Vilniaus universitetas
Matematikos ir informatikos fakultetas
Informatikos katedra

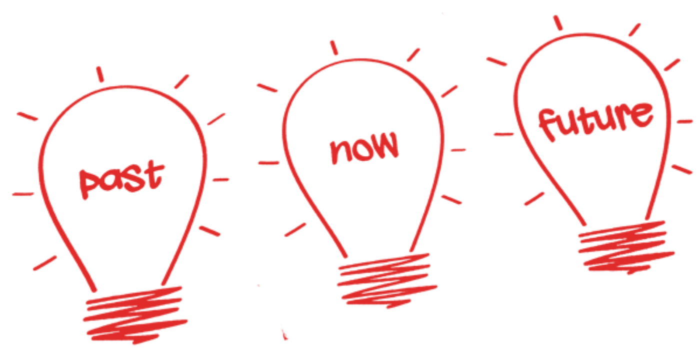
Dirbtiniai neuroniniai tinklai duomenims prognozuoti

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

2018

Neuroniniai tinklai prognozavimui

- Dar vienas duomenų analizės uždavinys – **duomenų prognozavimas** (*prediction*, *forecasting*).
- **Prognozavimo uždaviniai** ypač aktualūs ekonomikoje, finansuose, meteorologijoje ir kt.
- Turint vadinamuosius **istorinius duomenis**, reikia kiek galima tiksliau **numatyti** tam tikro požymio reikšmes ateityje.



Prognozavimo metodai

- Prognozavimo uždaviniai gali būti sprendžiami taikant įvairius **statistinius metodus**, pvz., regresija, slenkančio vidurkio, ARMA, ARIMA, SARIMA ir kt.
- **Tiesioginio sklidimo neuroniniai tinklai** taip pat yra sėkmingai taikomi prognozavimo uždaviniams spręsti.

Regresija – paprasčiausias prognozavimo metodas

- **Regresinės analizės** paskirtis – numatyti priklausomojo kintamojo reikšmę mažiausiai vieno nepriklausomojo kintamojo atžvilgiu ir paaiškinti, kaip nepriklausomo kintamojo pokyčiai veikia priklausomą kintamąjį.
- Turint tokią priklausomybę, galime **prognozuoti** priklausomo kintamojo reikšmes.
- Atliekant regresinę analizę priklausomumas tarp dviejų kintamųjų yra išreiškiamas matematine lygtimi, kuri vadinama **regresijos lygtimi**.
- Išskiriamos **tiesinė** ir **netiesinė** regresijos. Pirmuoju atveju priklausomumas išreiškiamas **tiesės lygtimi**, o antruoju atveju kitokia lygtimi, pvz., polinomu, eksponente ir kt.

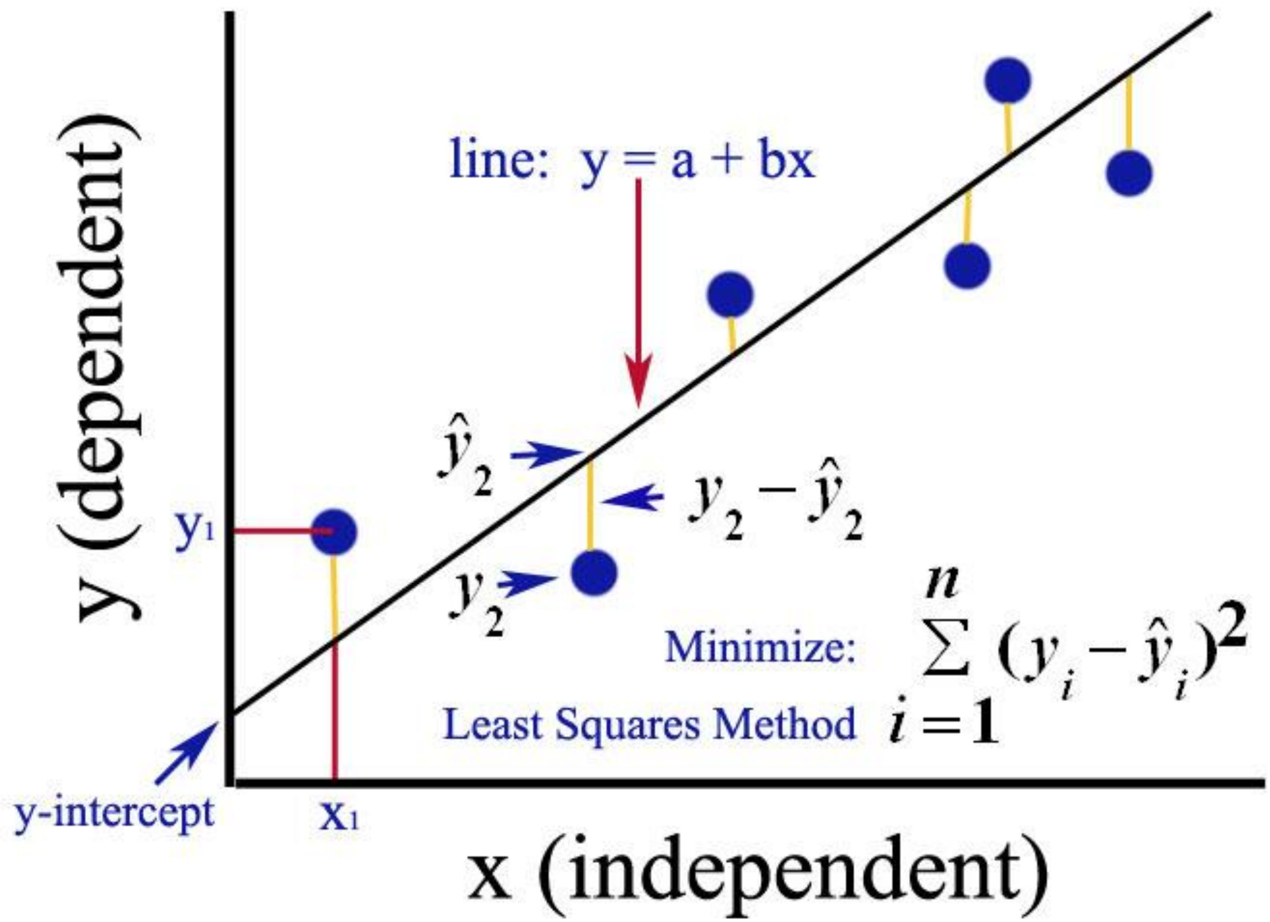
Tiesinė regresijos lygtis

- Jei yra viena priklausomas y ir vienas nepriklausomas kintamasis (požymis) x, tuomet **tiesinės regresijos lygtis** užrašoma taip:

$$y = a + bx + e$$

- Čia e yra atsitiktinė paklaida, atsirandanti dėl matavimo ar kitų duomenų gavimo paklaidų.
- Kai yra žinomi koeficientai a ir b, galima **prognozuoti**, kaip keisis priklausomojo požymio y reikšmės, keičiantis nepriklausomajam požymiui x.
- Naudodamiesi lygtimi galime paskaičiuoti, kaip keisis y reikšmės, esant tokioms x reikšmėms, kurių mes netyrėme, t. y., **galėsime prognozuoti** y reikšmes.

Tiesinė regresija grafiškai



Tiesinės regresijos lygties koeficientų radimas

- Tarkime, turime m stebėjimų metu gautas **duomenų poras** $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$.
- **Tikslas** – rasti koeficientų a ir b įverčius $\hat{a}$ ir $\hat{b}$ tokius, kad funkcijos $\hat{y}(x) = \hat{a} + \hat{b}x$ reikšmės taškuose x_j kiek galima mažiau skirtųsi nuo y_j reikšmių.
- Gautoji funkcija bus naudojama priklausomojo kintamojo nežinomoms reikšmėms **prognozuoti**.
- Kiekvieną x_j atitinka y_j ir funkcijos $\hat{y}(x)$ reikšmės taškuose x_j.
- Geriausia tinkanti funkcija yra tokia, kurios **skirtumai** $\hat{e} = y_j - \hat{y}(x_j)$ būtų mažiausi, $j = 1, \ldots, m$.

Tiesinės regresijos lygties koeficientų radimas

- Įverčiai $\hat{a}$ ir $\hat{b}$ randami vadinamuoju **mažiausių kvadratų metodu**, t. y., minimizuojama kvadratinių sumų paklaidos (KSP) funkcija:

$$\mathrm{KSP} = \sum_{j=1}^{m} \left(y_j - \hat{y}(x_j) \right)^2 = \sum_{j=1}^{m} \left(y_j - \hat{a} - \hat{b} x_j \right)^2$$

- Šią funkciją reikia minimizuoti pagal du parametrus $\hat{a}$ ir $\hat{b}$, t. y., **skaičiuoti dalines išvestines** ir jas prilyginti nuliui.

Tiesinės regresijos lygties koeficientai

- Išsprendus **gautą lygčių sistemą** gaunami tokie sprendiniai:

$$\hat{a} = \frac{1}{m}\sum_{j=1}^{m} y_j - \hat{b}\frac{1}{m}\sum_{j=1}^{m} x_j$$

$$\hat{b} = \frac{\sum_{j=1}^{m} x_j y_j - \frac{1}{m}\left(\sum_{j=1}^{m} x_j \sum_{j=1}^{m} y_j\right)}{\sum_{j=1}^{m} x_j^2 - \frac{1}{m}\left(\sum_{j=1}^{m} x_j\right)^2}$$

Tiesinės regresijos įvertinimas

- Reikia nepamiršti, kad parametrai $\hat{a}$ ir $\hat{b}$ yra tik parametrų a ir b įverčiai, kurie bendru atveju gali ir nesutapti, t. y., gautis **liekamoji paklaida**, parodanti, kiek stebėtoji y_j reikšmė skiriasi nuo reikšmės, kurią gautume prognozuodami pagal regresijos tiesę.

- **Liekamųjų paklaidų kvadratų suma** (SSE), skaičiuojama pagal formulę:

$$\mathrm{SSE} = \sum_{j=1}^{m}\left(y_j - \hat{y}(x_j)\right)^2 = \sum_{j=1}^{m}\left(y_j - (\hat{a} + \hat{b}x_j)\right)^2$$

Determinacijos koeficientas

- Dar dažnai vertinamas **determinacijos koeficientas** R^2, įgyjantis reikšmes nuo 0 iki 1, t. y. $0 < R^2 \leq 1$. Idealiu atveju, $R^2 = 1$.

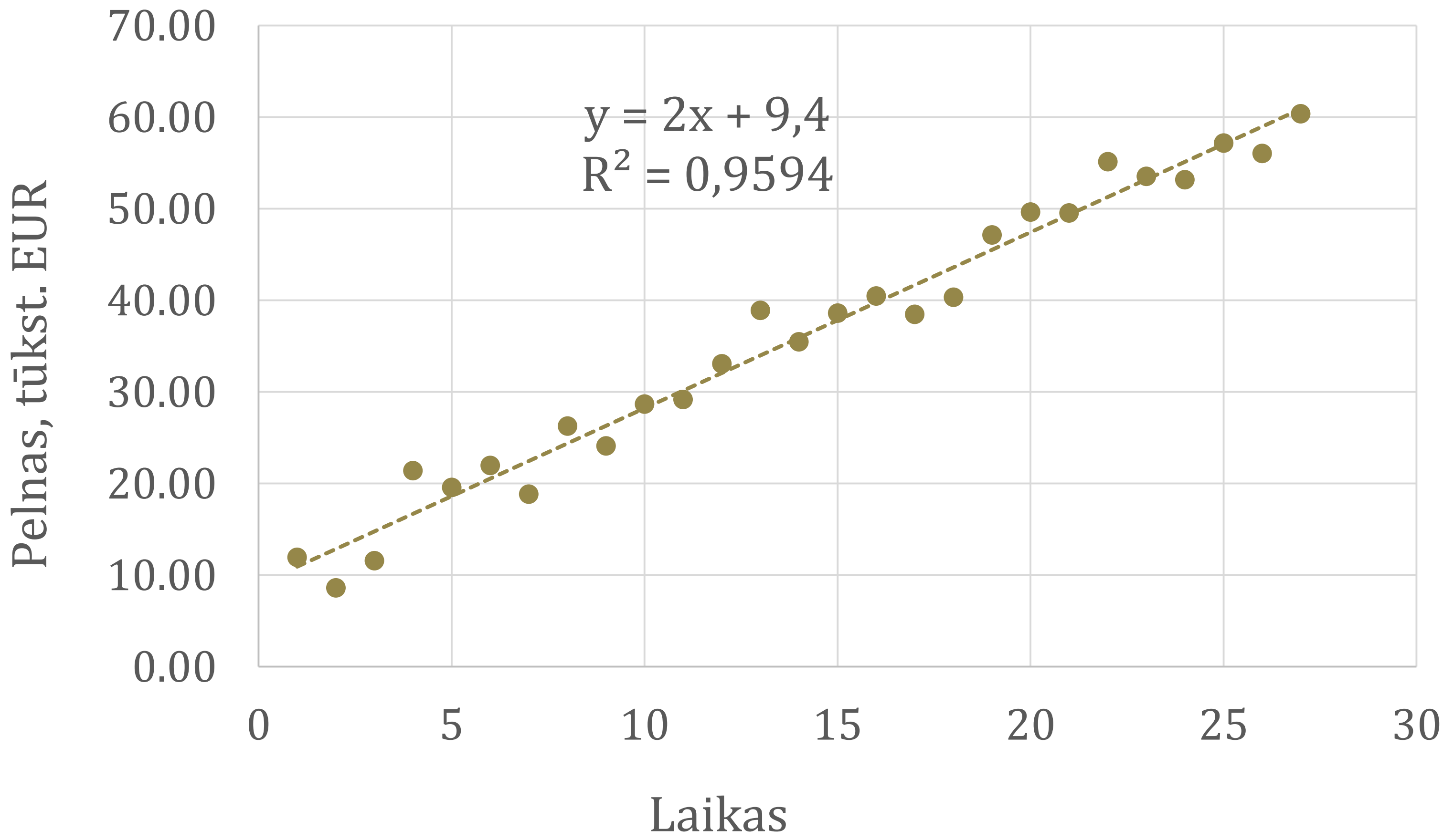
$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}}$$

- Čia SSR – regresijos kvadratų suma, SST – visa kvadratų suma

$$\mathrm{SST} = \sum_{j=1}^{m} \left(y_j - \frac{1}{m} \sum_{k=1}^{m} y_k \right)^2,$$

$$\mathrm{SSR} = \sum_{j=1}^{m} \left(\hat{y}(x_j) - \frac{1}{m} \sum_{k=1}^{m} y_k \right)^2.$$

Tiesinė regresija



Tiesinė regresija kelių kintamųjų atveju

- Jei ieškome sąryšio tarp vieno priklausomo kintamojo y ir kelių kitų nepriklausomų $x_1, x_2, \dots, x_n$, **turime praplėsti** prieš tai aptartą modelį.

- Tuomet **tiesinės regresijos** modelis yra aprašomas tokia lygtimi:

$$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n + e,$$

- čia $a, b_1, b_2, \dots, b_n$ yra **regresijos koeficientai**, e – atsitiktinė paklaida.

Kito tipo regresijos

Tiriant vieno nepriklausomo kintamojo x sąryšį nuo priklausomo kintamojo y, galimi šie **regresijos tipai**:

- **eksponentinė**:

$$y = ae^{bx} + e,$$

- **logaritminė**:

$$y = a\ln(x) + b + e,$$

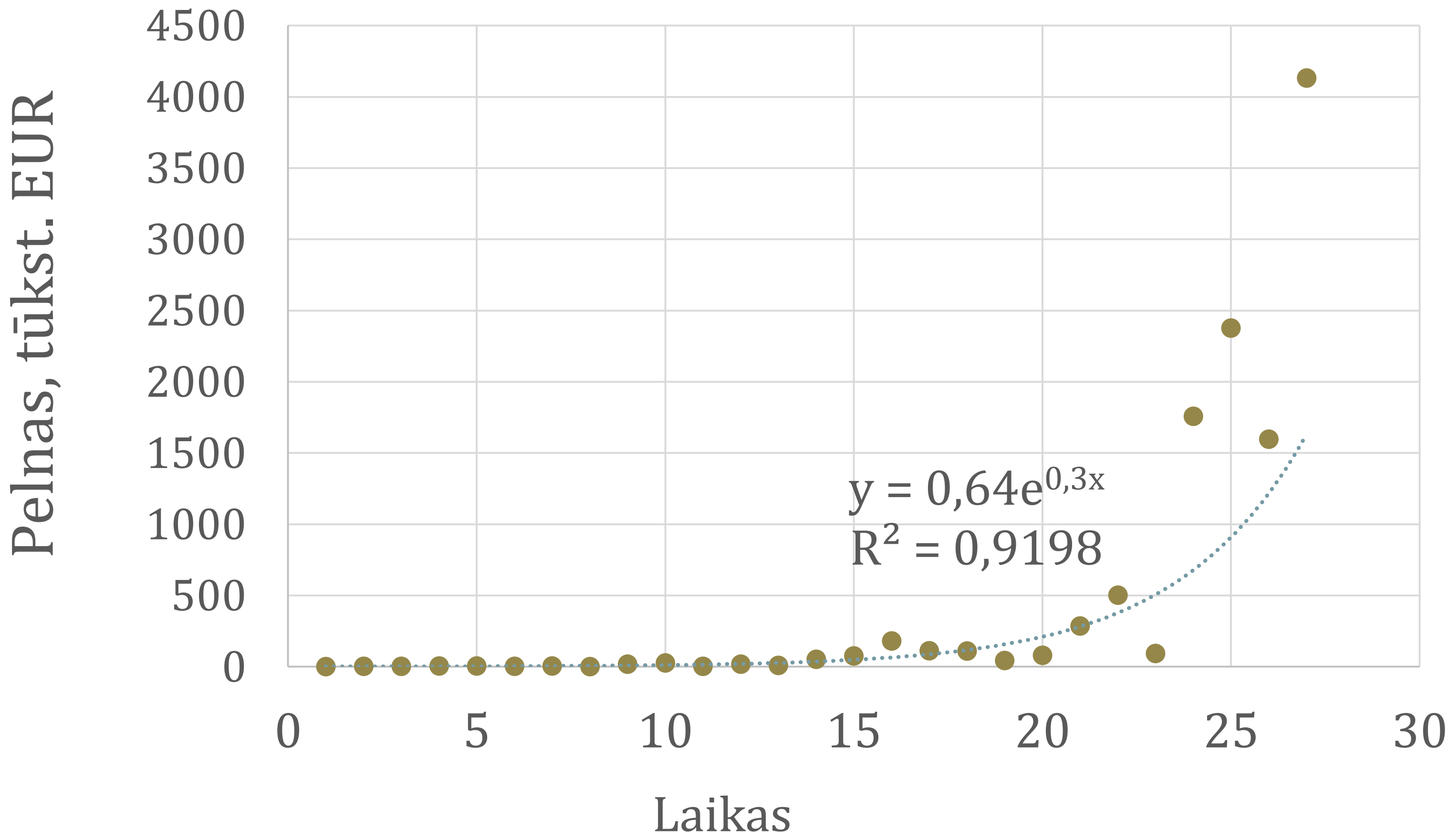
- **laipsninė** (**rodiklinė**):

$$y = ax^{b} + e,$$

- **polinominė** (n-tojo laipsnio):

$$y = a + b_1 x + b_2 x^2 + \cdots + b_n x^n + e.$$

Eksponentinė regresija



Laiko eilutės

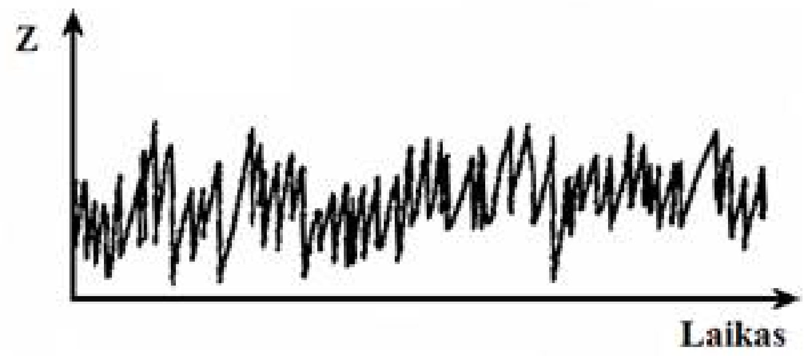
- Įprastai prognozavimo uždaviniai sprendžiami nagrinėjant vadinamąsias **laiko eilutes**.
- Tarkime tam tikro atsitiktinio dydžio X reikšmės stebimos laikui bėgant. Tokio atsitiktinio dydžio reikšmių seka $(X_1, X_2, \dots, X_t)$ vadinama **laiko eilute** (*time series*).
- Įprastai laikoma, kad yra žinomos reikšmės $X(t_i)$ laiko momentais $t_1 < t_2 < \cdots < t_n$, o visi stebėjimai atliekami **vienodais laiko intervalais**, $t_{i+1} - t_i = \Delta t$.
- **Prognozavimo tikslas** – žinant reikšmes $X(t_1), X(t_2), \dots, X(t_n)$, nustatyti reikšmę $X(t_{n+1})$.

Laiko eilučių dedamosios

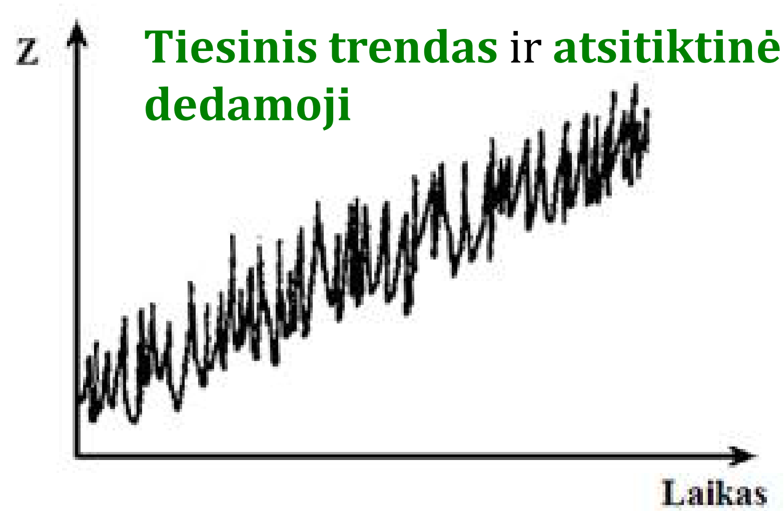
- Dažnai laiko eilutėse stebimos dvi dedamosios: **atsitiktinė** ir **apibrėžtoji**.
- **Apibrėžtosios** dedamosios dalys:
  - **trendas** (atspindi pagrindines bei ilgalaikes laiko eilutės tendencijas, esminius tiriamo proceso bruožus; trendas gali būti tiesinis, eksponentinis ir kt.),
  - **sezoniniai svyravimai** (reguliarus stebimo kintamojo reikšmių didėjimas bei mažėjimas griežtai apibrėžtais laiko periodais),
  - **cikliniai svyravimai** (yra panašūs į sezoninius, tačiau neturi tokio griežto matematinio aprašymo, jų pasikartojimo periodas nėra toks apibrėžtas).

Laiko eilučių dedamosios

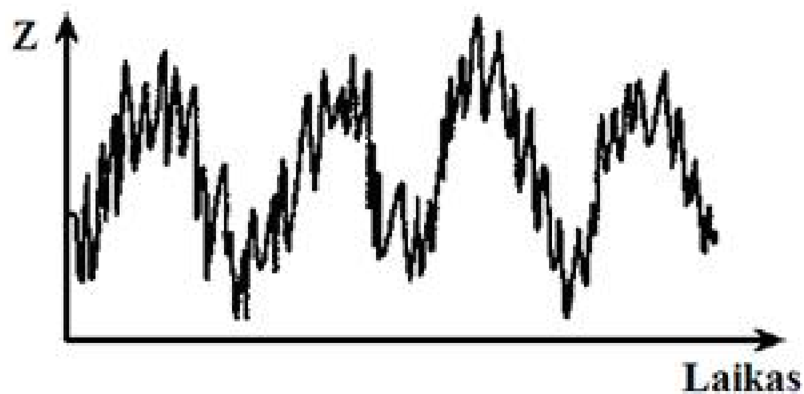
Vien tik **atsitiktinė dedamoji**



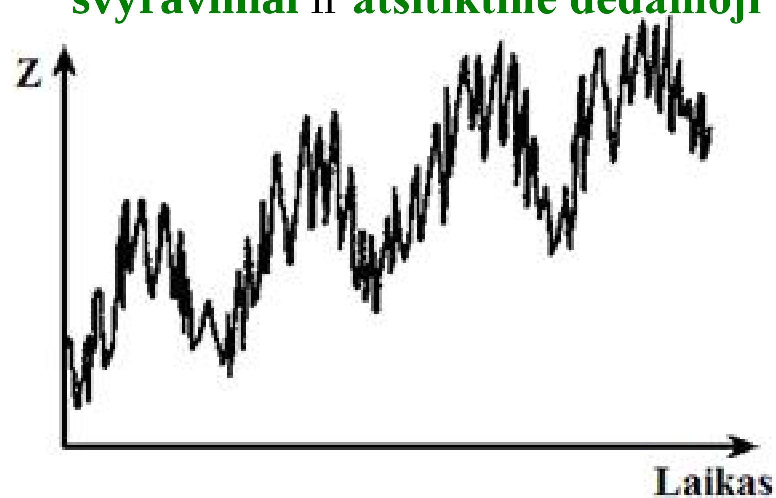
Tiesinis trendas ir **atsitiktinė dedamoji**



Sezoniniai svyravimai ir **atsitiktinė dedamoji**

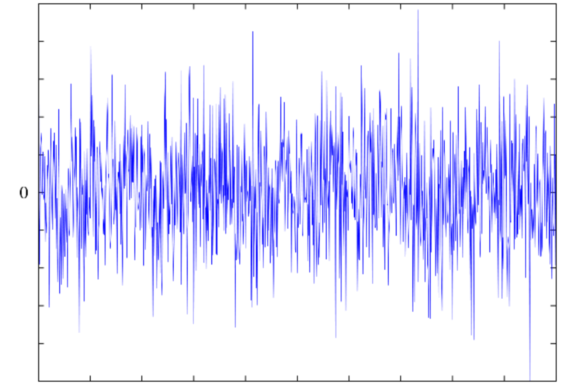


Tiesinis trendas, **sezoniniai svyravimai** ir **atsitiktinė dedamoji**

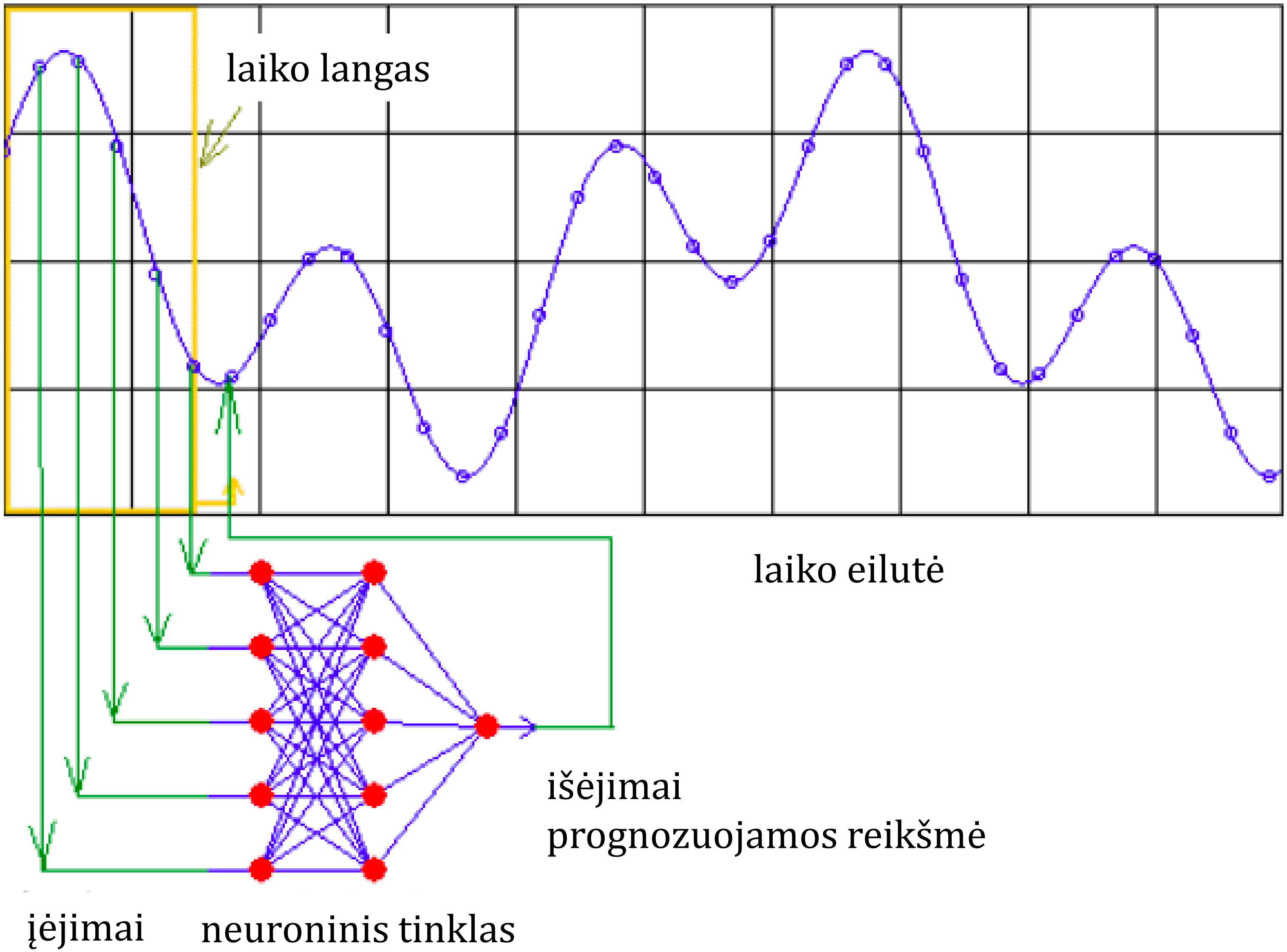


Kodėl DNT?

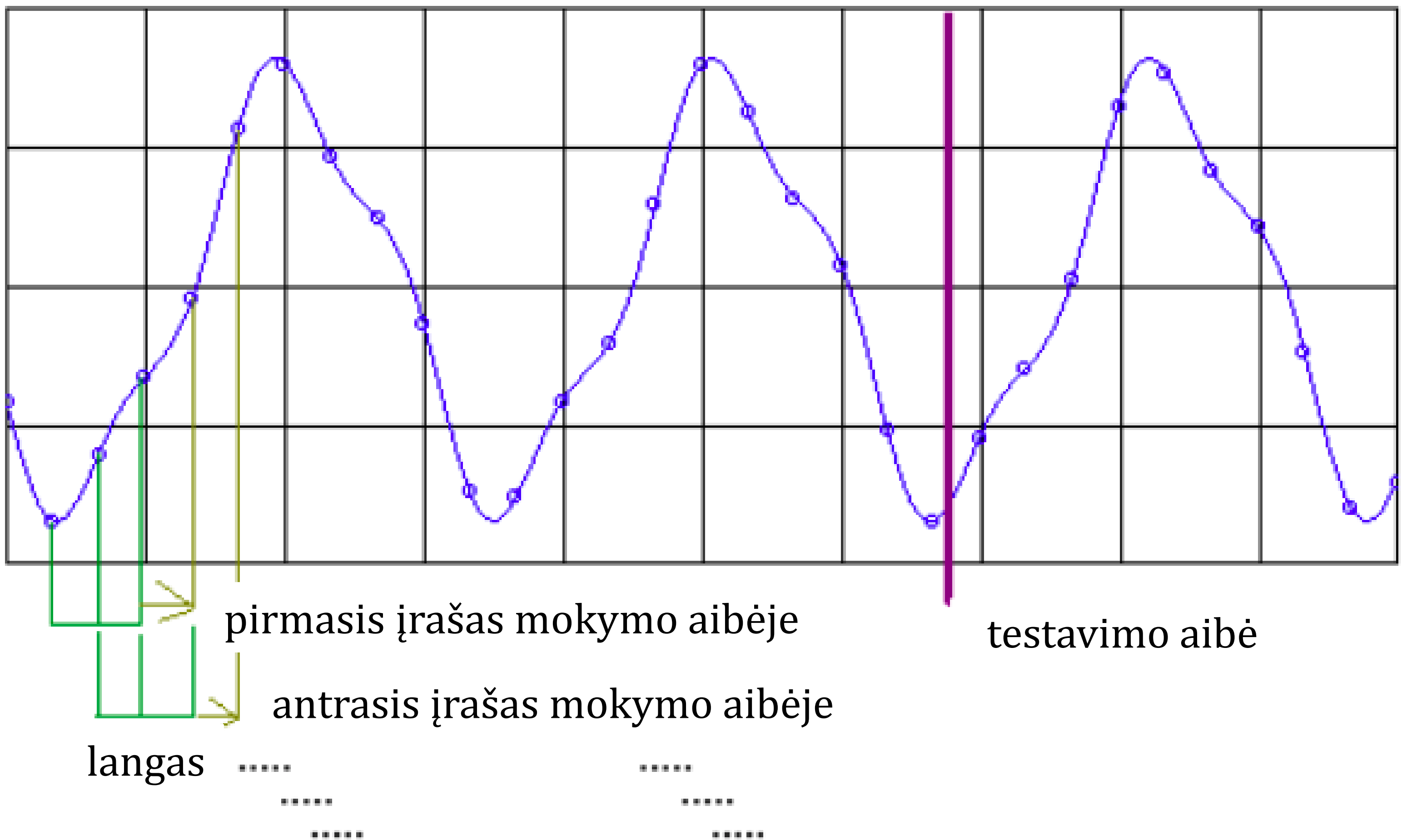
- DNT taikomi duomenis prognozuoti, kai statistiniai metodai **nepajėgūs to padaryti**.
- Sprendžiant realius uždavinius yra sunku nustatyti ar **tai triukšmas**, ar **tikros reikšmės**.
- Taikant **statistinius** prognozavimo metodus, būtina „**atpažinti**“ triukšmo tipą.
- **Baltasis triukšmas** – tai atsitiktinių dydžių seka, kurios vidurkis lygus nuliui, o standartinis nuokrypis lygus vienam.



DNT prognozavimui



Mokymo ir testavimo duomenys

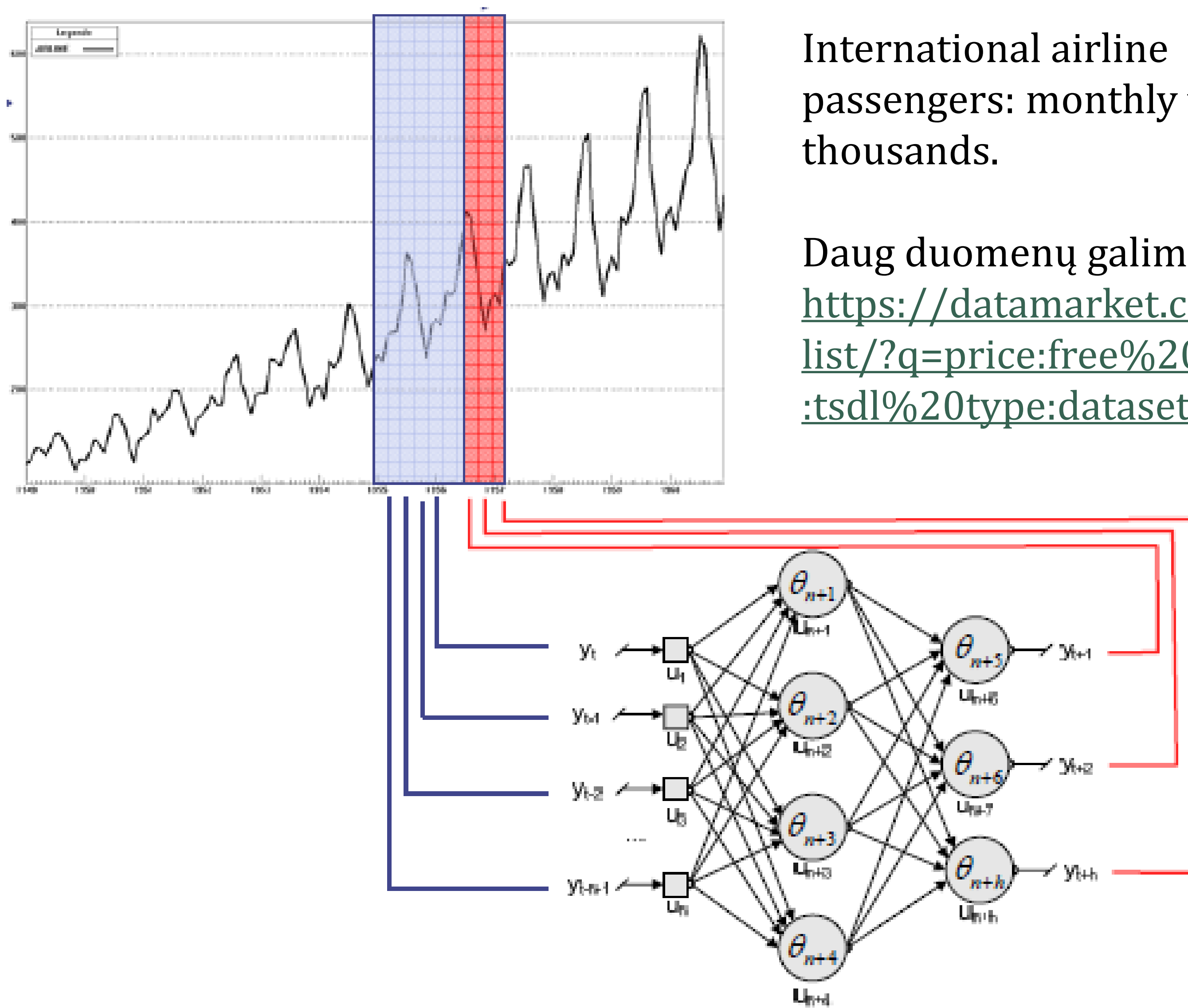


Prognozavimo pavyzdžiai

- Matematinės funkcijos atvejis:
  - http://www.obitko.com/tutorials/neural-network-prediction/function-prediction.html

- NASDAQ akcijų rinka:
  - http://www.obitko.com/tutorials/neural-network-prediction/nasdaq-prediction.html

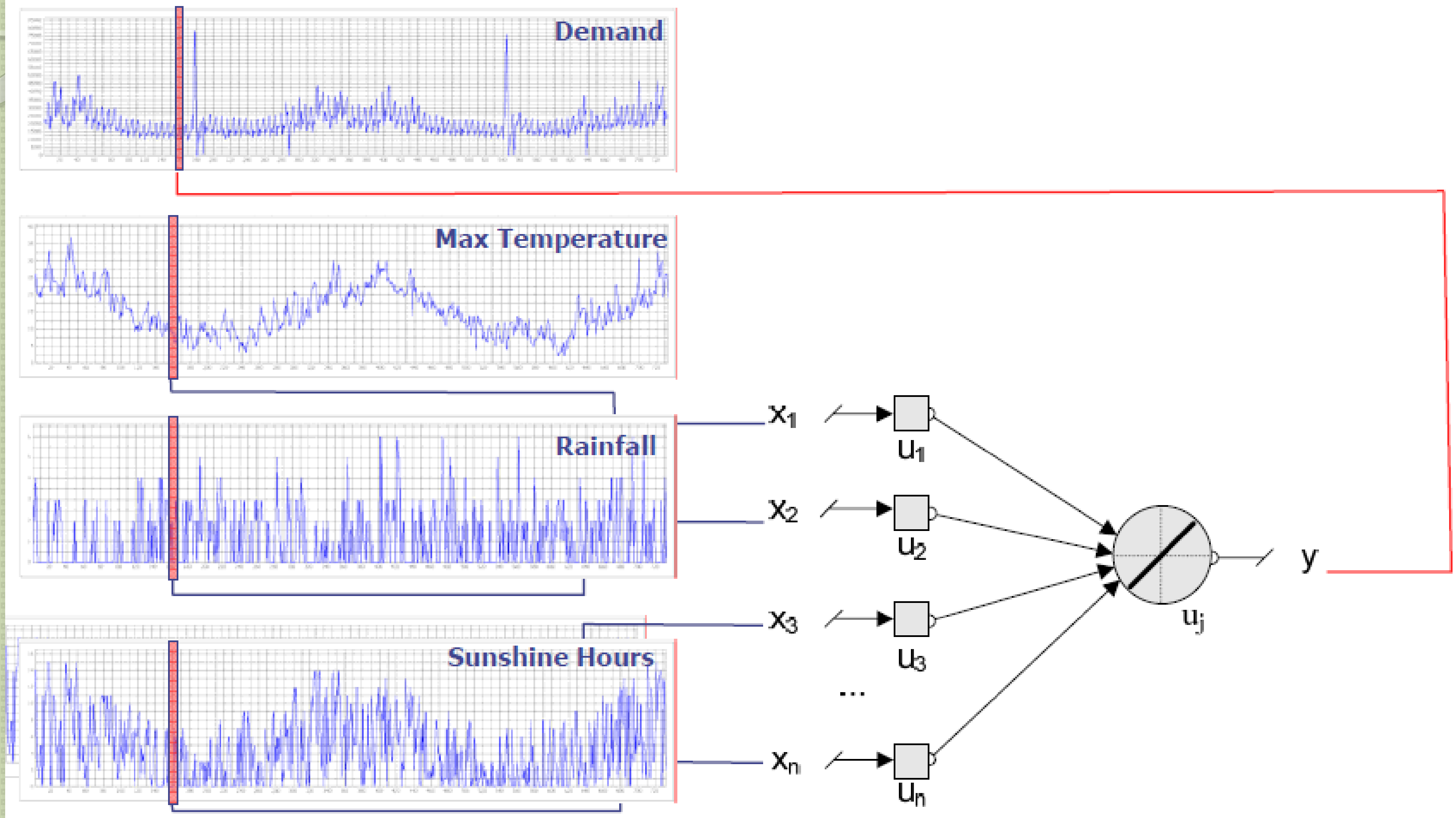
DNT prognozuojantis kelis išėjimus



International airline passengers: monthly totals in thousands.

Daug duomenų galima rasti: https://datamarket.com/data/list/?q=price:free%20provider:tsdl%20type:dataset

DNT kelių laiko eilučių atveju



Prognozavimo tikslumo matai

- Tegu y_i yra prognozuojama, o t_i – tikra reikšmė, m – duomenų kiekis.
- Vidutinė absoliuti paklaida (*mean absolute error*)

$$\mathbf{MAE} = \frac{1}{m}\sum_{i=1}^{m}|t_i - y_i|,$$

- Vidutinė kvadratinė paklaida (*mean squared error*)

$$\mathbf{MSE} = \frac{1}{m}\sum_{i=1}^{m}(t_i - y_i)^2,$$

- Šaknis iš vidutinės kvadratinės paklaidos (root mean squared error)

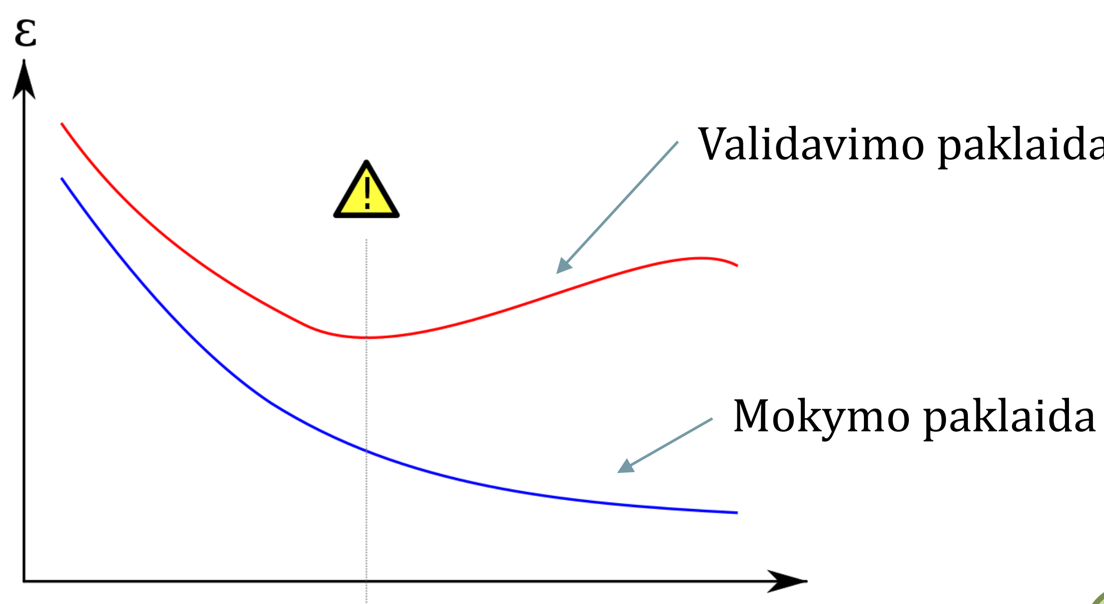
$$\mathbf{RMSE} = \frac{1}{m}\sqrt{\sum_{i=1}^{m}(t_i - y_i)^2}.$$

Prognozavimo taikant DNT etapai

- **Mokymas** (*training*) – tai procesas, kurio metu į DNT pateikiami mokymo aibės duomenys, siekiant nustatyti tinkamas DNT **svorių reikšmės**.
- **Validavimas** (*validation*) – tai procesas, kurio metu į DNT pateikiami validavimo aibės duomenys siekiant nustatyti tinkamus **kitus** DNT **parametrus** (ne svorius). Šis procesas taip pat reikalingas siekiant „**nepermokinti**“ (*overfitting*) neuroninį tinklą.
- **Testavimas** (*testing*) – tai procesas, kurio metu į DNT pateikiami testavimo aibės duomenys, kurie nebuvo naudojami DNT mokyme, siekiant nustatyti **prognozavimo tikslumą**.

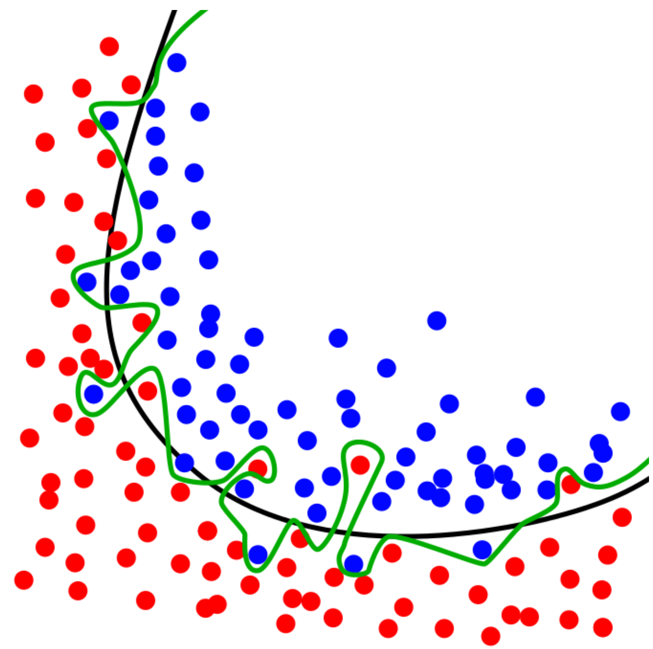
Neuroninio tinklo permokymas

- Mokant neuroninį tinklą, reikia stengtis jo „**nepermokti**“, kuomet tinklas labai prisiderina prie mokymo duomenų, tačiau jis nebus pajėgus tiksliai prognozuoti (klasifikuoti) naujus duomenis.



Permokytas modelis

- Nagrinėjamas **dviejų** klasių atvejis.
- **Žalias** skiriamas paviršius gautas „permokyto“ modelio, **juodas** – tinkamo modelio.
- Nors žalia kreivė **geriausiai** skiria mokymo aibės klases, tačiau ji **nepajėgs tiksliai skirti** naujų duomenų klases.



DNT, taikomi duomenims prognozuoti

- Daugiasluoksnis **perceptronas**
- **Rekurentiniai** neuroniniai tinklai:
  - Long short-term memory (LSTM)

Rekurentiniai neuroniniai tinklai

- **Rekurentiniai tinklai** turi ciklus, kurie leidžia informacijai būti perduotai iš vieno tinklo žingsnio į kitą.