija (matmenų skaičius) $n = 9$

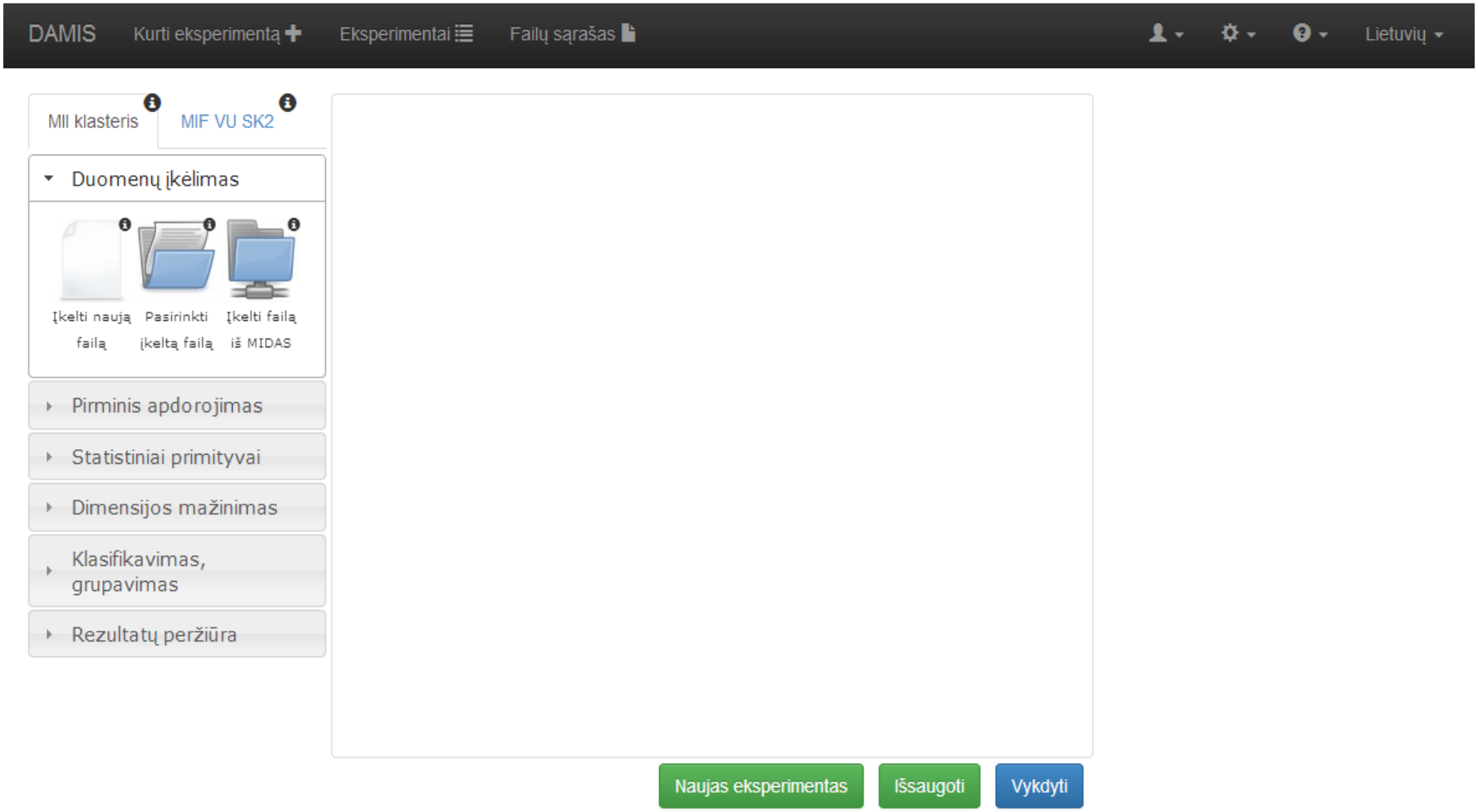
viena eilutė – vienos pacientės duomenys

Duomenų analizė

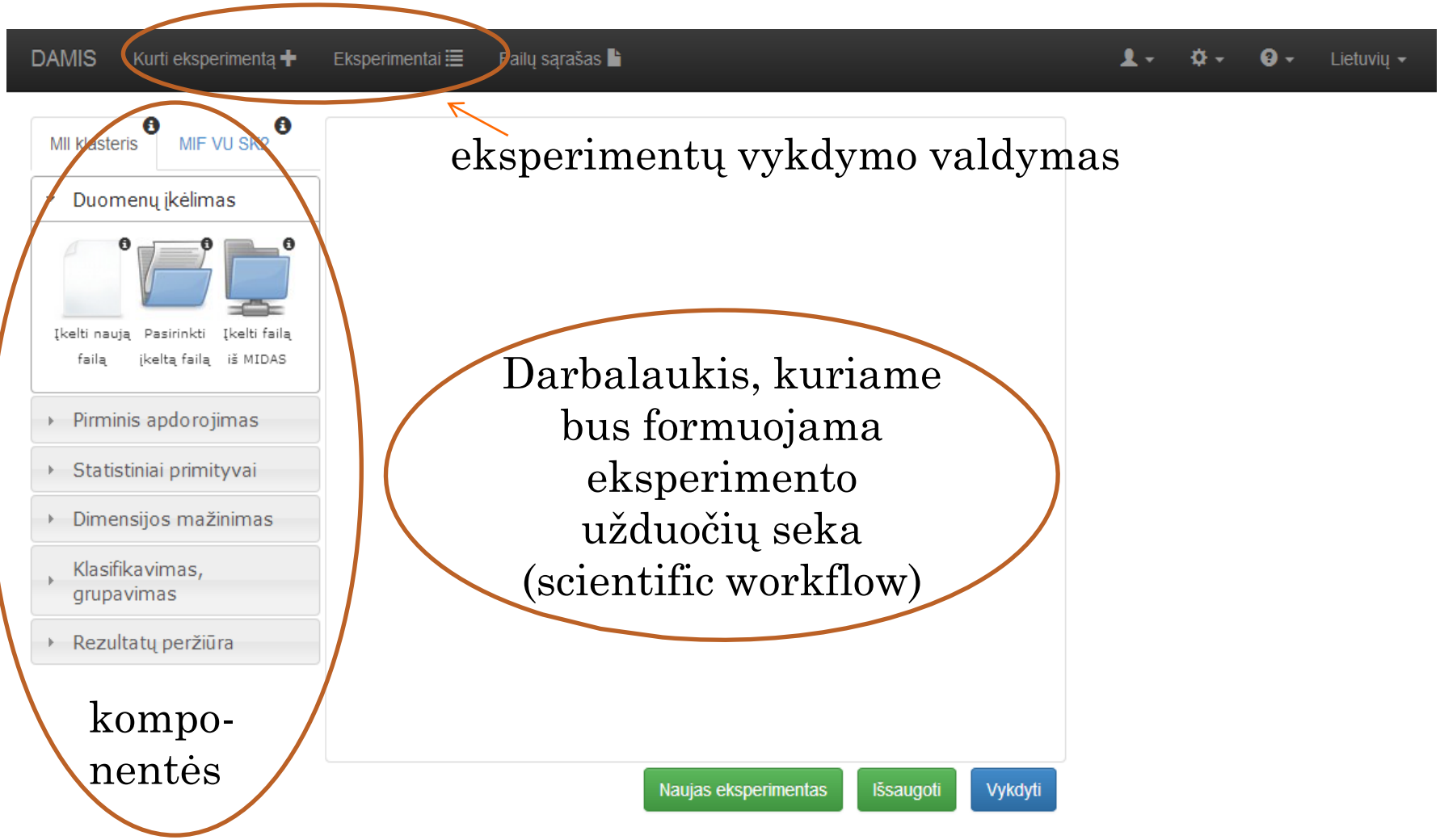
MIDAS
mokslo duomenų archyvas



DAMIS grafinė naudotojo sąsaja



DAMIS grafinė naudotojo sąsaja

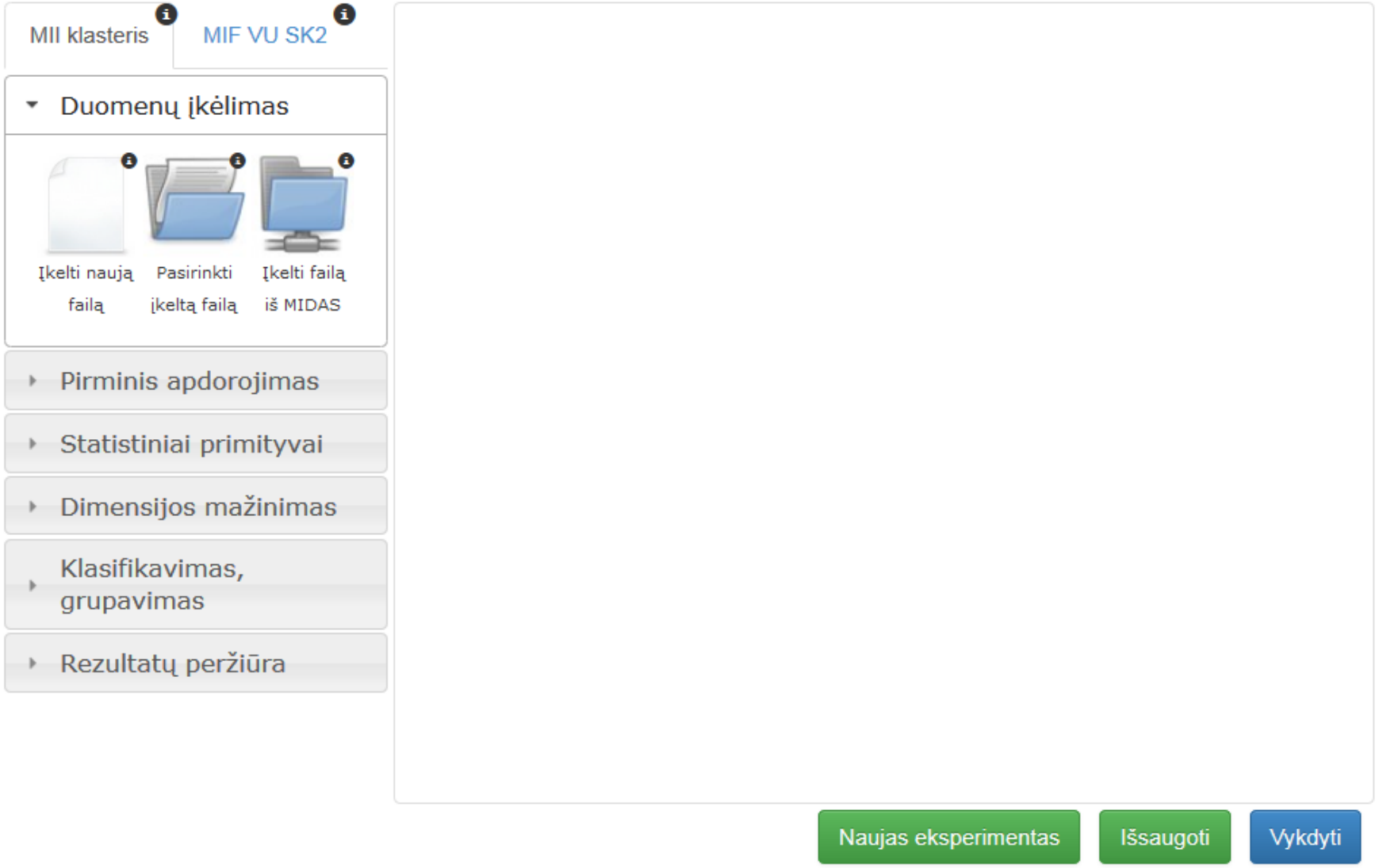


DAMIS duomenų analizei

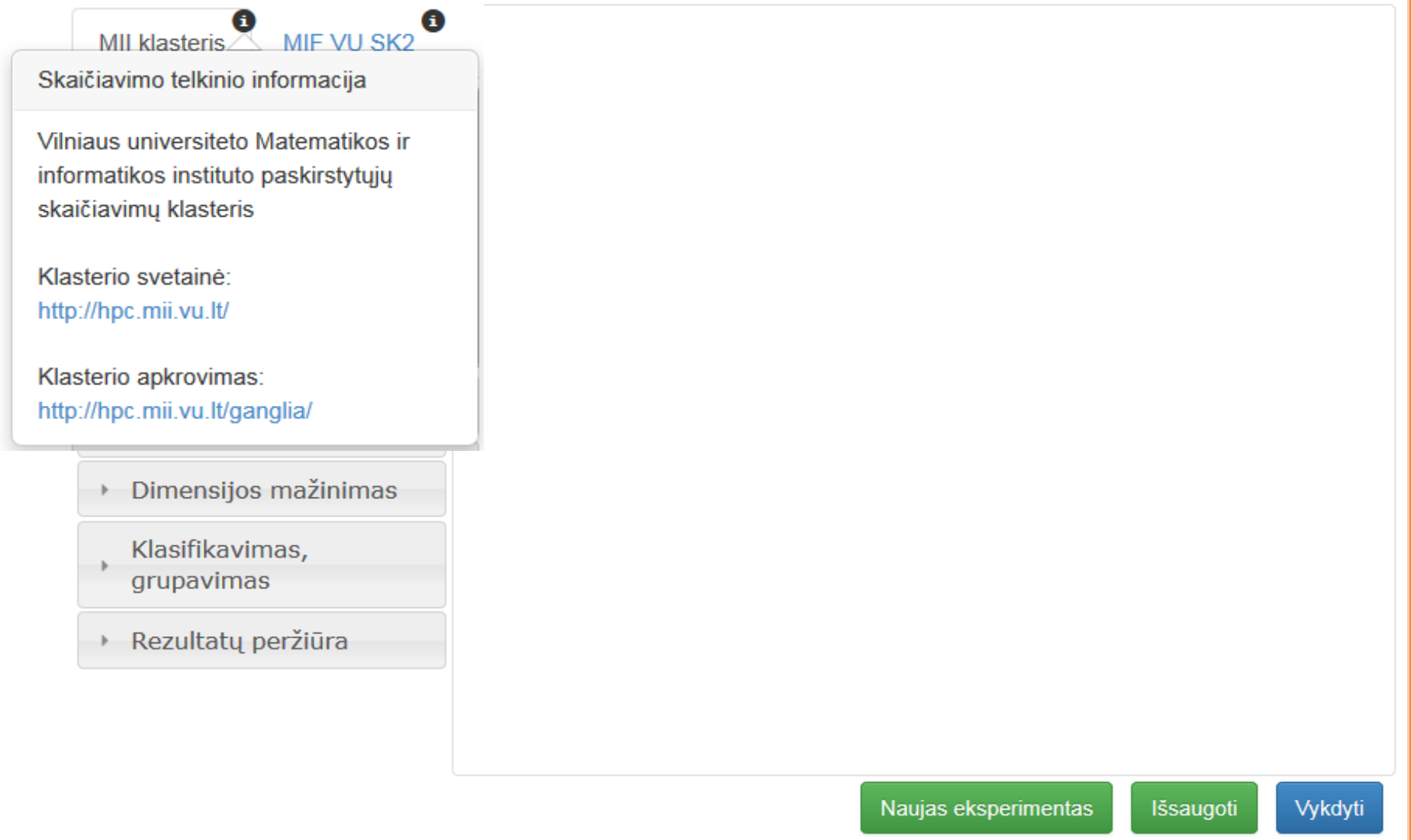
MIDAS
mokslo duomenų archyvas

- 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.

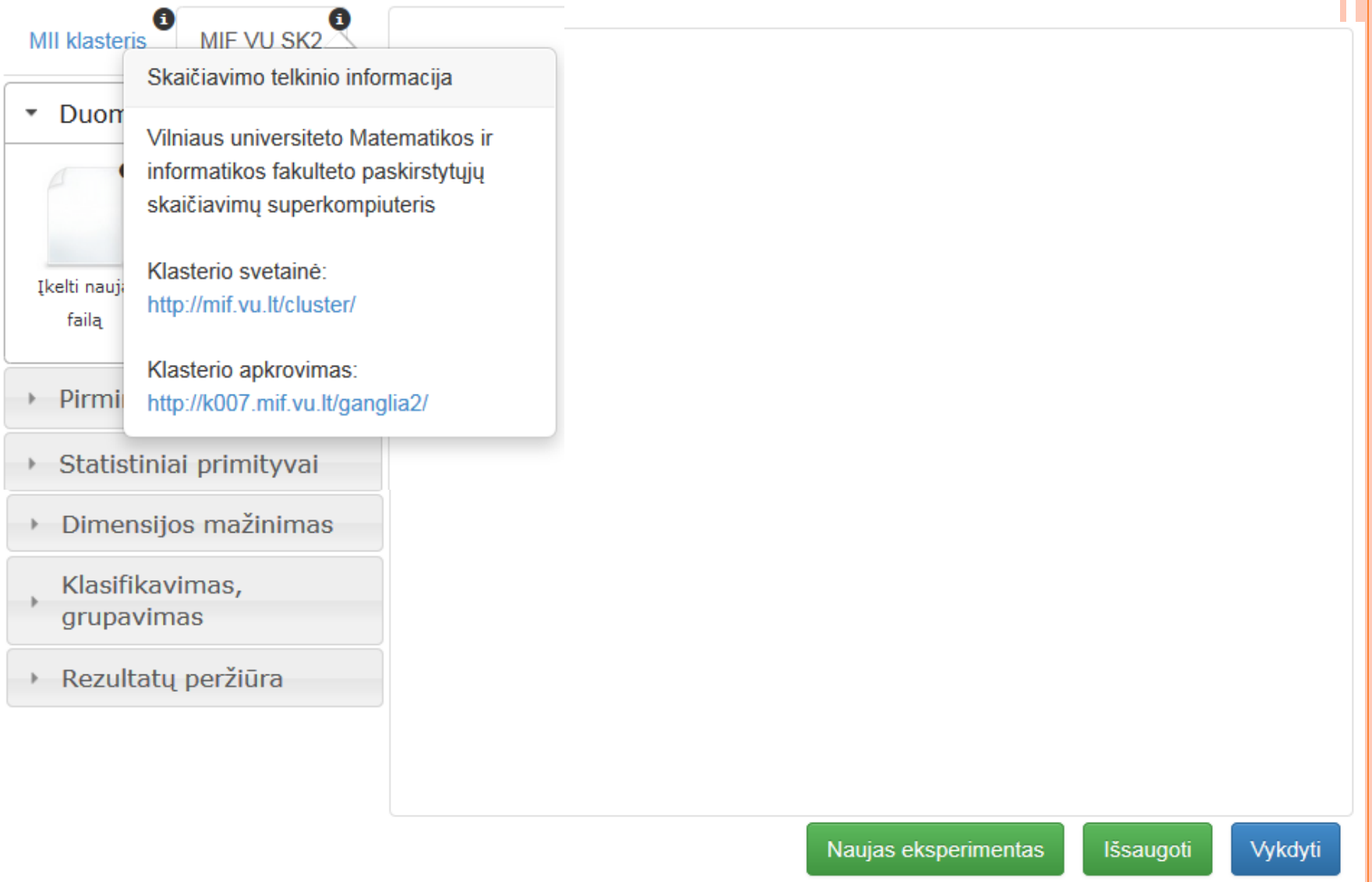
Skaičiavimų resursų parinkimas



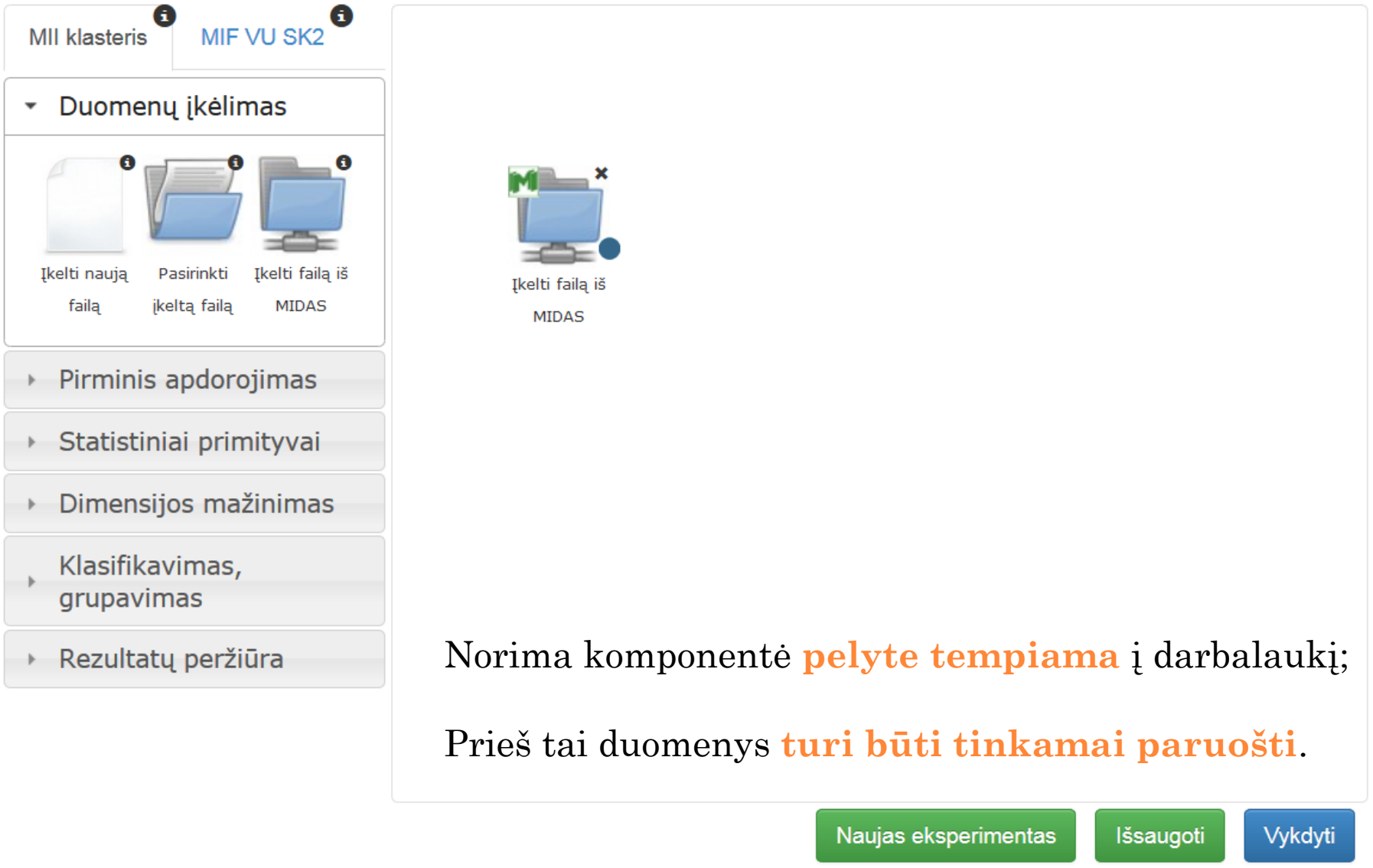
Skaičiavimų resursų parinkimas



Skaičiavimų resursų parinkimas



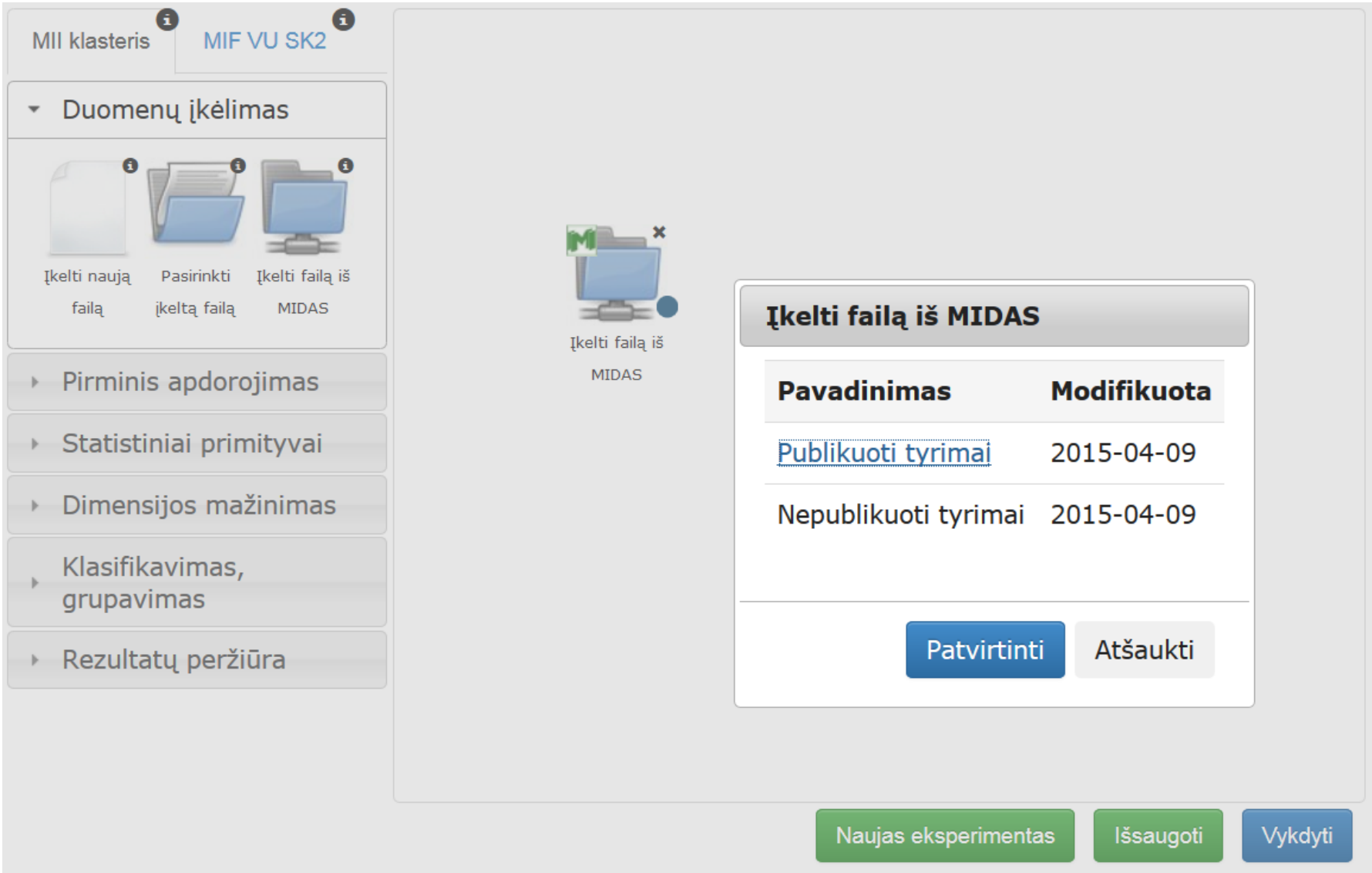
Duomenų įkėlimas



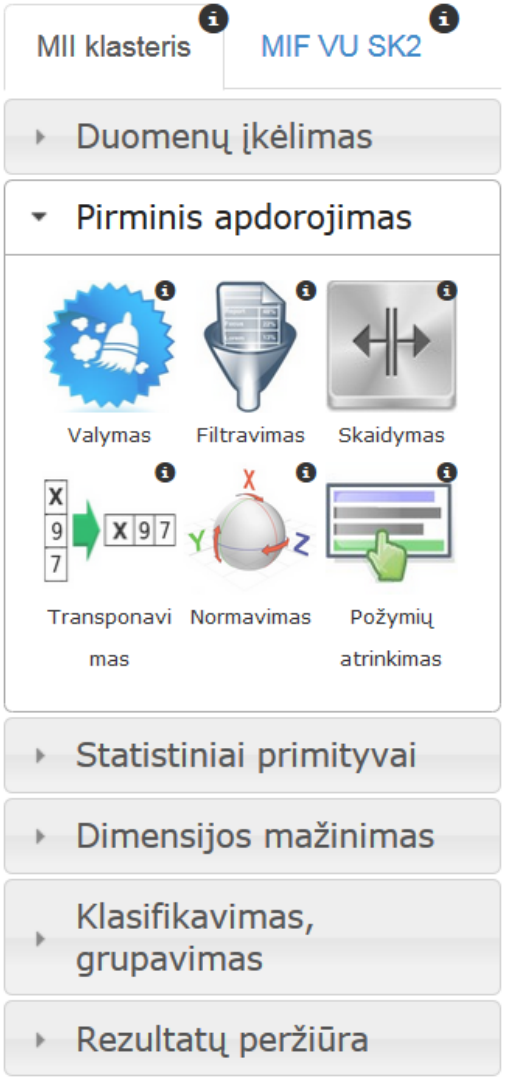
Norima komponentė **pelyte tempiama** į darbalaukį;

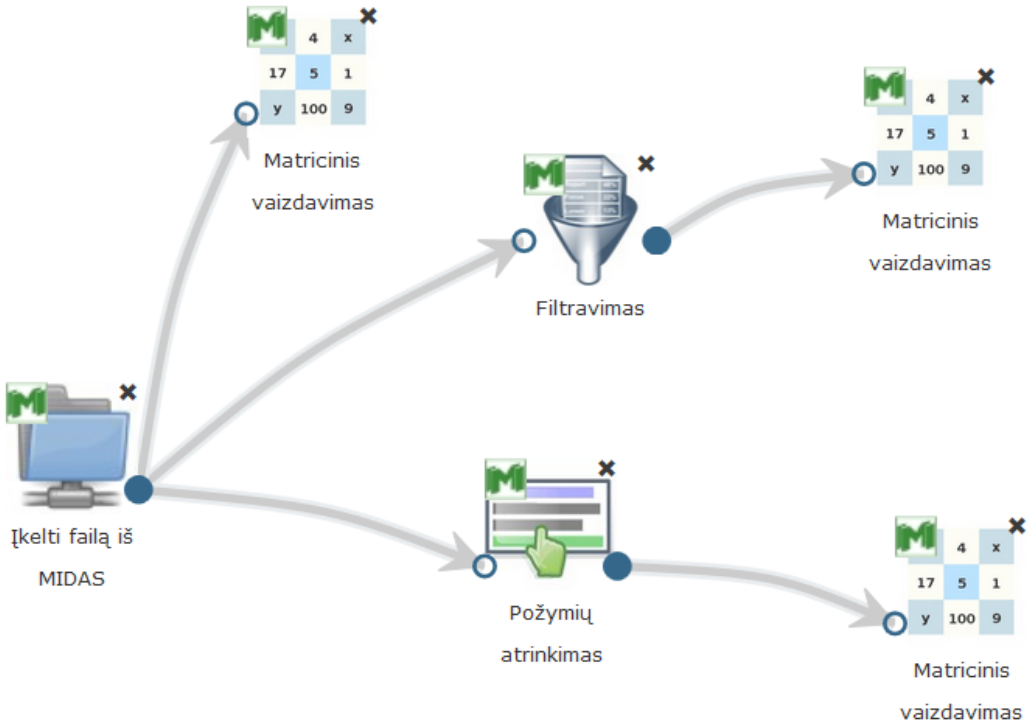
Prieš tai duomenys **turi būti tinkamai paruošti**.

Duomenų įkėlimas iš MIDAS



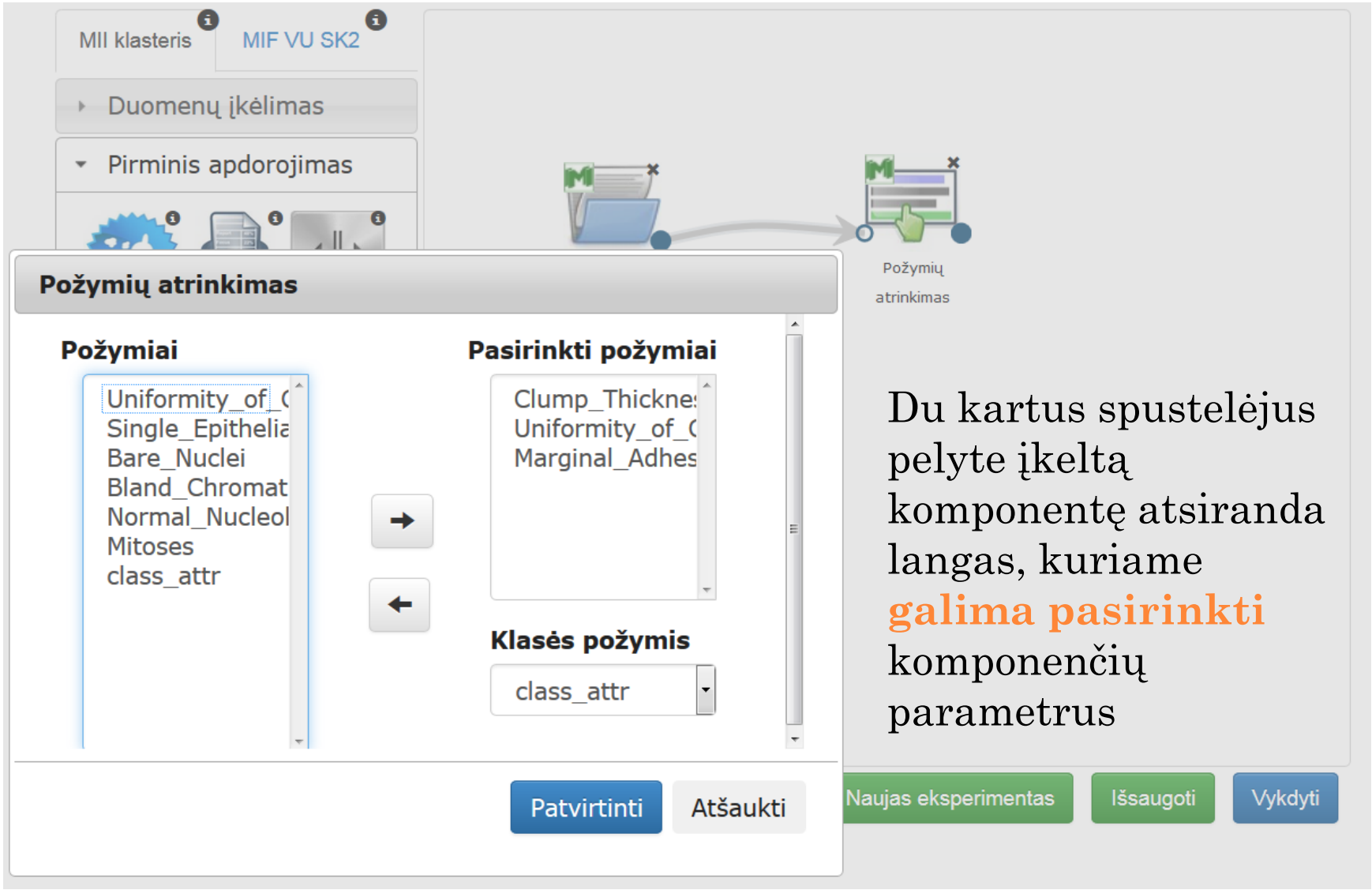
Duomenų pirminis apdorojimas





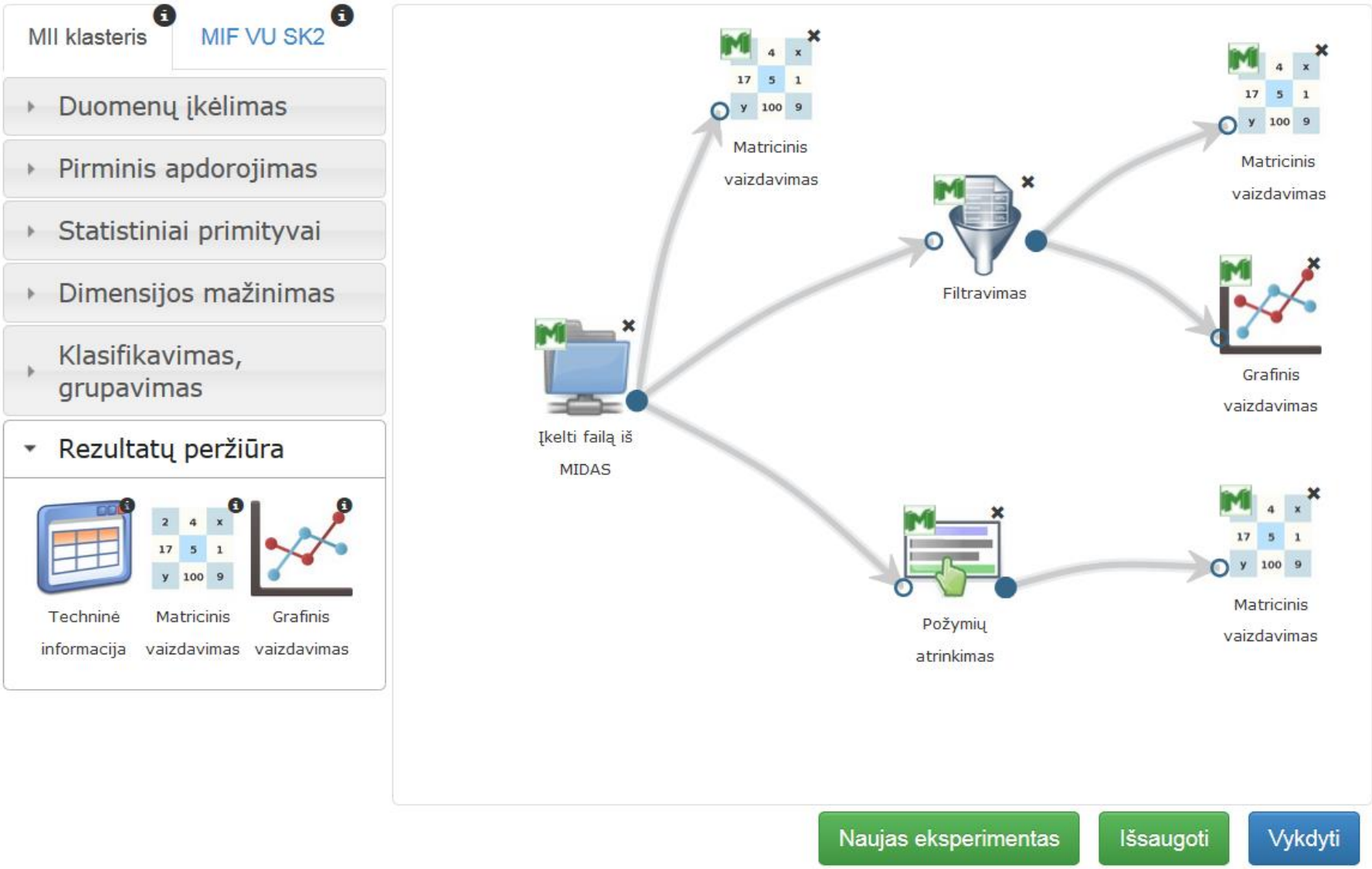
Įkeltos komponentės sujungiamos į vadinamąsias **mokslinių užduočių sekas** (scientific workflows)

Komponenčių parametrų parinkimas

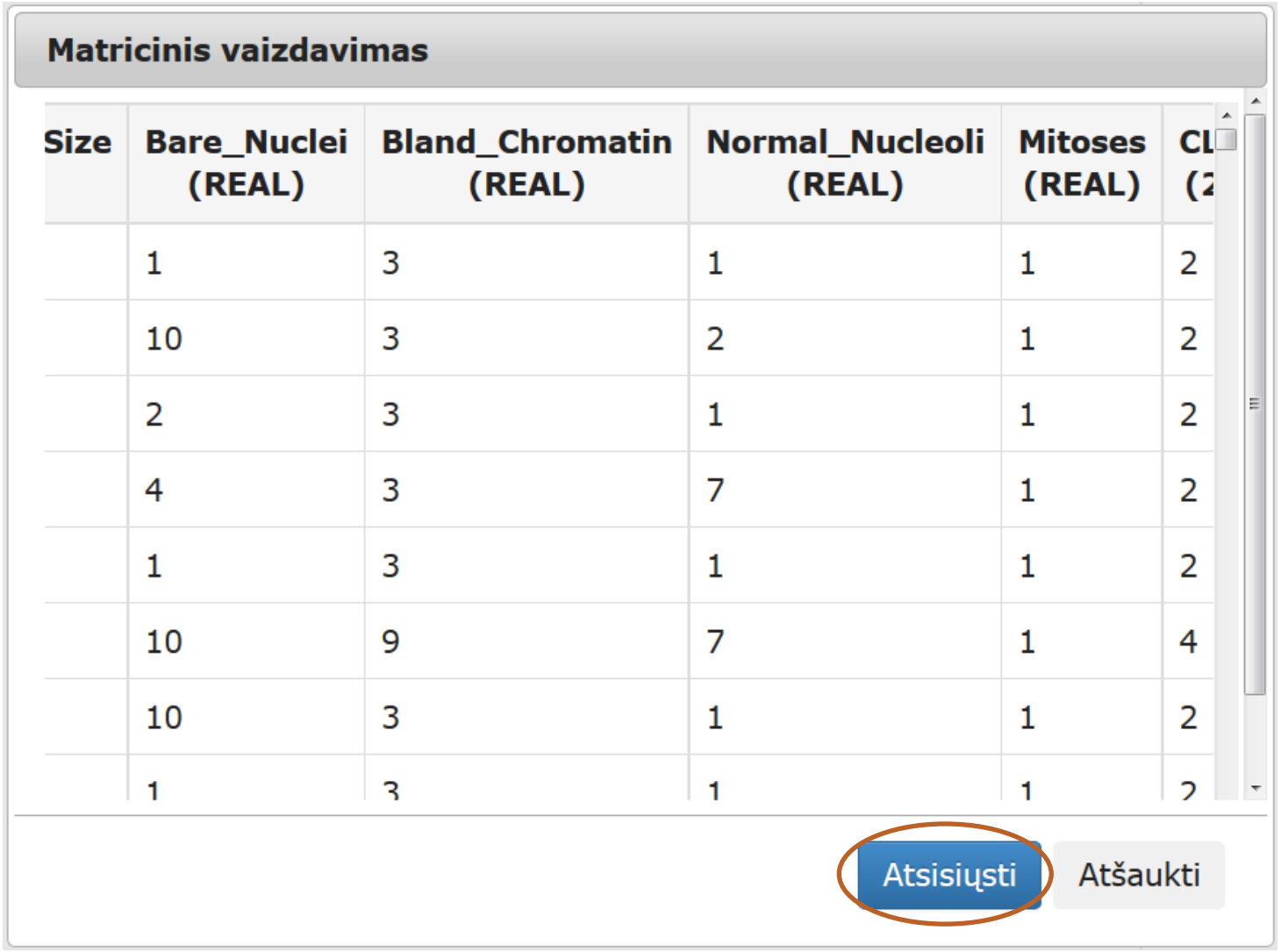


Du kartus spustelėjus pelyte įkeltą komponentę atsiranda langas, kuriame **galima pasirinkti** komponenčių parametrus

Rezultatų peržiūra



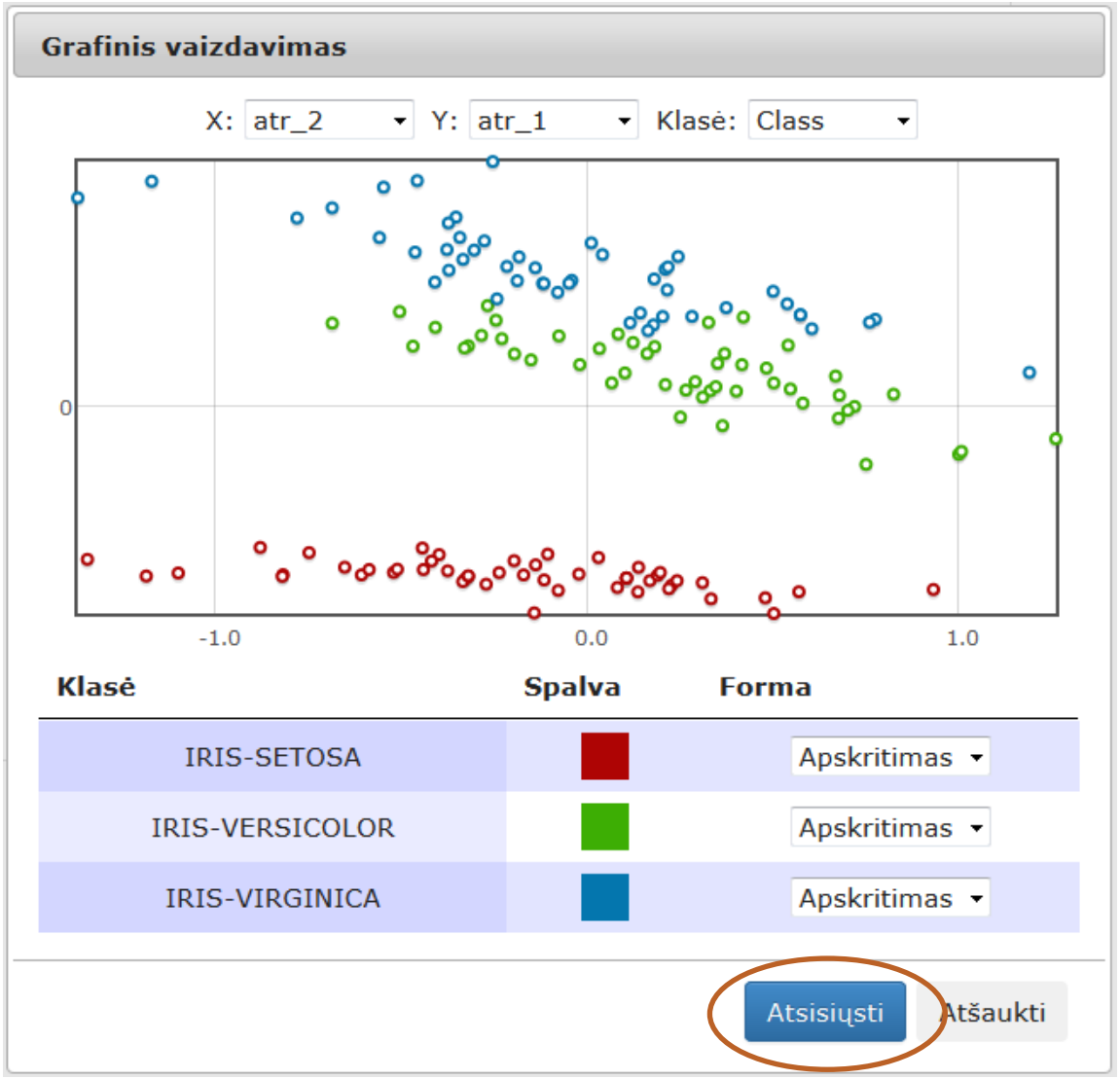
Rezultatų matricinis vaizdavimas



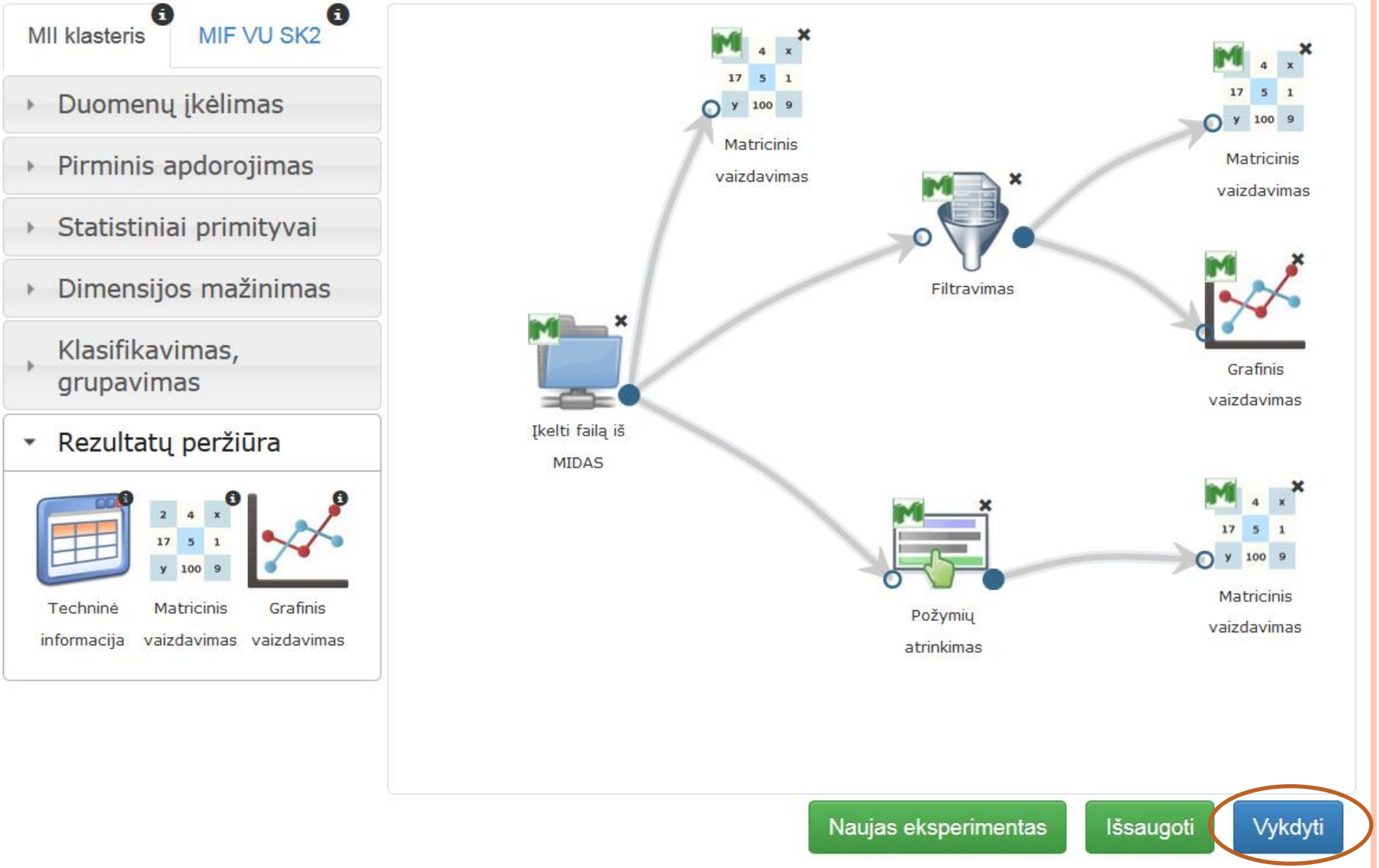
Rezultatų grafinis vaizdavimas

- 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**.

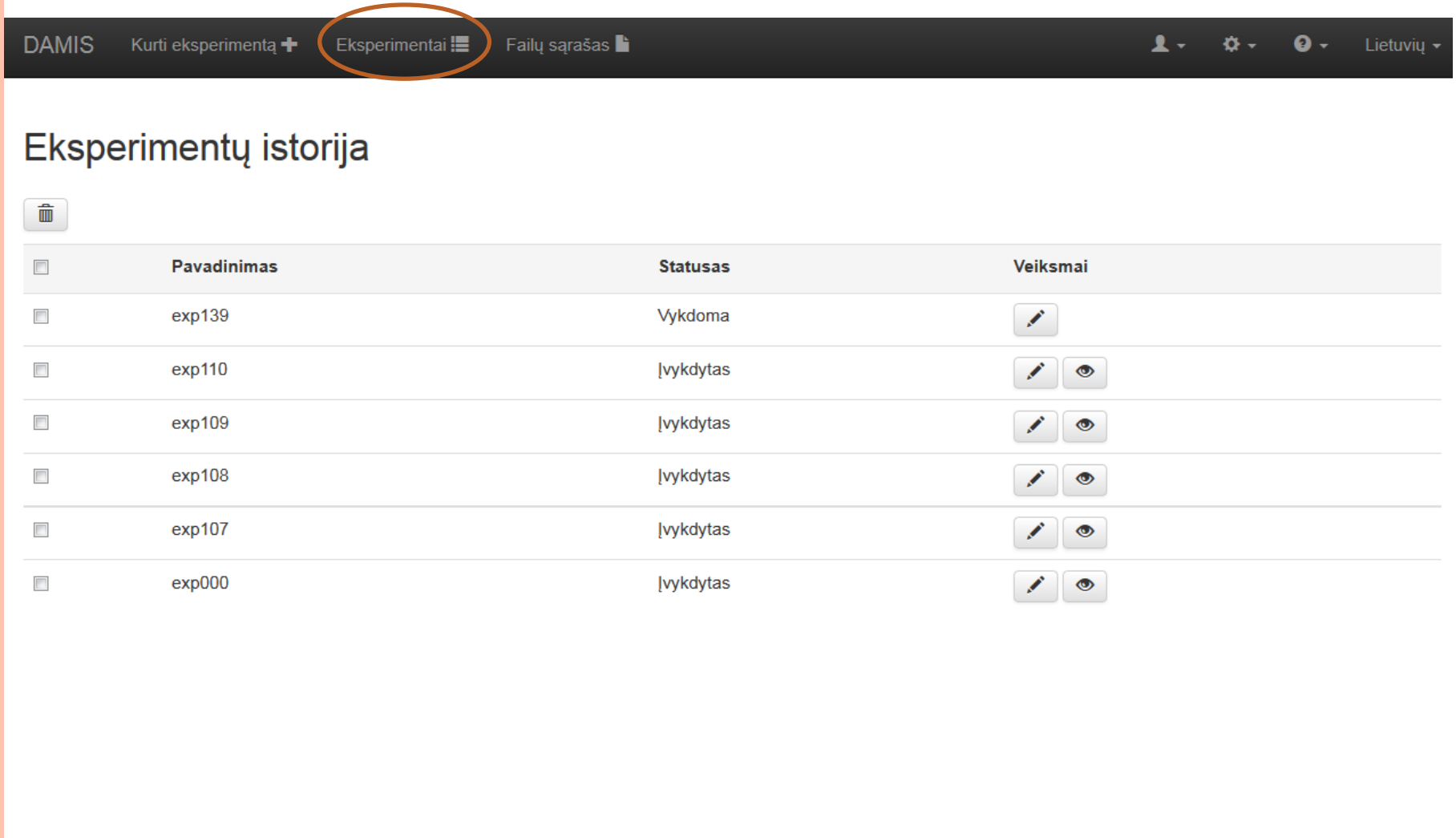
Rezultatų grafinis vaizdavimas



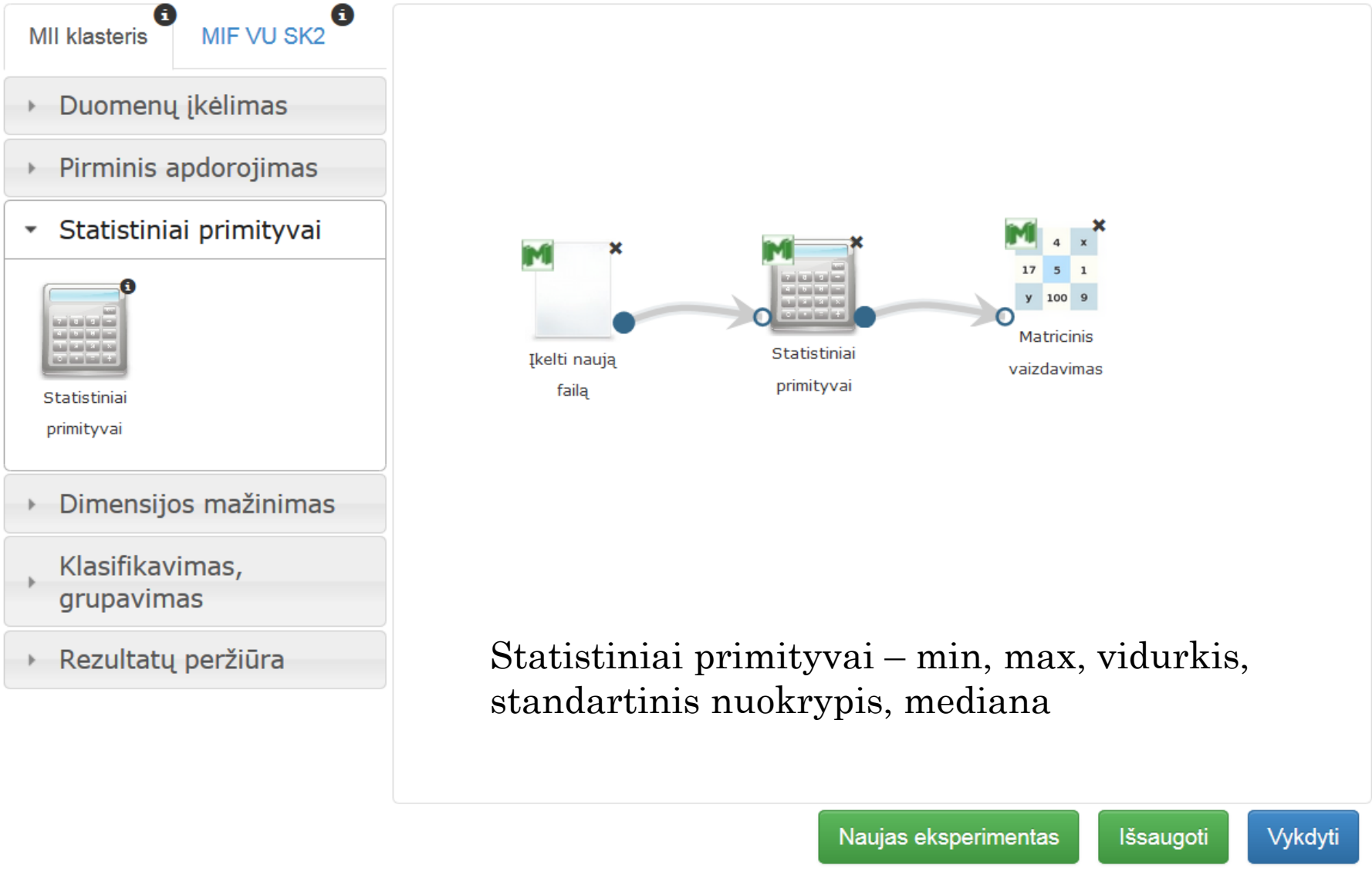
Eksperimento vykdymas



Eksperimentų vykdymo istorija

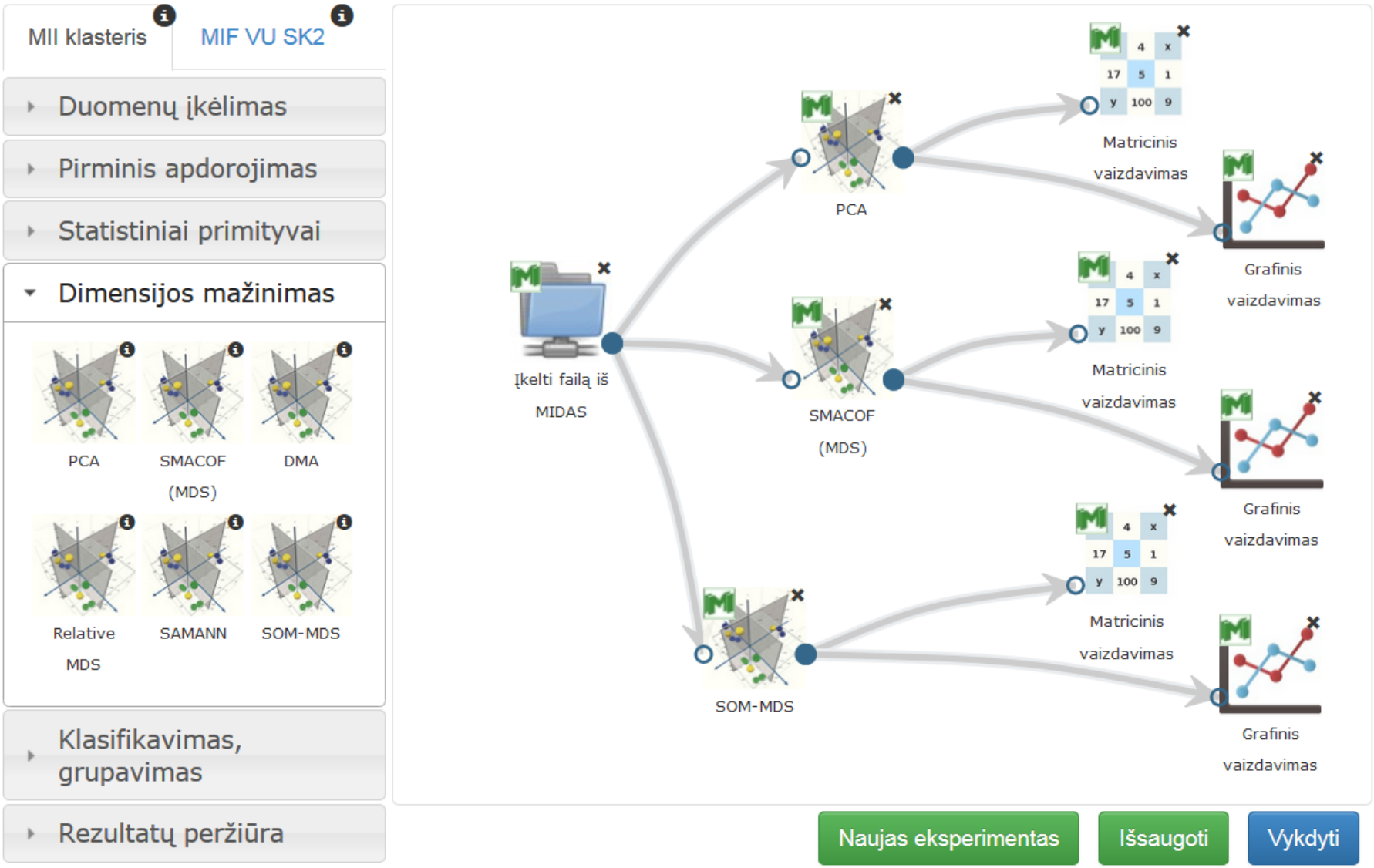


Duomenų statistinės charakteristikos



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

Duomenų dimensijos mažinimas



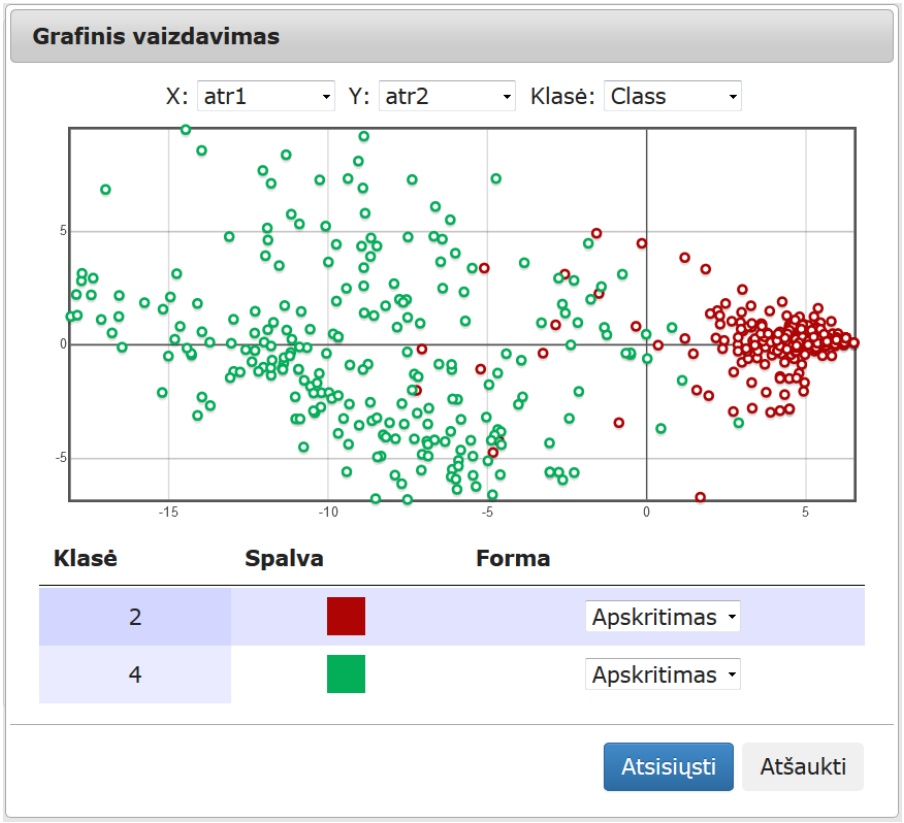
Duomenų dimensijos mažinimas

- 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 komponenčių 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

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



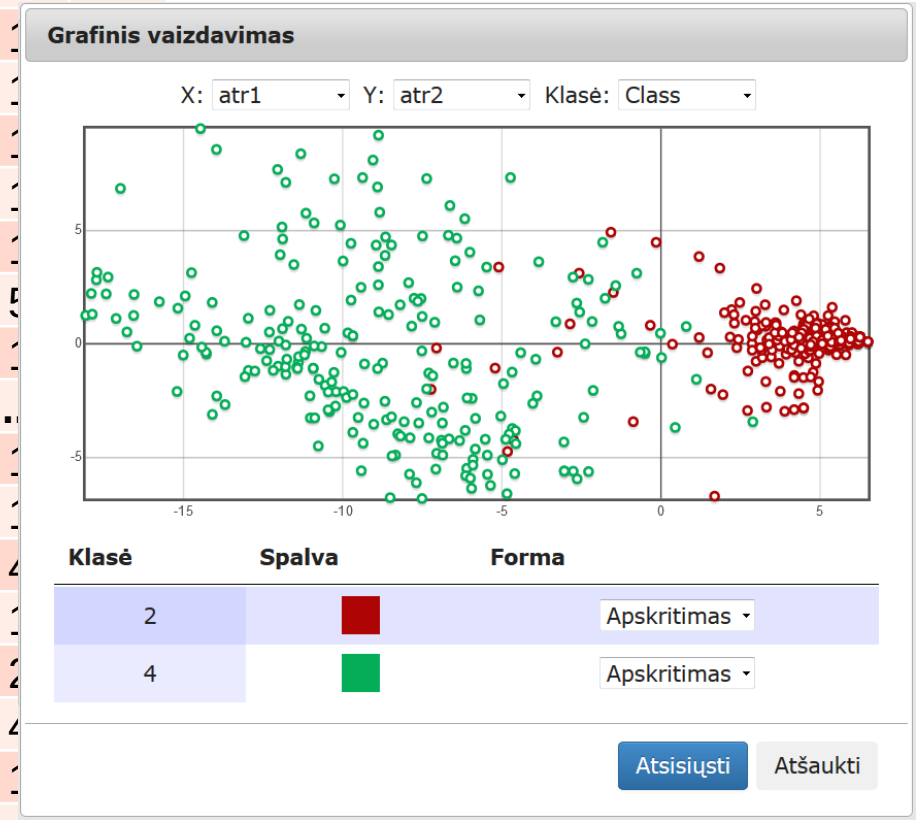
- vienas **taškas** atitinka vieną **pacientę**
- vienas **taškas** apjungia visus **devynis požymius**
- vizualus pateikimas padeda lengviau suvokti **informacijos visumą**

2 – **b**enign, 4 – **m**alignant

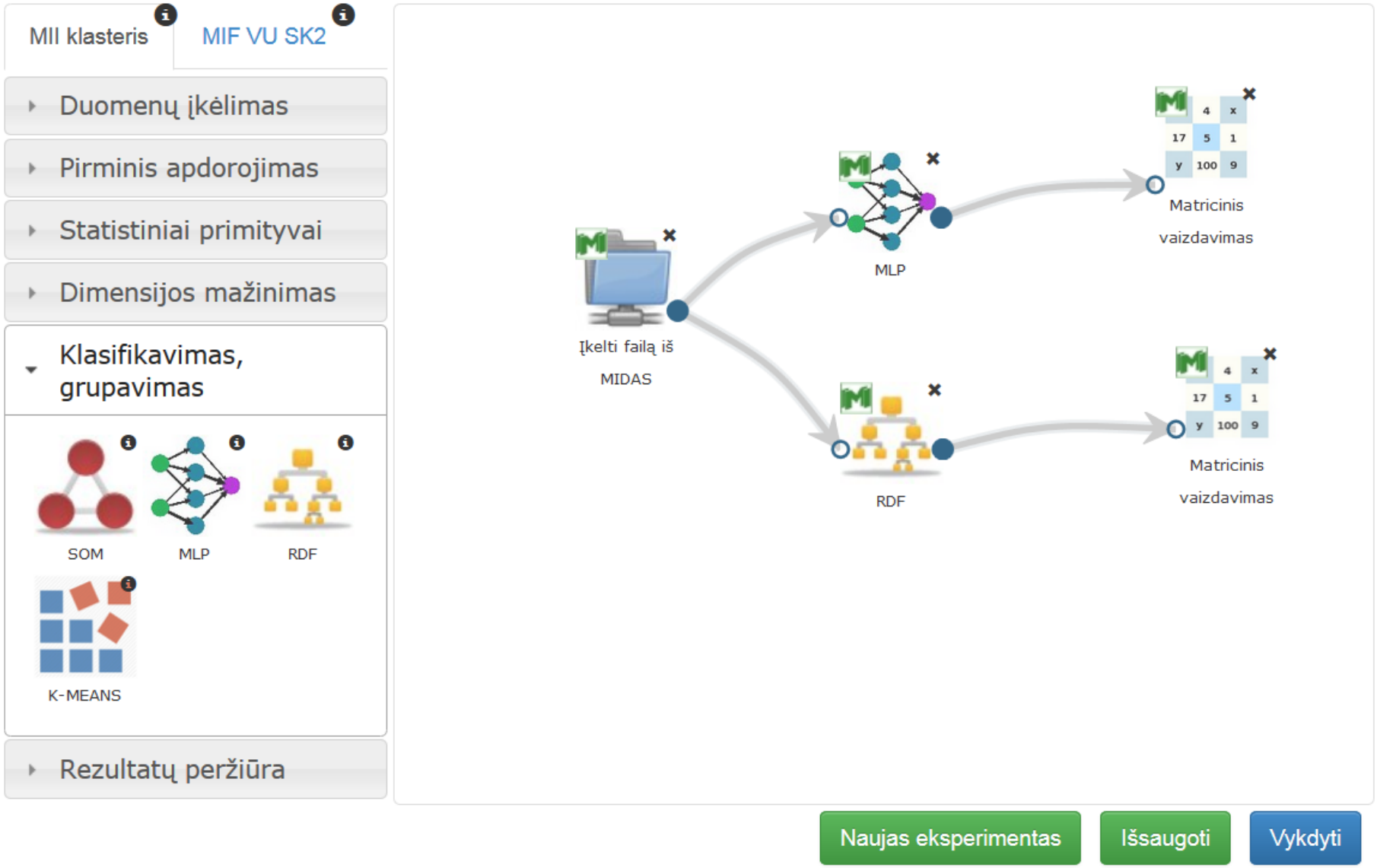
Krūties vėžio duomenų dimensijos mažinimas (vizualizavimas) MDS metodu

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	
6	8	8	1	3	4	3	7	1	
4	1	1	3	2	1	3	1	1	
1	1	1	1	2	10	3	1	1	
2	1	2	1	2	1	3	1	1	
2	1	1	1	2	1	1	1	5	
4	2	1	1	2	1	2	1	1	
…	…	…	…	…	…	…	…	…	
8	10	10	8	7	10	9	7	1	
5	3	3	3	2	3	4	4	1	
8	7	5	10	7	9	5	5	4	
7	4	6	4	6	1	4	3	1	
10	7	7	6	4	10	4	1	2	
7	3	2	10	5	10	5	4	4	
10	5	5	3	6	7	7	10	1	
…	…	…	…	…	…	…	…	…	…
4	8	8	5	4	5	10	4	1	m

Class (C):
2 – **b**enign
4 – **m**alignant



Duomenų klasifikavimas



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

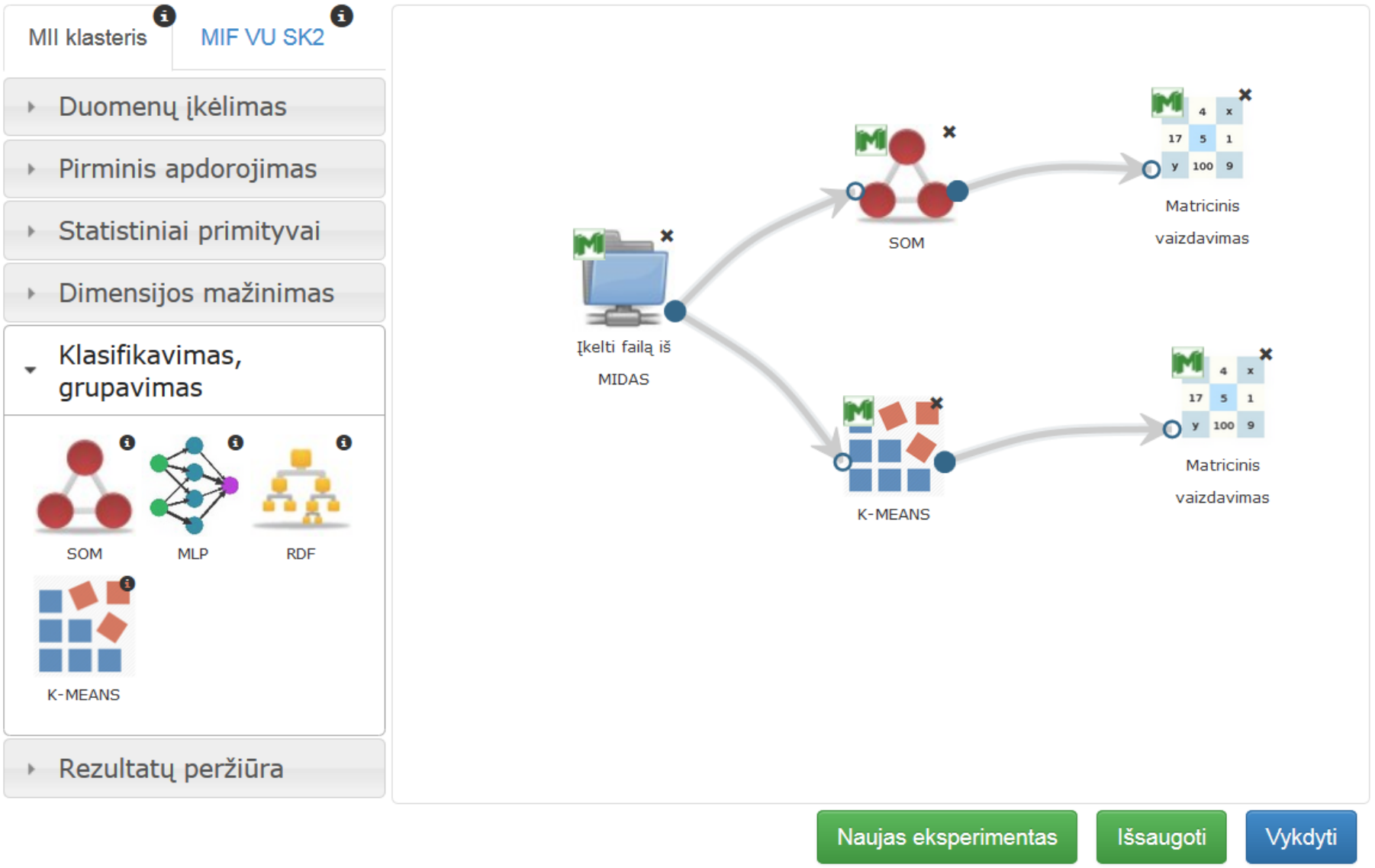
Matricinis vaizdavimas

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

Atsisiųsti Atšaukti

Class:
2 – **b**enign
4 – **m**alignant

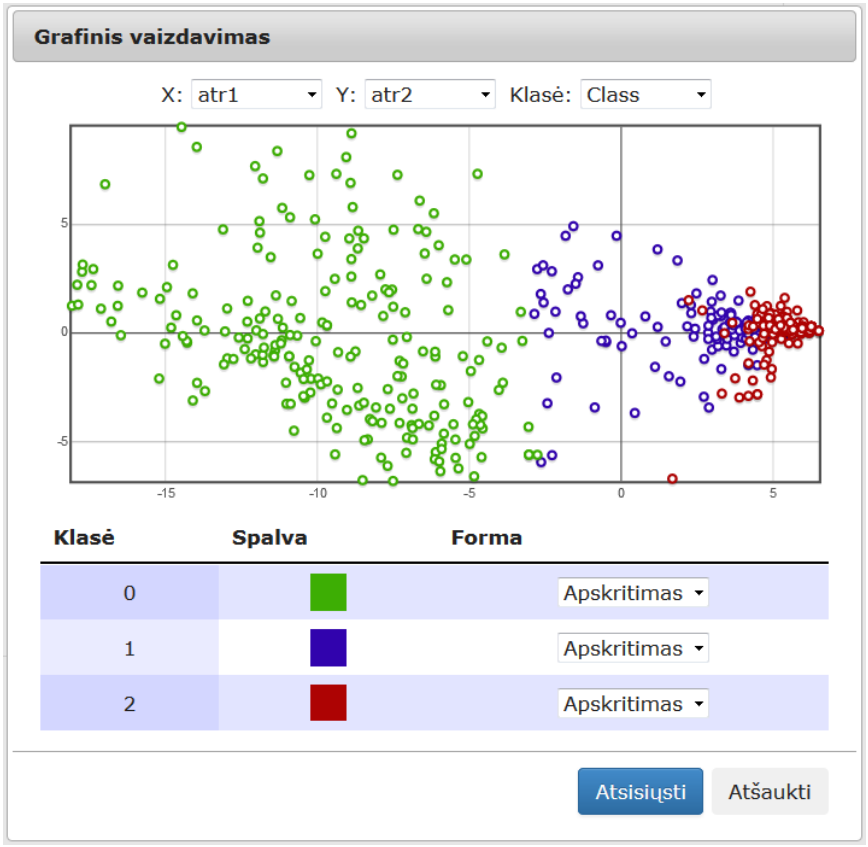
Duomenų grupavimas (klasterizavimas)



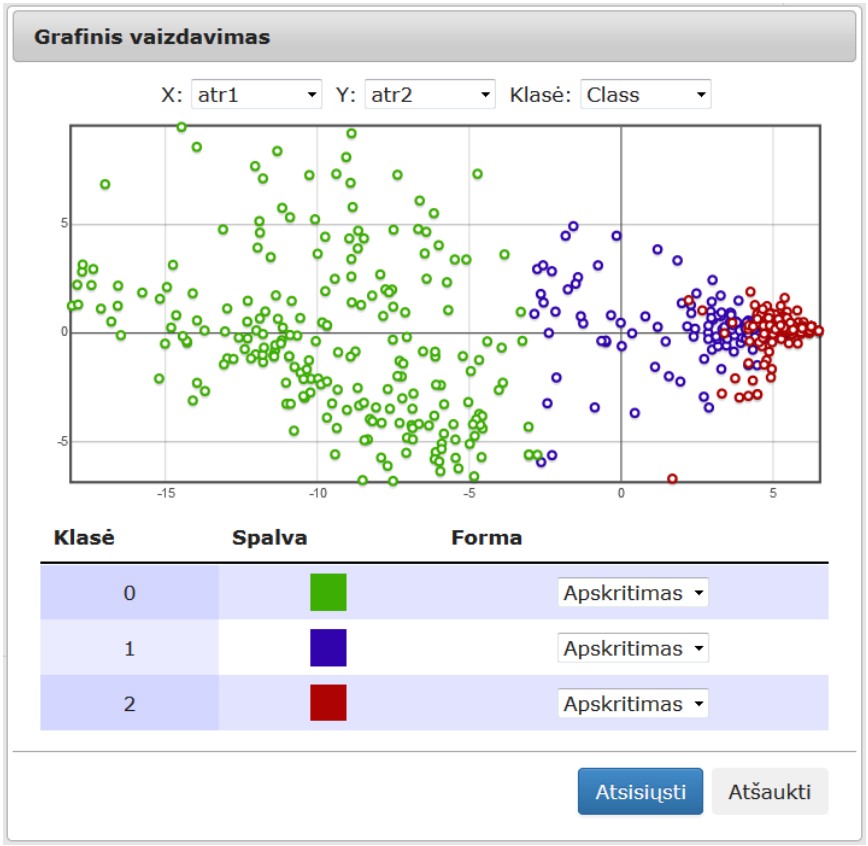
Duomenų grupavimas (klasterizavimas)

- **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.

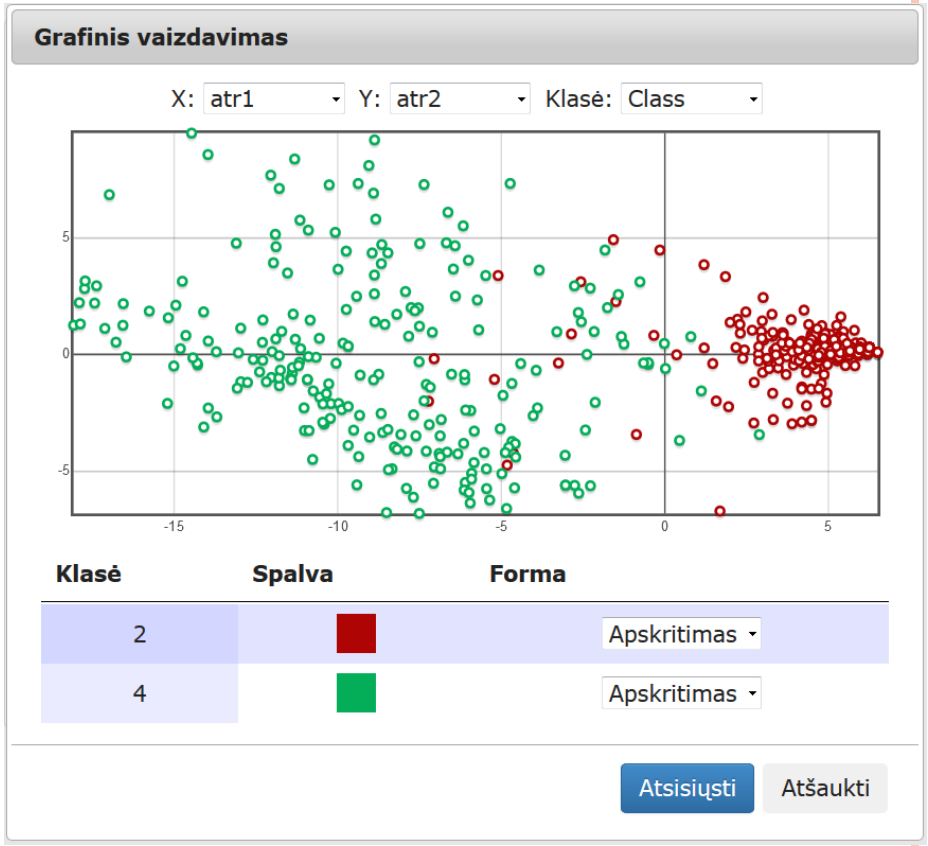
Krūties vėžio duomenų grupavimas ir vizualizavimas



Krūties vėžio duomenų grupavimas, klasifikavimas ir vizualizavimas



Grupavimo į tris grupes rezultatas

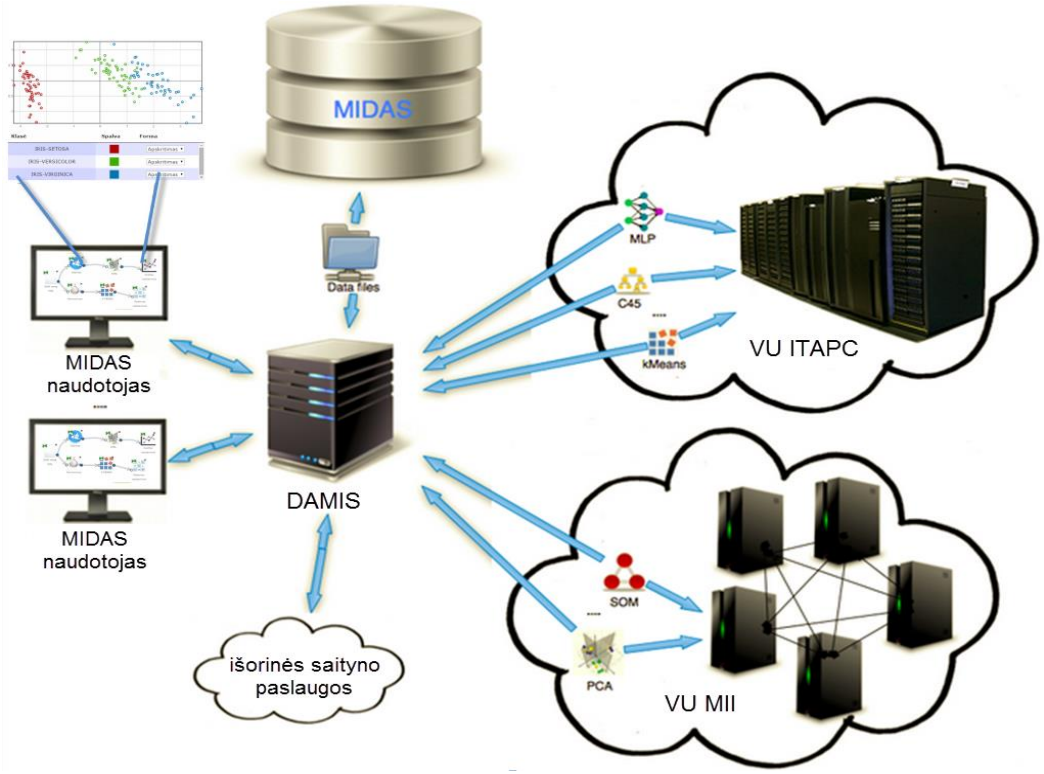


Klasių vizualizavimas

Kodėl verta rinktis DAMIS?

MIDAS
mokslo duomenų archyvas

- **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