Vilniaus universitetas
Matematikos ir informatikos fakultetas
Informatikos katedra

Dirbtiniai neuroniniai tinklai daugiamačių duomenų dimensija mažinti (vizualizuoti)

prof. dr. Olga Kurasova
Olga.Kurasova@mii.vu.lt

2018

Daugiamačiai duomenys. kas tai?

- **Daugiamačiai duomenys** – tai objektą charakterizuojančių **savybių** (rodiklių, parametrų) reikšmių rinkinys.
- Jei savybių reikšmės $x_1, x_2, \dots, x_n$ yra skaičiai, daugiamačius duomenis atitinka **taškai** (**vektoriai**) $X_i = (x_{i1}, x_{i2}, \dots, x_{in}), i = 1, \dots, m,$ čia m – objektų skaičius, n – savybių skaičius.
- Analizuojamų daugiamačių duomenų aibė yra **matrica** (lentelė): $\boldsymbol{X} = \{X_1, X_2, \dots, X_m\} = \{x_{ij},\ i = 1, \dots, m,\ j = 1, \dots, n\},$ kurios i-oji eilutė yra vektorius $X_i \in R^n$, eilutės atitinka analizuojamus objektus, stulpeliai – savybes.

Daugiamačių duomenų pavyzdys (vėžio duomenys)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	C
X_1	5	1	1	1	2	1	3	1	1	b
X_2	5	4	4	5	7	10	3	2	1	b
X_3	3	1	1	1	2	2	3	1	1	b
X_4	6	8	8	1	3	4	3	7	1	b
X_5	4	1	1	3	2	1	3	1	1	b
X_6	1	1	1	1	2	10	3	1	1	b
X_7	2	1	2	1	2	1	3	1	1	b
X_8	2	1	1	1	2	1	1	1	5	b
X_9	4	2	1	1	2	1	2	1	1	b
…	…	…	…	…	…	…	…	…	…	…
X_{460}	8	10	10	8	7	10	9	7	1	m
X_{461}	5	3	3	3	2	3	4	4	1	m
X_{462}	8	7	5	10	7	9	5	5	4	m
X_{463}	7	4	6	4	6	1	4	3	1	m
X_{464}	10	7	7	6	4	10	4	1	2	m
X_{465}	7	3	2	10	5	10	5	4	4	m
X_{466}	10	5	5	3	6	7	7	10	1	m
…	…	…	…	…	…	…	…	…	…	…
X_{699}	4	8	8	5	4	5	10	4	1	m

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

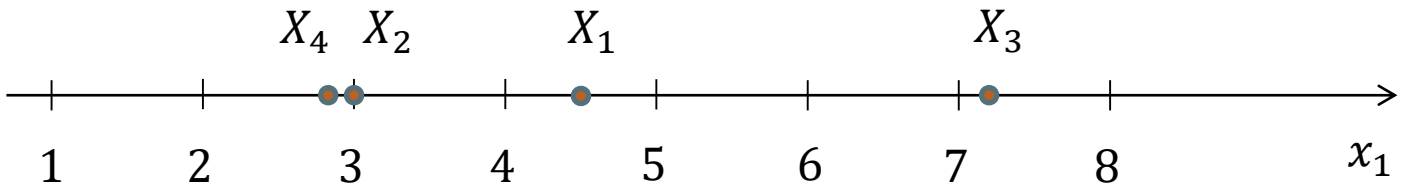
dimensija (matmenų skaičius)
$n = 9$
objektų (duomenų elementų) skaičius $m = 699$

Kokios daugiamačių duomenų ypatybės?

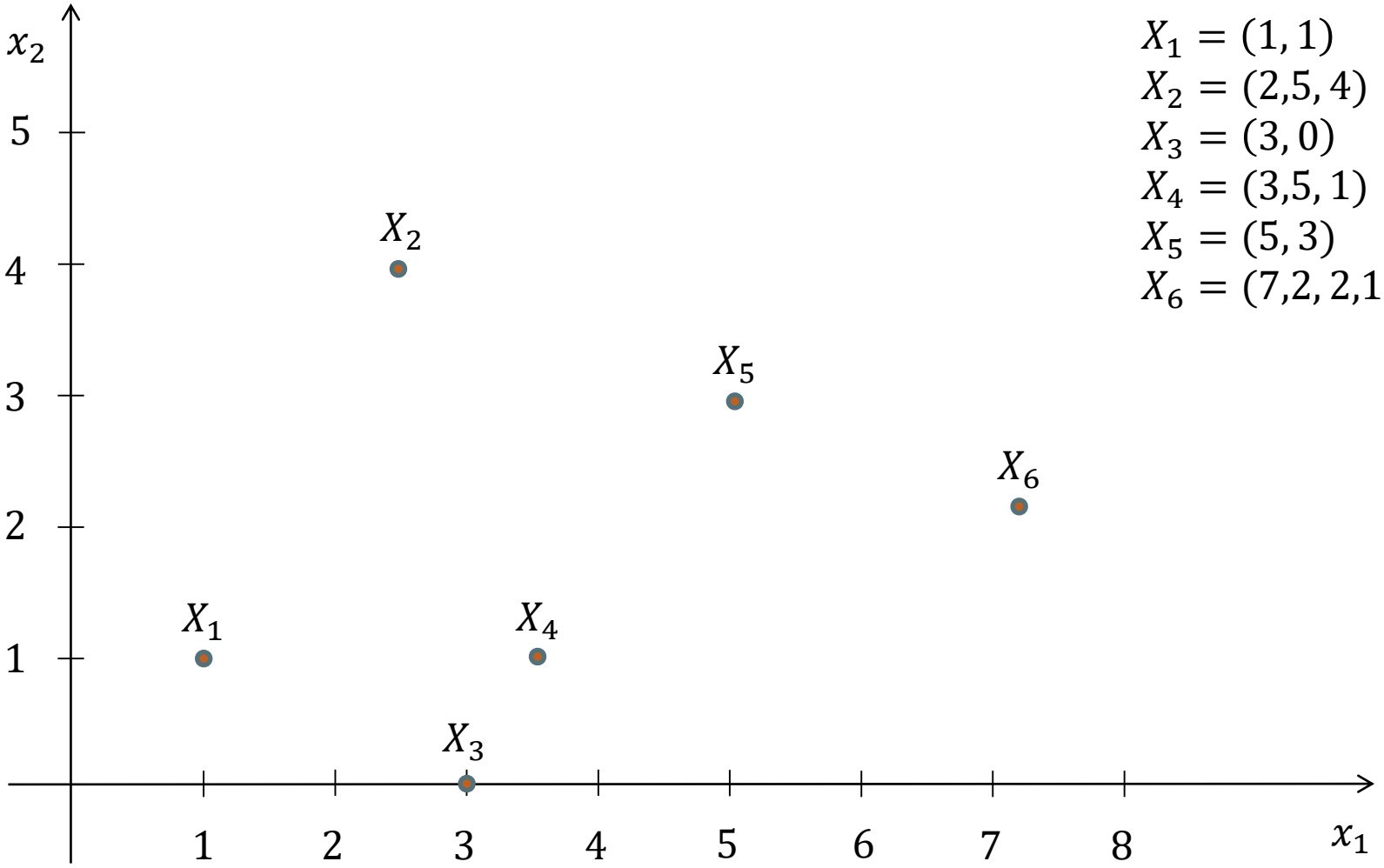
- Turint duomenis, kai $n = 1$, taškus galima atvaizduoti **tiesėje**.
- Turint duomenis, kai $n = 2$ (arba $n = 3$), taškus galima atvaizduoti Dekarto **plokštumoje** (arba erdvėje).
- **Ką daryti**, kai $n > 3$? ☹ ☹ ☹
- Pasitelksime įvairius daugiamačių duomenų **vizualizavimo metodus**.

Kai $n = 1$

$X_1 = (4,5)$
$X_2 = (3)$
$X_3 = (7,2)$
$X_4 = (2,8)$



Kai $n = 2$. taškinis grafikas



$X_1 = (1, 1)$
$X_2 = (2{,}5, 4)$
$X_3 = (3, 0)$
$X_4 = (3{,}5, 1)$
$X_5 = (5, 3)$
$X_6 = (7{,}2, 2{,}1)$

Daugiamačių duomenų vizualizavimo metodai

- **Tiesioginio** vizualizavimo metodai (nėra griežto matematinio kriterijaus):
  - Černovo veidai
  - Žvaigždžių metodas
  - Andrews kreivės
  - Lygiagrečiosios koordinatės
  - Kt.

- **Dimensijų mažinimu** grįsti vizualizavimo (projekcijos) metodai:
  - Pagrindinių komponenčių metodas
  - Daugiamatės skalės
  - **Dirbtiniais neuroniniais tinklais** grįsti metodai
  - Kt.

Irisų gėlių duomenys

- Buvo išmatuoti irisų gėlių žiedų:
  - vainiklapių pločiai (x_1)
  - vainiklapių ilgiai (x_2)
  - taurėlapių ilgiai (x_3)
  - taurėlapių pločiai (x_4)
- Matuotos trijų veislių gėlės (po 50 kiekvienos veislės, viso 150)
- $\boldsymbol{X} = \{X_1, X_2, \dots, X_{150}\} = \{x_{ij},\ i = 1, \dots, 150,\ j = 1, \dots, 4\}$

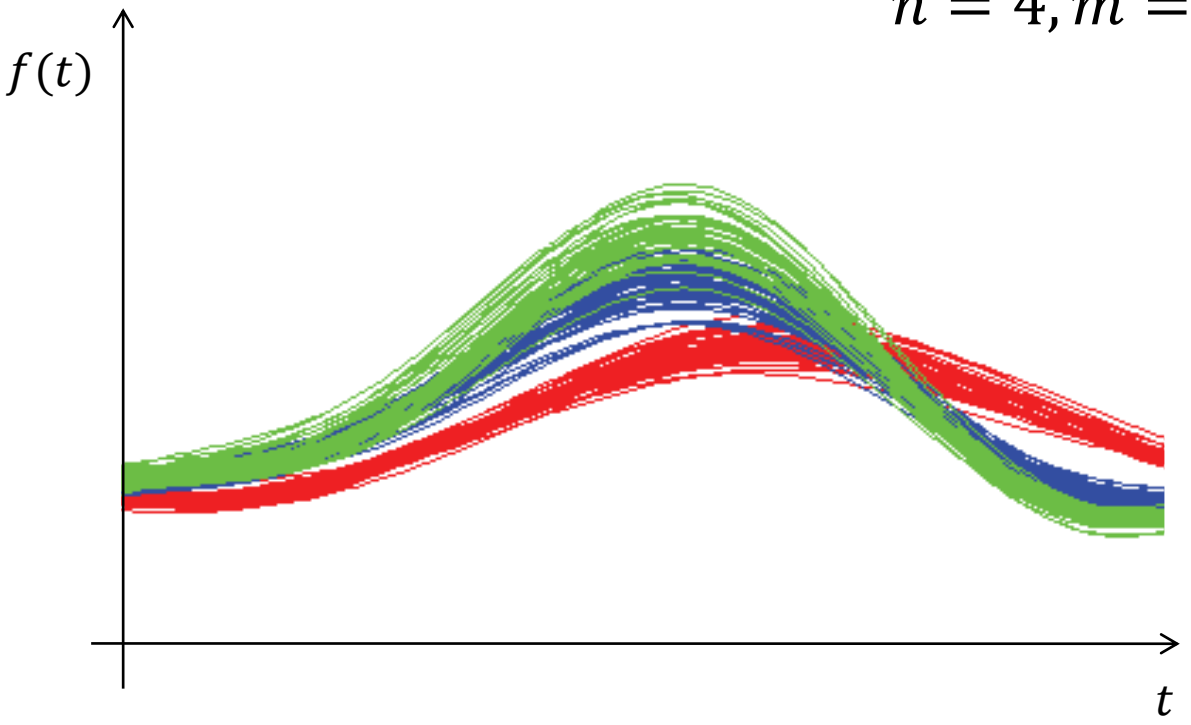
Iris-Setosa

Iris-Versicolor

Iris-Virginica

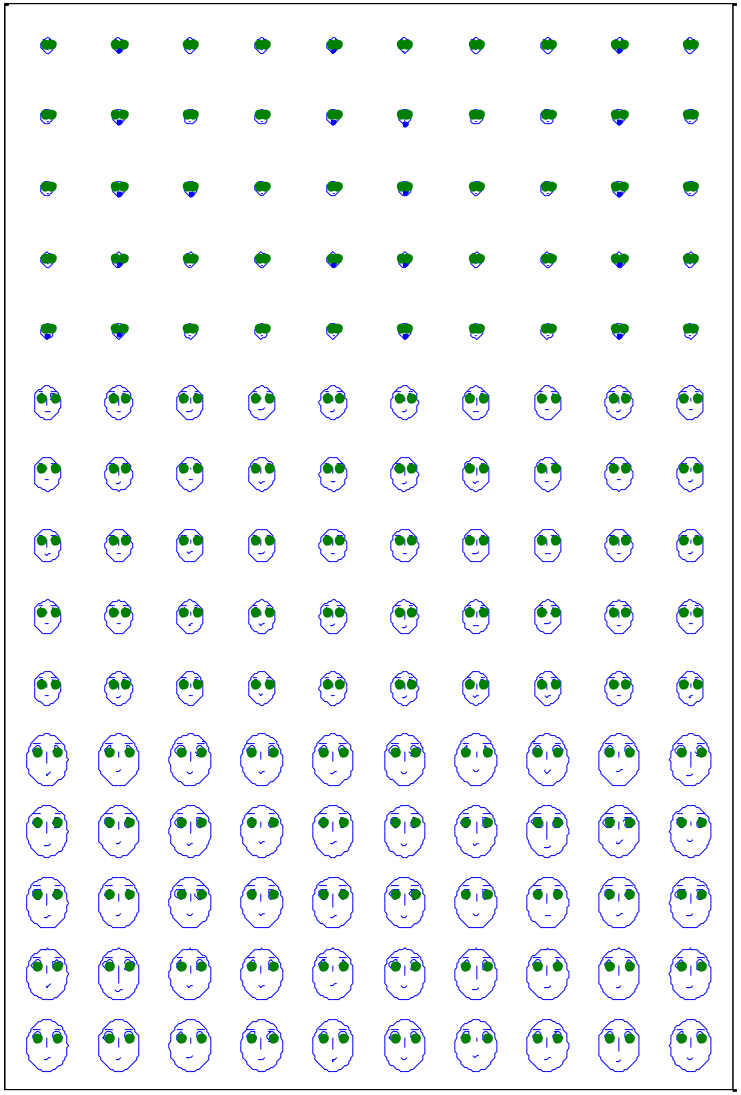
Irisų duomenų vizualizavimas Andrews kreivėmis

$$n = 4, m = 150$$



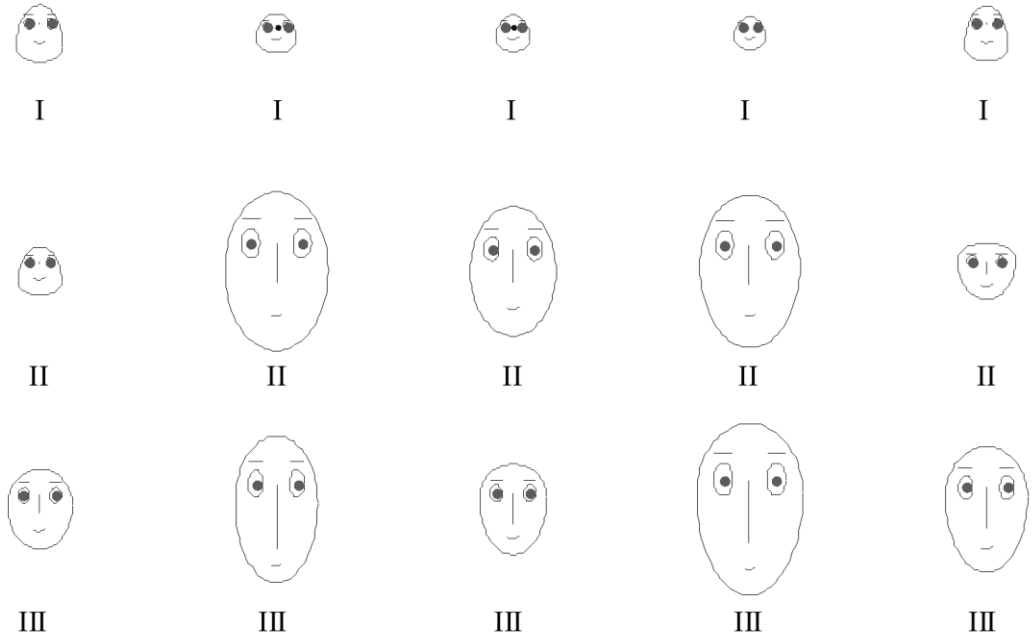
$$f_i(t) = \frac{x_{i1}}{\sqrt{2}} + x_{i2}\sin(t) + x_{i3}\cos(t) + x_{i4}\sin(2t) + x_{i5}\cos(2t) + \cdots \quad -\pi < t < \pi$$

Irisų duomenų vizualizavimas Černovo veidais



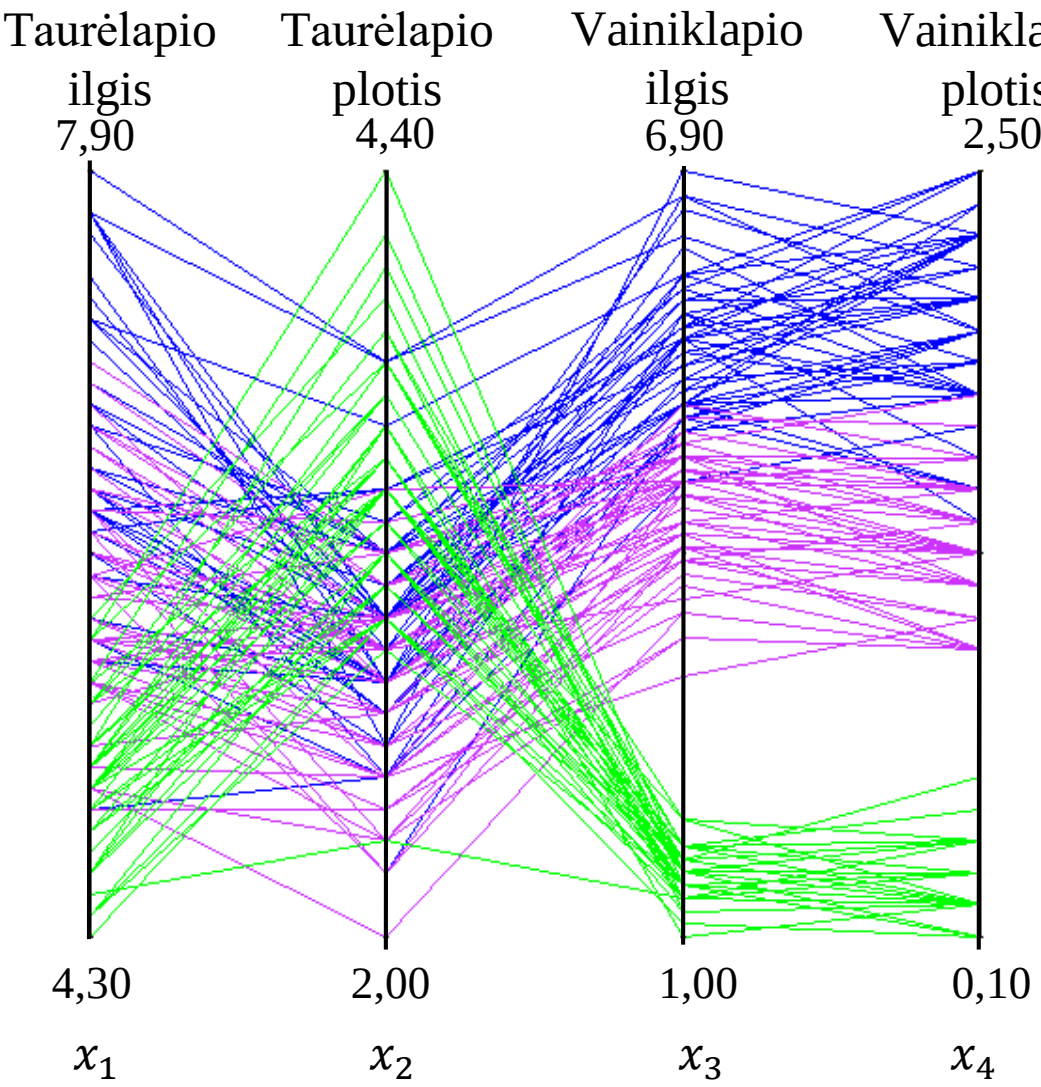
$n = 4, m = 150$

Irisų duomenų vizualizavimas
Černovo veidais



x_1 – veido dydis
x_2 – santykis tarp kaktos/smakro arkų ilgių
x_3 – kaktos forma
x_4 – smakro dydis

Irisų duomenys lygiagrečiose koordinatėse



$n = 4, m = 150$

Dimensijų mažinimu grįstas vizualizavimas

- **Tikslas** – transformuoti daugiamačius duomenis

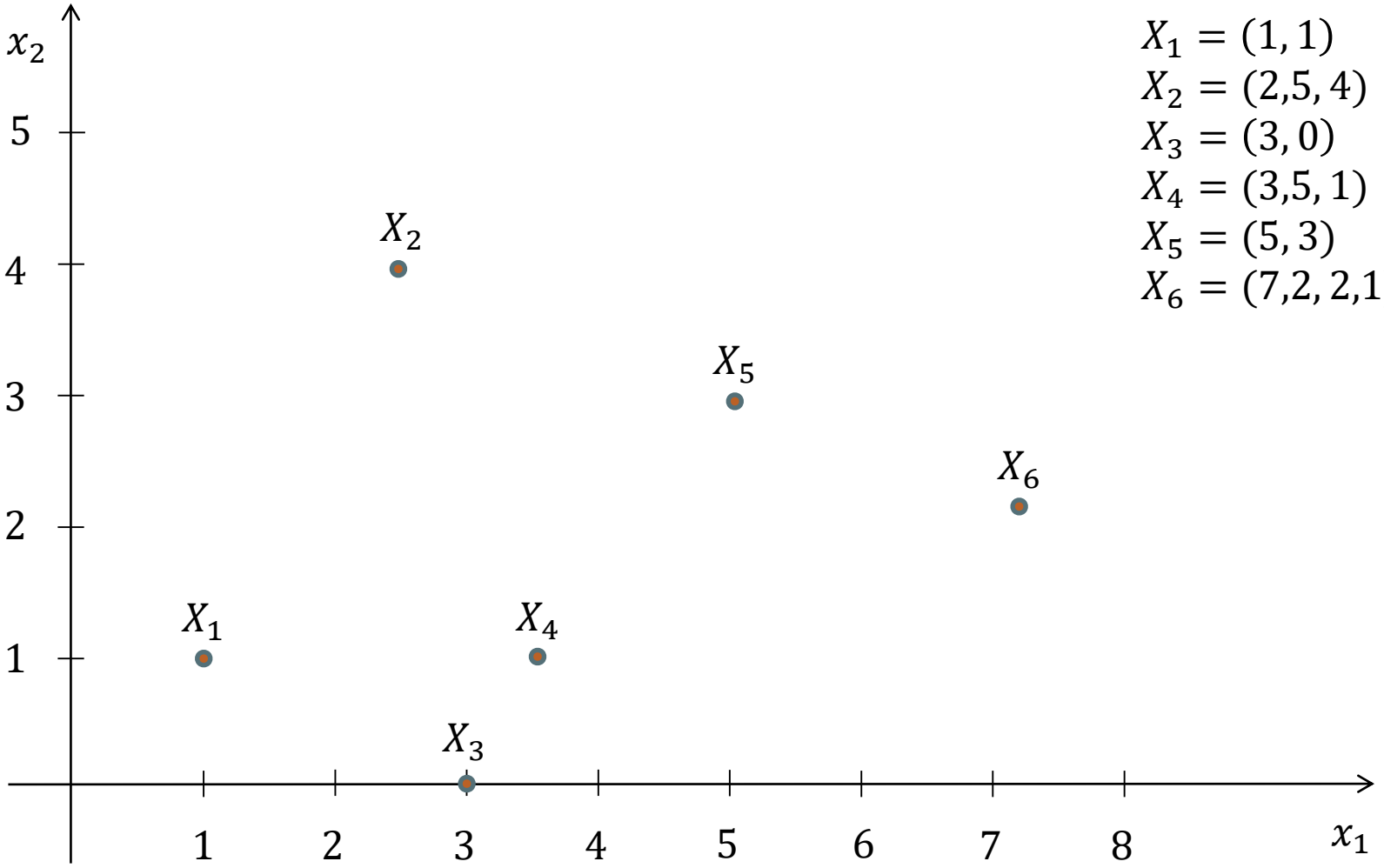
$$X_i = (x_{i1}, x_{i2}, \dots, x_{in}) \in R^n, i = 1, \dots, m,$$

į mažesnės dimensijos erdvės duomenis

$$Y_i = (y_{i1}, y_{i2}, \dots, y_{id}) \in R^d, d < n.$$

- Kai $d = 2$, gautus duomenis (vektorius, taškus) galima **atvaizduoti** įprastoje Dekarto koordinačių sistemoje.

Kai $n = 2$. taškinis grafikas



$X_1 = (1, 1)$
$X_2 = (2{,}5, 4)$
$X_3 = (3, 0)$
$X_4 = (3{,}5, 1)$
$X_5 = (5, 3)$
$X_6 = (7{,}2, 2{,}1)$

Dimensijų mažinimu grįsti vizualizavimo metodai

- Pagrindinių komponenčių analizė (*principal component analysis*)
  - tikslas – **išlaikyti dispersijas**
- Daugiamačių skalių metodas (*multidimensional scaling*)
  - tikslas – **išlaikyti panašumus** (pvz., atstumus)
- Dirbtiniais neuroniniais tinklais grįsti metodai
  - Įprasti tiesioginio sklidimo neuroniniai tinklai
  - Autoasociatyvieji neuroniniai tinklai
  - Saviorganizuojantys neuroniniai tinklai (*self-organizing maps*),
  - Įvairūs junginiai
- Kiti metodai

Vėžio duomenys

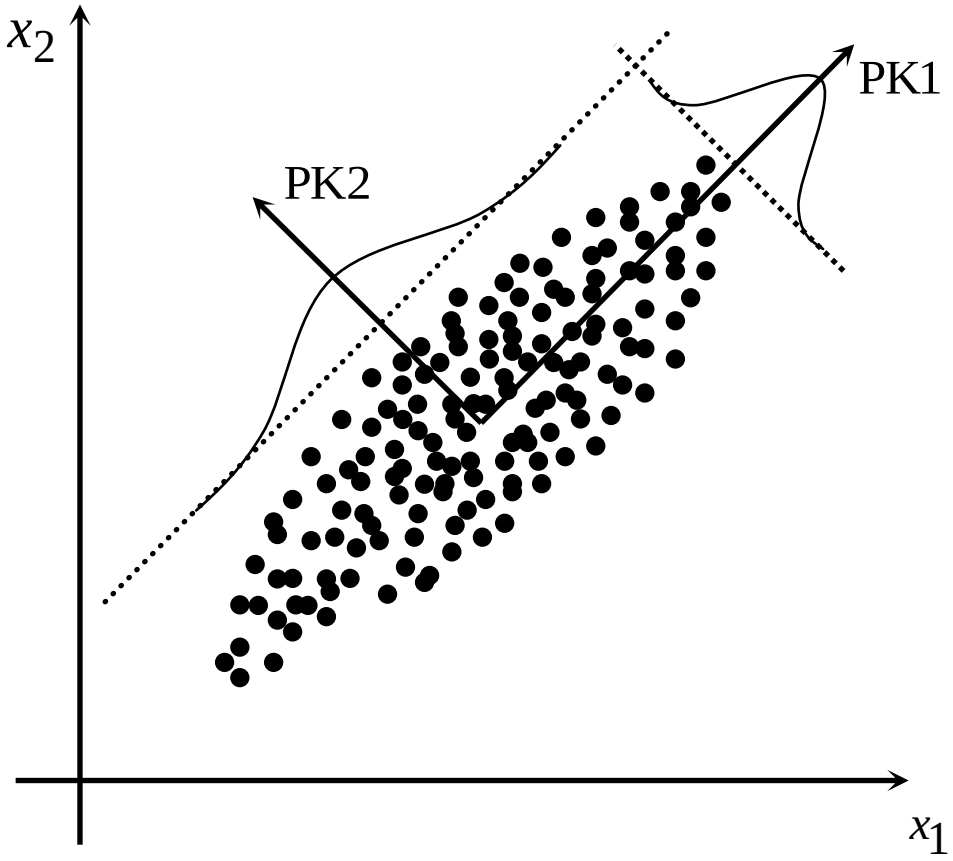
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	C
X_1	5	1	1	1	2	1	3	1	1	b
X_2	5	4	4	5	7	10	3	2	1	b
X_3	3	1	1	1	2	2	3	1	1	b
X_4	6	8	8	1	3	4	3	7	1	b
X_5	4	1	1	3	2	1	3	1	1	b
X_6	1	1	1	1	2	10	3	1	1	b
X_7	2	1	2	1	2	1	3	1	1	b
X_8	2	1	1	1	2	1	1	1	5	b
X_9	4	2	1	1	2	1	2	1	1	b
…	…	…	…	…	…	…	…	…	…	…
X_{460}	8	10	10	8	7	10	9	7	1	m
X_{461}	5	3	3	3	2	3	4	4	1	m
X_{462}	8	7	5	10	7	9	5	5	4	m
X_{463}	7	4	6	4	6	1	4	3	1	m
X_{464}	10	7	7	6	4	10	4	1	2	m
X_{465}	7	3	2	10	5	10	5	4	4	m
X_{466}	10	5	5	3	6	7	7	10	1	m
…	…	…	…	…	…	…	…	…	…	…
X_{699}	4	8	8	5	4	5	10	4	1	m

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

Pagrindinių komponenčių analizė (PCA)

- Esminė PCA idėja yra sumažinti duomenų matmenų skaičių atliekant tiesinę transformaciją ir atsisakant dalies po transformacijos gautų naujų komponenčių, kurių **dispersijos yra mažiausios**.
- Didžiausią dispersiją turinti kryptis vadinama **pirmąja pagrindine komponente** (PK1). Ji eina per duomenų centrinį tašką. Tai taškas, kurio komponentės yra analizuojamą duomenų aibę sudarančių taškų atskirų komponenčių vidurkiai. Visų taškų vidutinis atstumas iki šios tiesės yra minimalus, t. y. ši tiesė yra kiek galima arčiau visų duomenų taškų.
- **Antrosios pagrindinės komponentės** (PK2) ašis taip pat turi eiti per duomenų centrinį tašką ir ji turi būti statmena pirmosios pagrindinės komponentės ašiai.

Pagrindinių komponenčių analizė (PCA)



Daugiamačių skalių (MDS) metodas (1)

- Ieškoma daugiamačių duomenų projekcijų mažesnės dimensijos erdvėje (dažniausiai R^2 arba R^3), siekiant išlaikyti analizuojamos aibės objektų **artimumus** – panašumus arba skirtingumus.
- Gautuose vaizduose panašūs objektai išdėstomi **arčiau** vieni kitų, o skirtingi – **toliau** vieni nuo kitų.
- Dažniausia artimumo matas yra **atstumas** (**Euklido**, miesto kvartalų ir kt.).

Daugiamačių skalių (MDS) metodas (2)

- Tarkime, kiekvieną n-matį vektorių $X_i \in R^n$, $i = 1, \dots, m$, atitinka **mažesnės dimensijos vektorius** $Y_i \in R^d$, $d < n$.
- Atstumą tarp vektorių X_i ir X_j pažymėkime $d(X_i, X_j)$,

o atstumą tarp vektorių Y_i ir Y_j – $d(Y_i, Y_j)$.

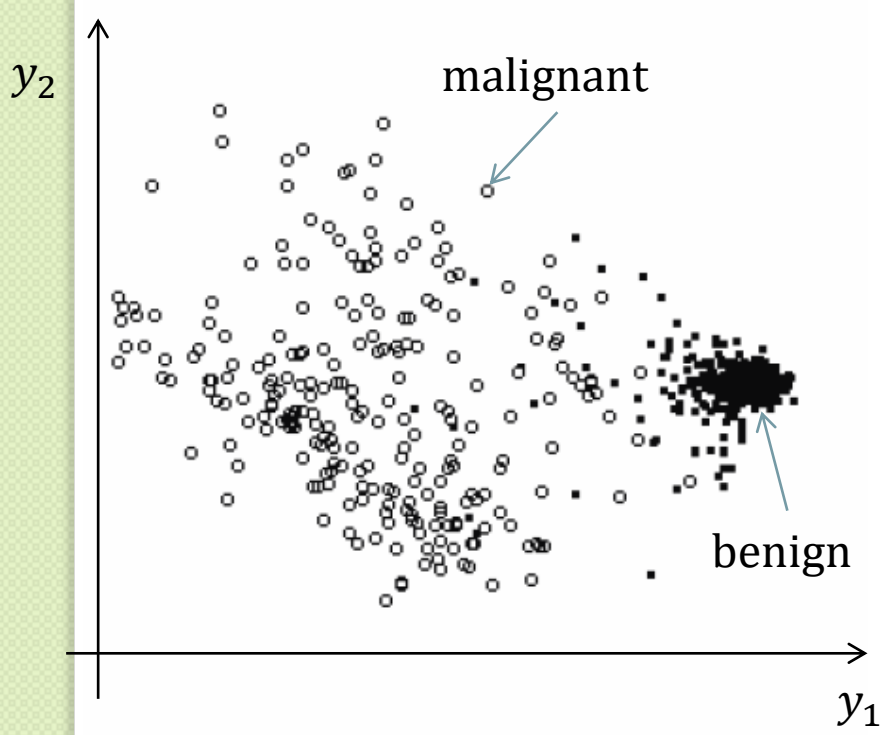
- Naudojantis MDS metodu, bandoma **atstumus** $d(Y_i, Y_j)$ **priartinti prie atstumų** $d(X_i, X_j)$.
- Jei naudojama kvadratinė paklaidos funkcija, tai **minimizuojama tikslo funkcija** užrašoma:

$$E_{MDS} = \sum_{i<j} w_{ij}\left(d(X_i, X_j) - d(Y_i, Y_j)\right)^2$$

Vėžio duomenų vizualizavimas MDS metodu

- Daugiamačių duomenų transformavimas į mažesnės dimensijos erdvę

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

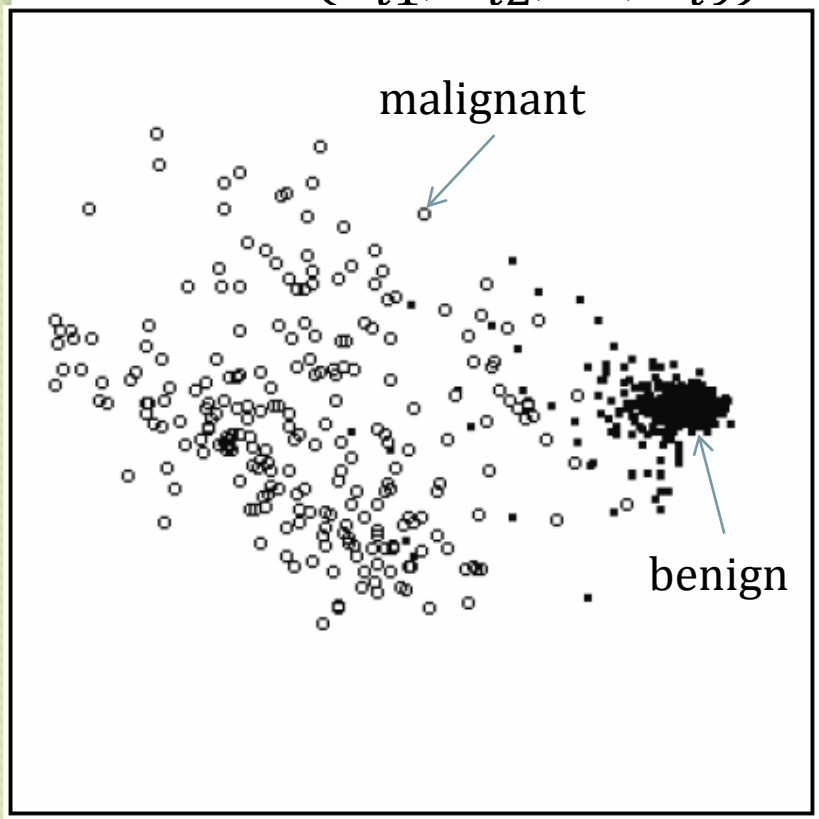


- vienas **taškas** atitinka vieną **pacientę**
- vienas **taškas** apjungia visas **devynias savybes**
- vizualus pateikimas padeda lengviau suvokti **informacijos visumą**

Vėžio duomenų vizualizavimas MDS metodu

- Daugiamačių duomenų transformavimas į mažesnės dimensijos erdvę

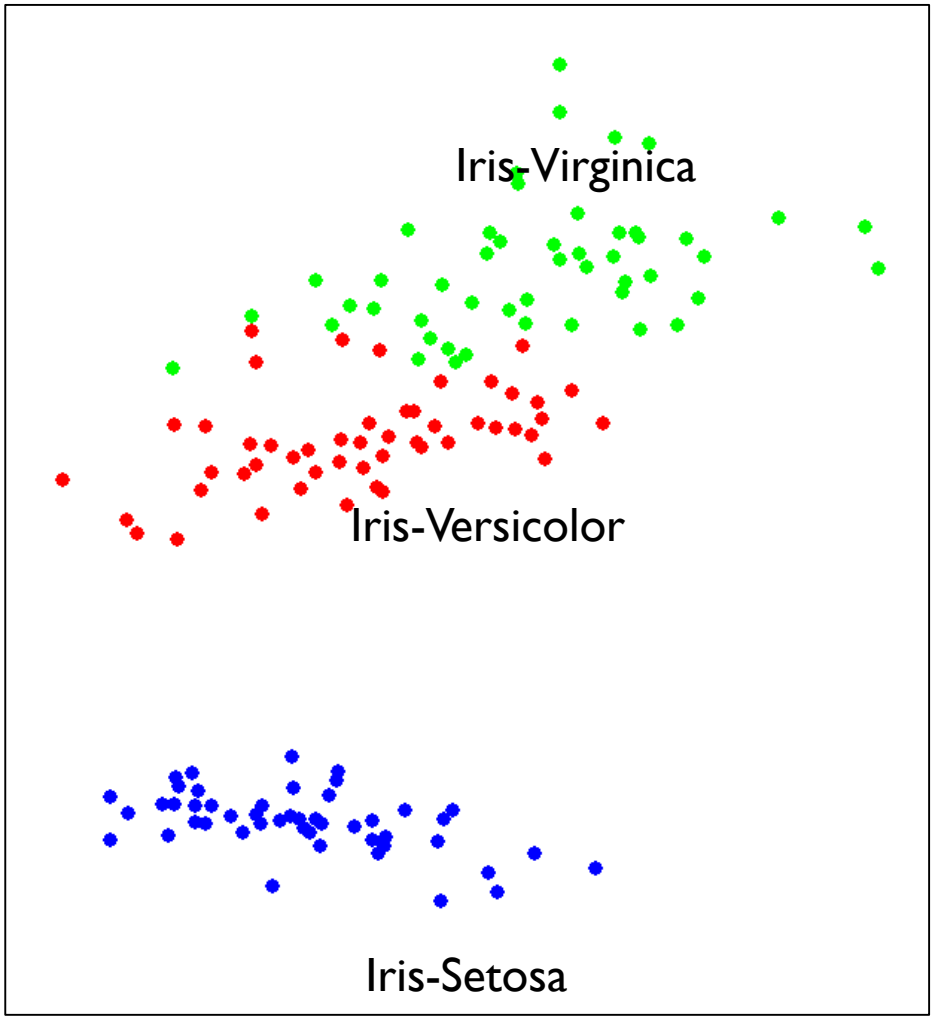
$$(x_{i1}, x_{i2}, \dots, x_{i9}) \rightarrow (y_{i1}, y_{i2})$$



- vienas **taškas** atitinka vieną **pacientę**
- vienas **taškas** apjungia visas **devynias savybes**
- vizualus pateikimas padeda lengviau suvokti **informacijos visumą**

Irisų duomenų vizualizavimas MDS metodu

	x_1	x_2	x_3	x_4	C
X_1	5,1	3,5	1,4	0,2	Set.
X_2	4,9	3,0	1,4	0,2	Set.
X_3	4,7	3,2	1,3	0,2	Set.
X_4	4,6	3,1	1,5	0,2	Set.
X_5	5,0	3,6	1,4	0,2	Set.
…	…	…	…	…	…
X_{51}	7,0	3,2	4,7	1,4	Ver.
X_{52}	6,4	3,2	4,5	1,5	Ver.
X_{53}	6,9	3,1	4,9	1,5	Ver.
X_{54}	5,5	2,3	4,0	1,3	Ver.
X_{55}	6,5	2,8	4,6	1,5	Ver.
…	…	…	…	…	…
X_{101}	5,7	2,8	4,1	1,3	Virg.
X_{102}	6,3	3,3	6,0	2,5	Virg.
X_{103}	5,8	2,7	5,1	1,9	Virg.
X_{104}	7,1	3,0	5,9	2,1	Virg.
X_{105}	6,3	2,9	5,6	1,8	Virg.
…	…	…	…	…	…

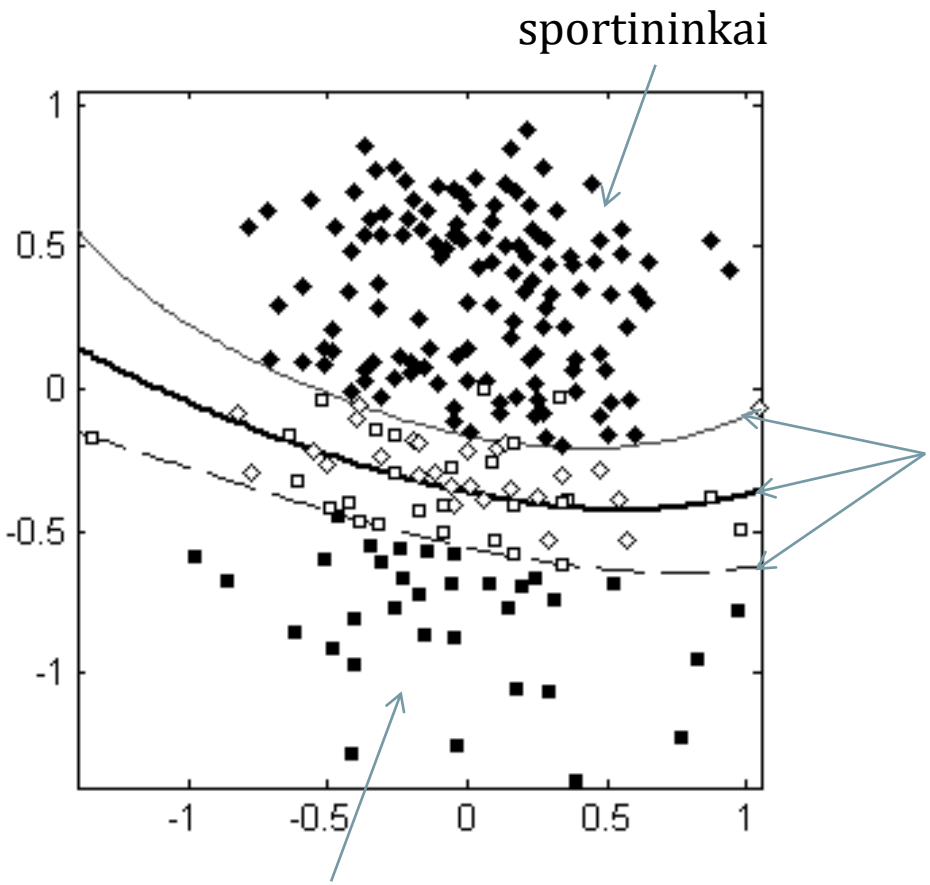


Daugiamačių skalių (MDS) metodas (3)

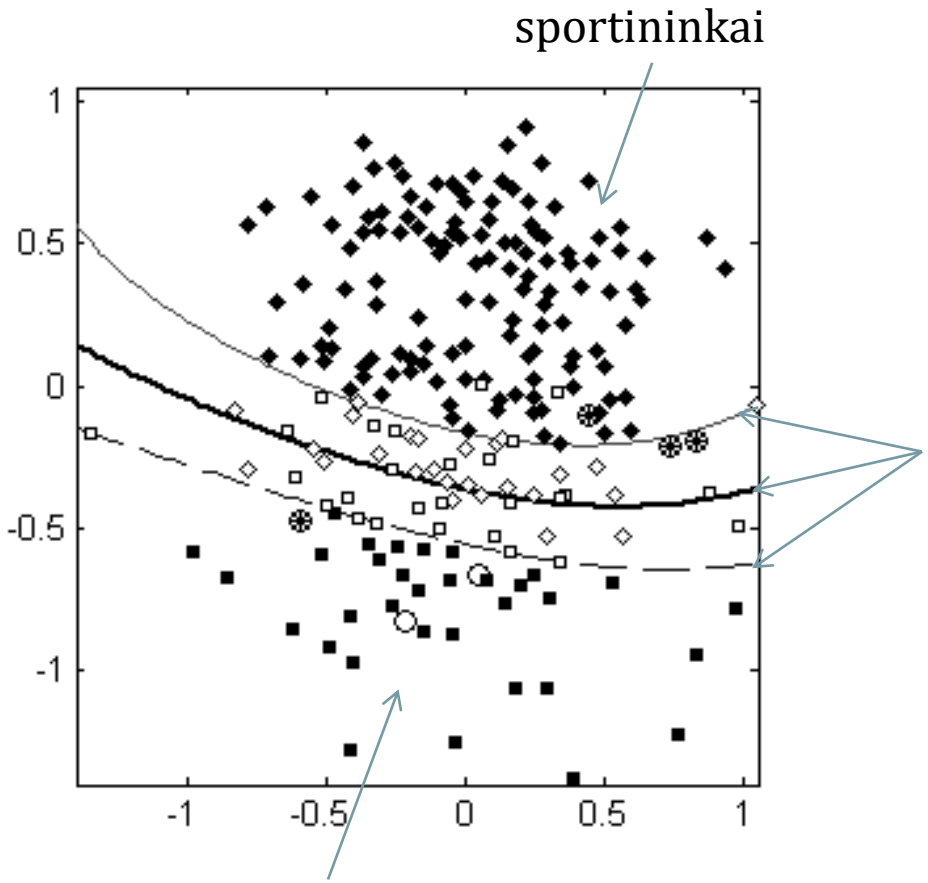
Įvairūs **optimizavimo** metodai:

- Gradientinis nusileidimas
- Genetinis algoritmas
- Šakų-rėžių algoritmas
- Funkcijos mažorizavimu grįstas algoritmas (SMACOF)
- Kt.

Klasių skiriamieji paviršiai

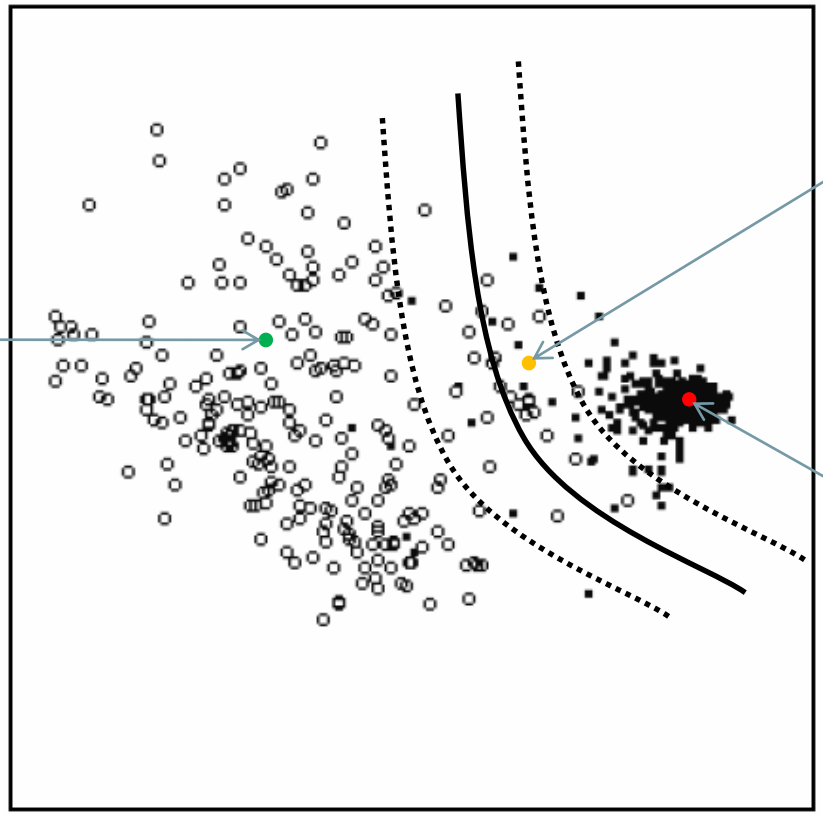


Naujų taškų atidėjimas



Panaudojimo galimybės

- Pirminės grandies gydytojams **ankstyvai diagnostikai nustatyti**

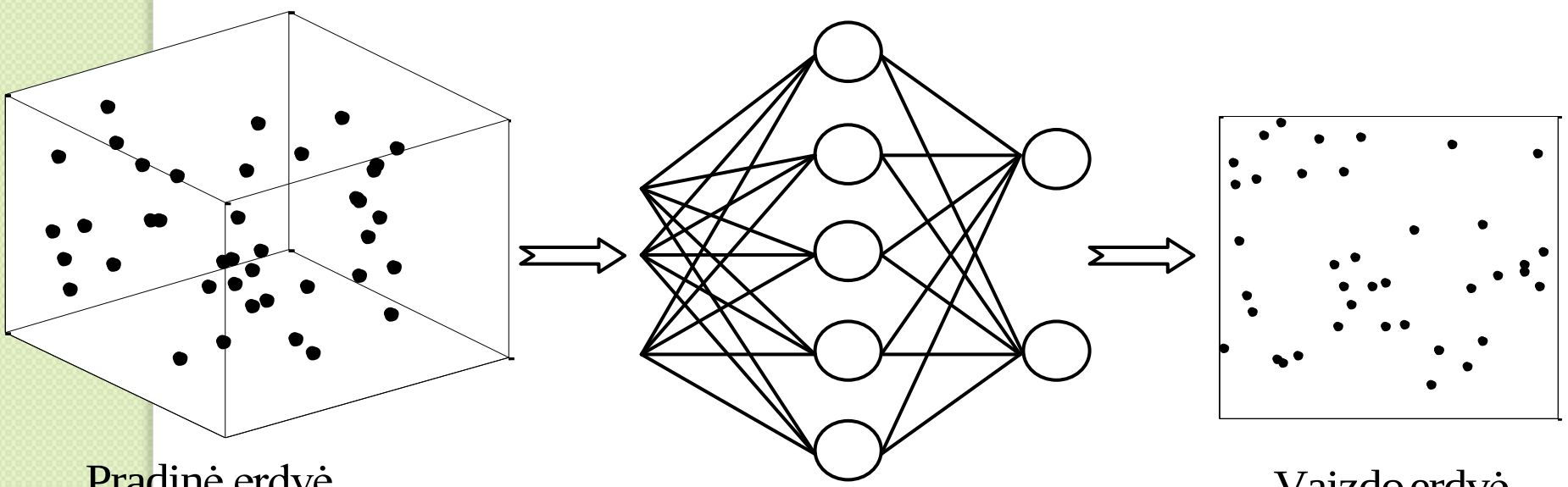


nauja pacientė 2 (būtina atlikti papildomus tyrimus)

nauja pacientė 3 (reikia nedelsti)

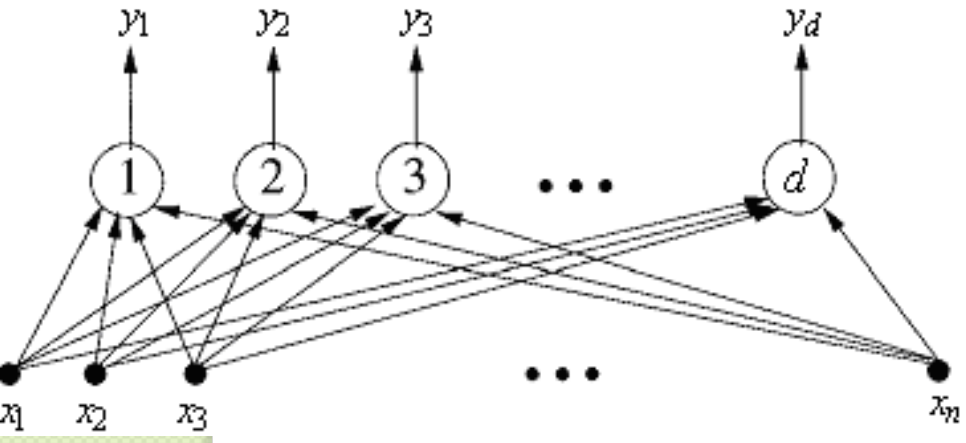
nauja pacientė 1 (viskas gerai)

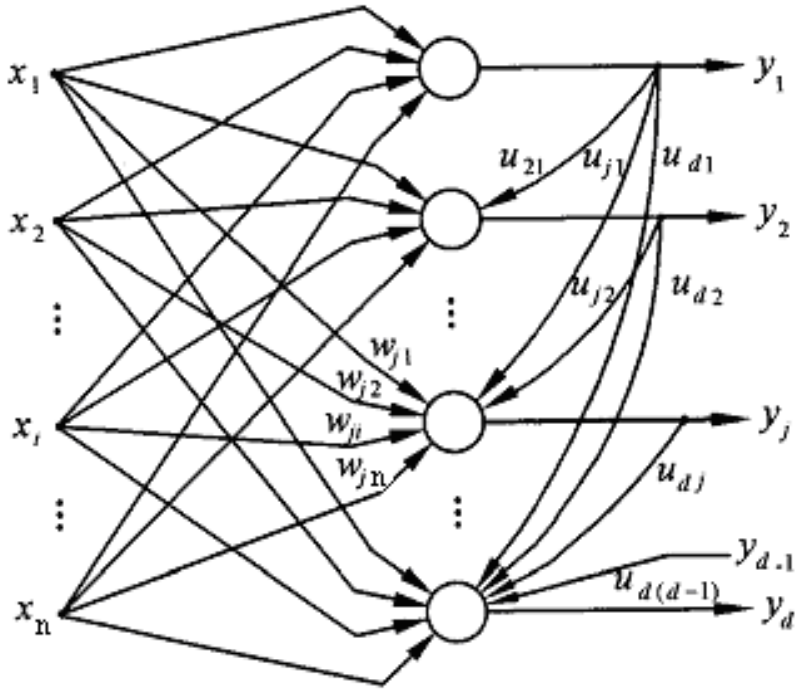
DNT daugiamačiams duomenims vizualizuoti (dimensijai mažinti)



DNT pagrindinėms komponentėms rasti

- **Hebbo** ir **Oja** mokymo taisyklės pagrindinėms komponentėms rasti





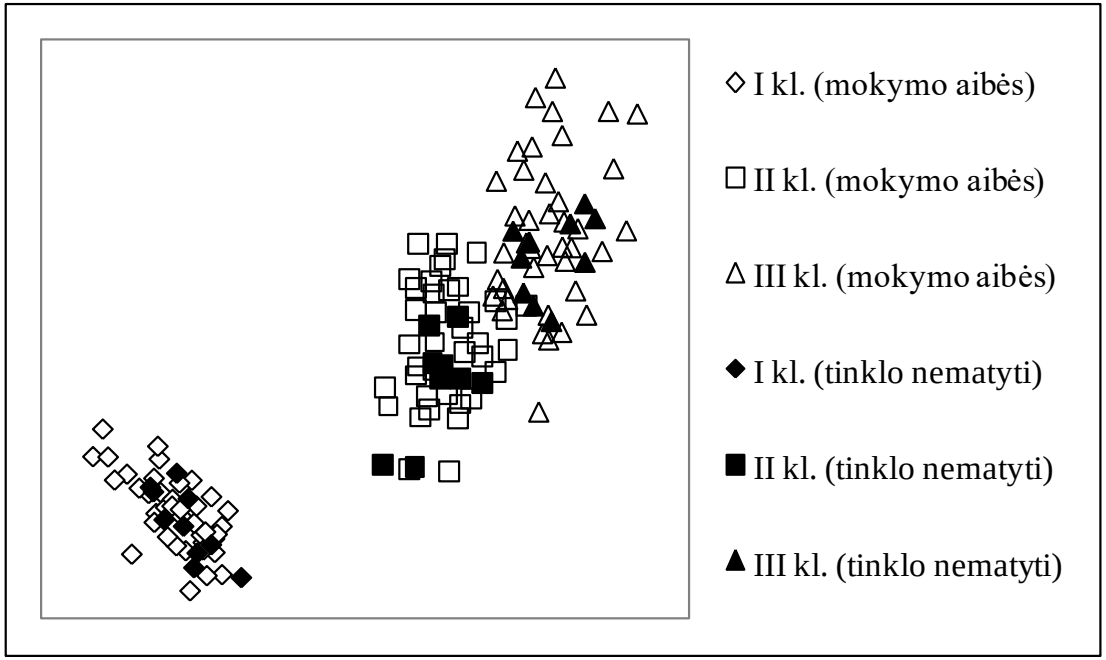
DNT ir daugiamatės skalės (1)

Daugiamačių skalių tipo projekcija yra randama taikant **tiesioginio sklidimo neuroninį tinklą**, mokomą įprastiniu „**klaidos skleidimo atgal**“ algoritmu (mokymas su mokytoju).

- Pradžioje gaunamos taškų projekcijos **daugiamačių skalių metodu**.
- Tinklas yra **apmokomas** duomenimis, sudarytais iš daugiamačių taškų koordinačių, kai norimos išėjimų reikšmės yra **taškų projekcijos**, gautos daugiamatėmis skalėmis.

DNT ir daugiamatės skalės (1)

- Apmokius tinklą, į jį **pateikiami daugiamačiai taškai**, kurių projekcijos dar nėra žinomos.
- Tinklo išėjimuose gaunamos **taškų projekcijos** mažesnės dimensijos erdvėje.



Autoasociatyvieji neuroniniai tinklai

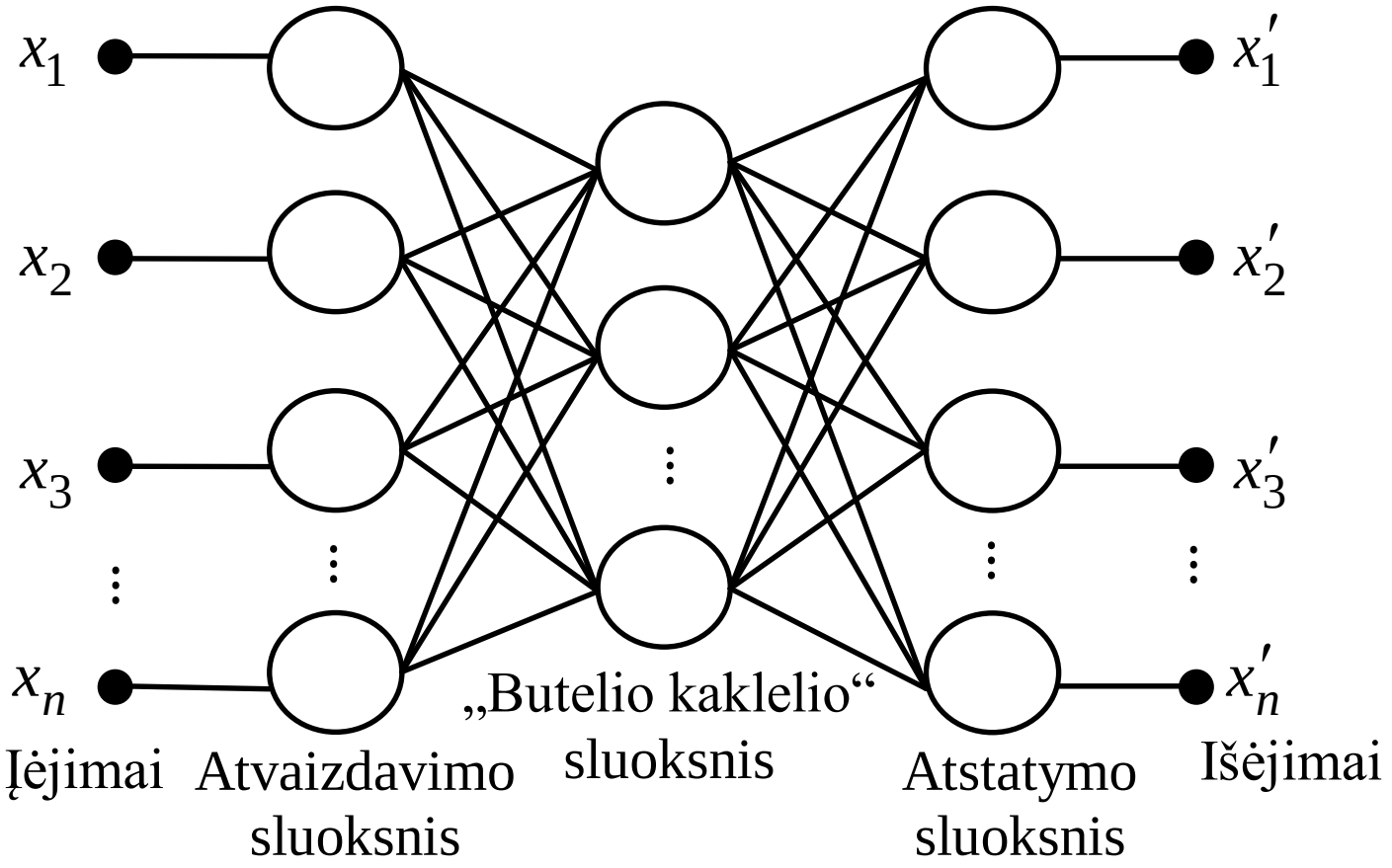
- **Autoasociatyvieji neuroniniai tinklai** (*autoassociative neural networks*) dar dažnai vadinami autokoderių tinklais.
- Jie naudojami matmenų skaičiui mažinti išskiriant *d* neuronų iš vadinamojo „**butelio kaklelio**“ (*bottleneck*) sluoksnio, sudaryto iš mažiau elementų nei įėjimo ir išėjimo sluoksniai, čia *d* yra vaizdo erdvės matmenų skaičius.

Autoasociatyvieji neuroniniai tinklai

Autoasociatyvusis neuroninis tinklas sudarytas iš **dviejų dalių**:

- pirma dalis transformuoja pradinius analizuojamus daugiamačius duomenis į mažesnio skaičiaus matmenų erdvę (**atvaizdavimo sluoksnis**),
- o antroji – rekonstruoja (atstato) pradinius duomenis iš gautų projekcijų (**atstatymo sluoksnis**).

Autoasociatyvieji neuroniniai tinklai



Ribota Boltzmano mašina

- **Riboto Boltzmano mašinos** (*Restricted Boltzman Machine*) neuroninis tinklas duomenų dimensijai mažinti veikimas panašus į autoasociatyvius neuroninius tinklus.