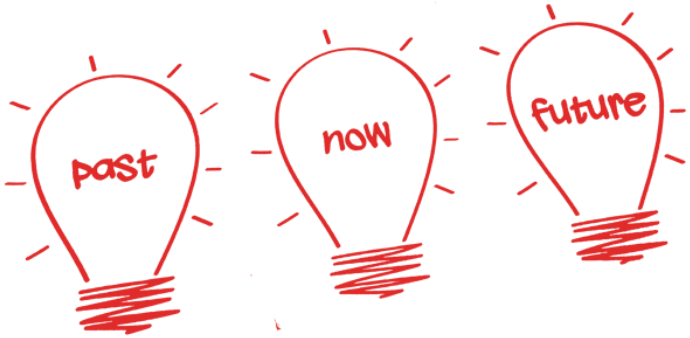
Vilniaus universitetas
Matematikos ir informatikos fakultetas

Dirbtiniai neuroniniai tinklai duomenims prognozuoti

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

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** (daugiasluoksniai perceptronai) 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.

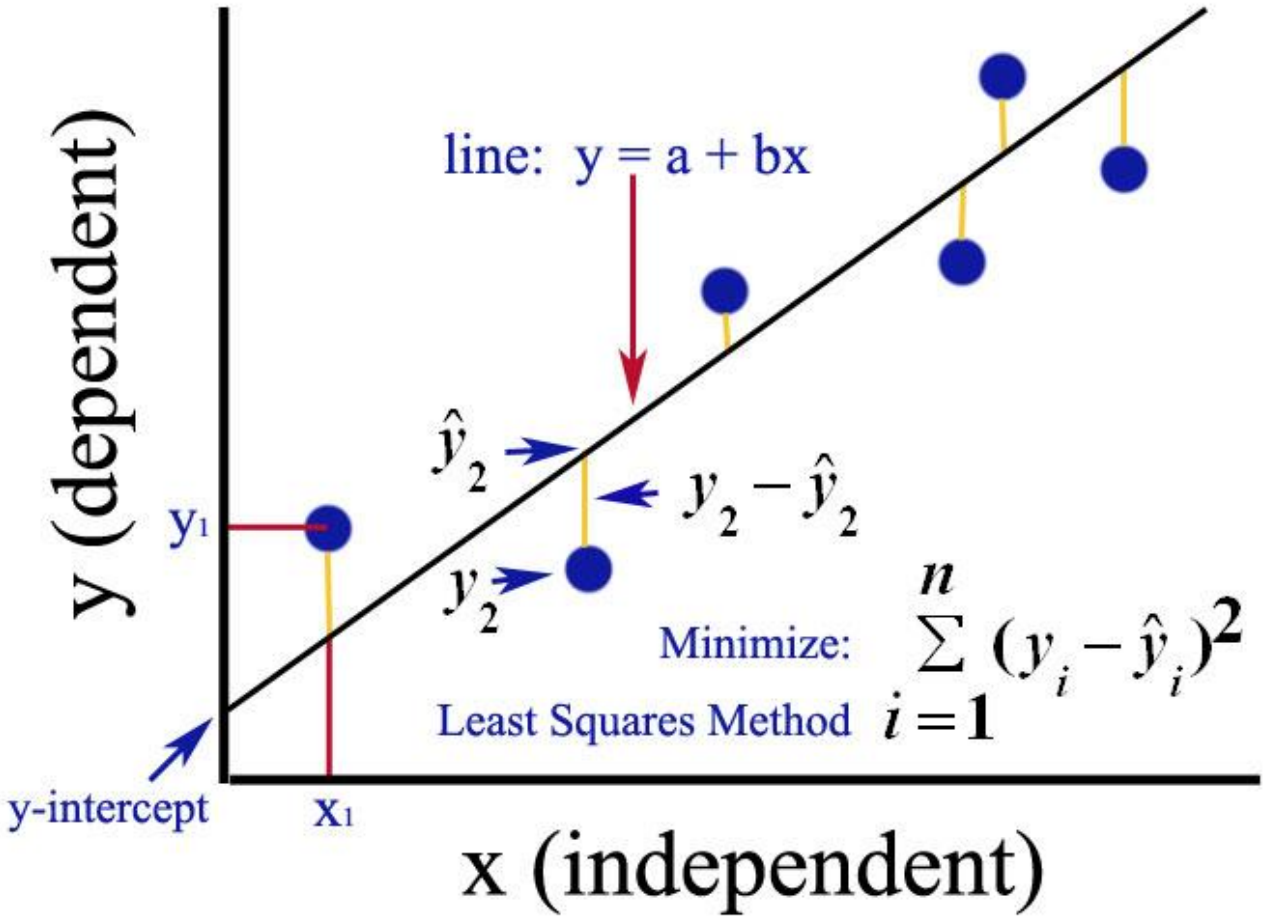
Tiesinė regresijos lygtis

- 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** $(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)$.
- **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, $j = 1, \dots, m$.

Tiesinės regresijos lygties koeficientų radimas

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

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

- 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:

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

Determinacijos koeficientas

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

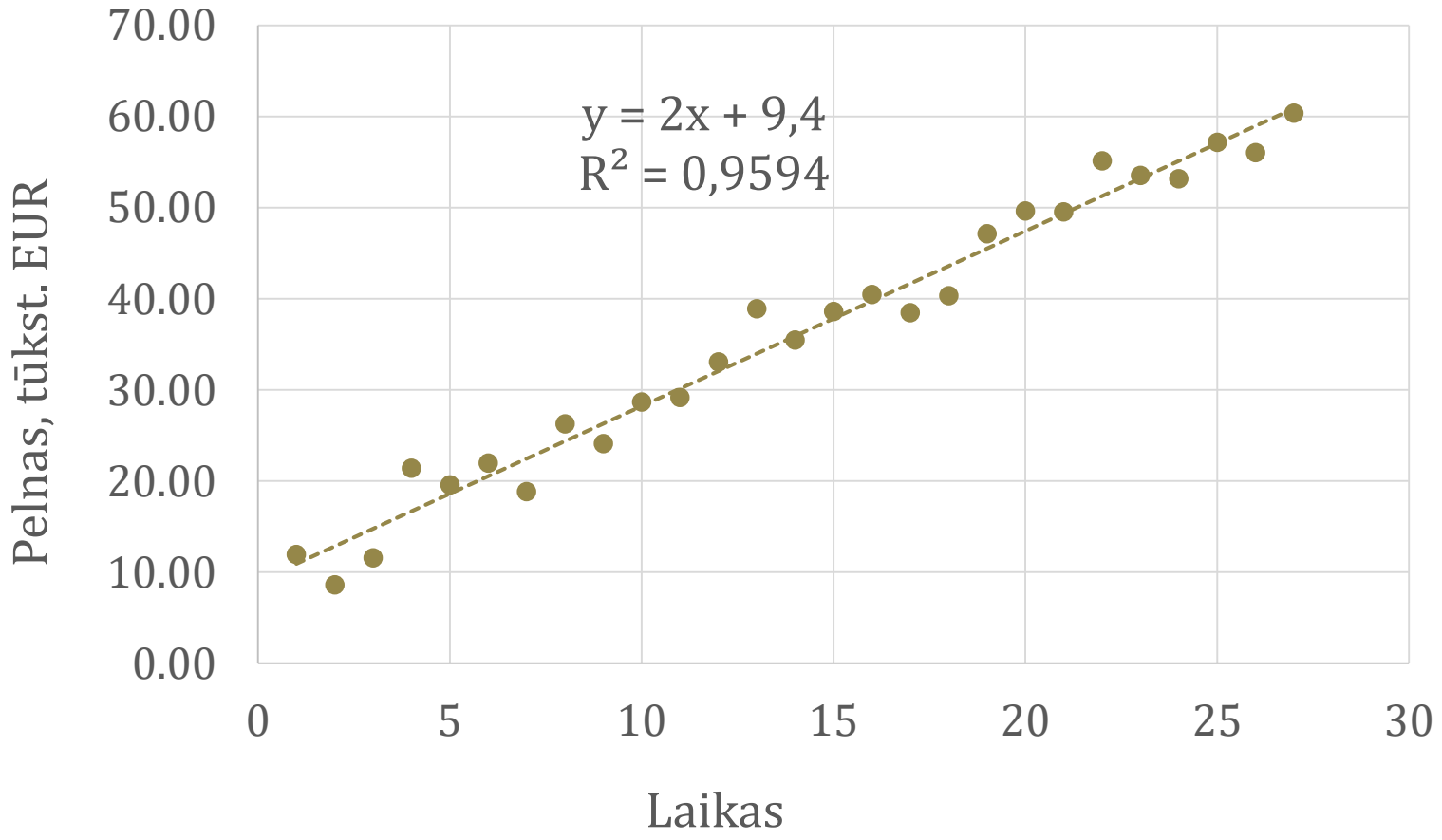
$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

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

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

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

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**:

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

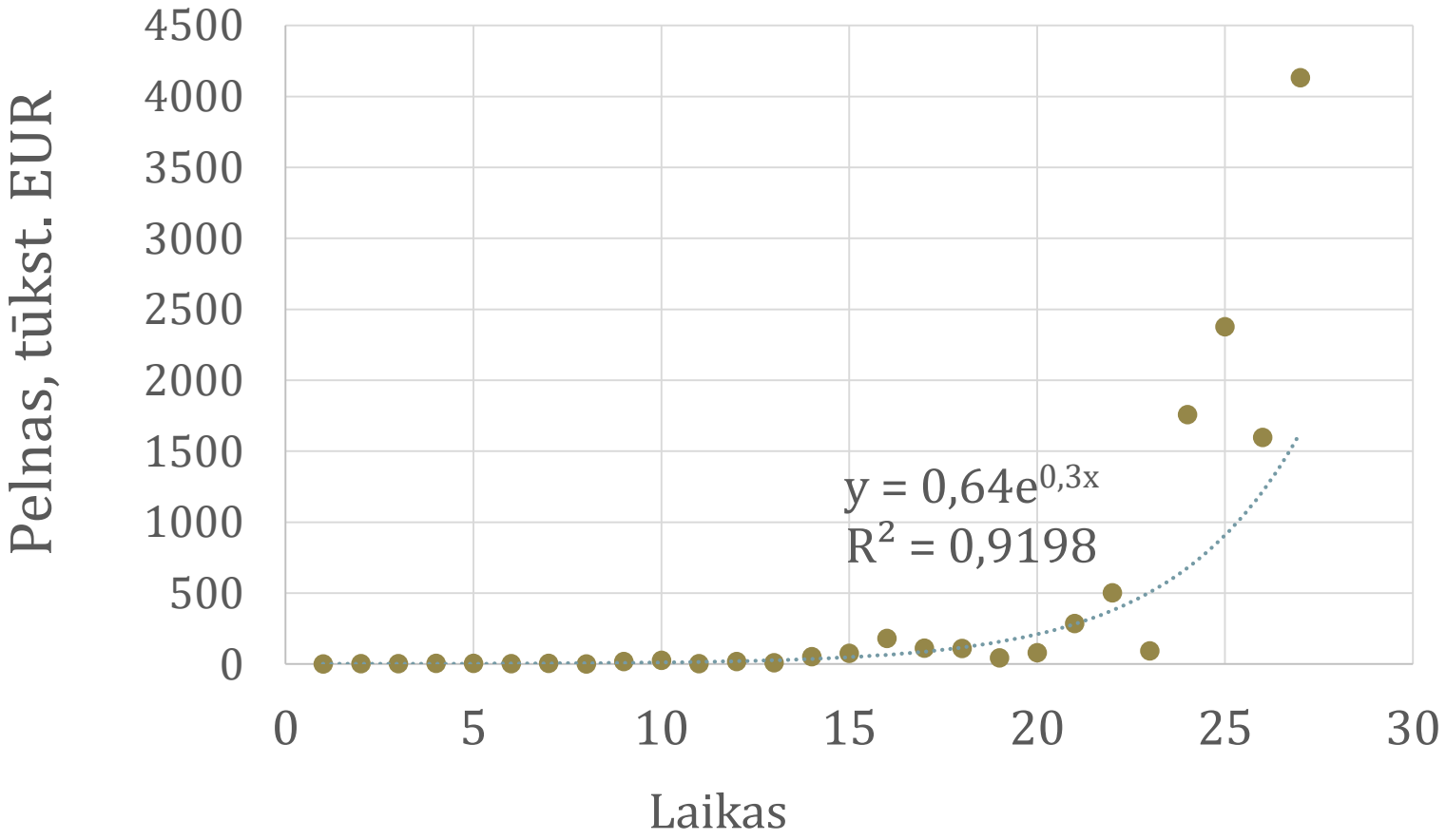
- **laipsninė** (**rodiklinė**):

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

- **polinominė** (n-tojo laipsnio):

$$y = a + b_1x + b_2x^2 + \cdots + b_nx^n + e.$$

Eksponentinė regresija



Laiko eilutės

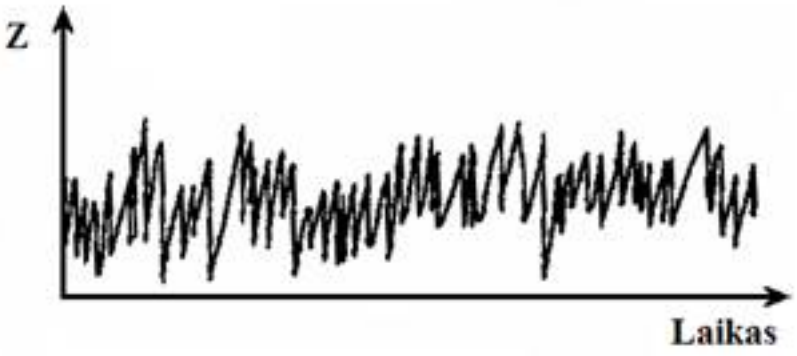
- Įprastai prognozavimo uždaviniai sprendžiami nagrinėjant vadinamąsias **laiko eilutes**.
- Tarkime tam tikro atsitiktinio dydžio X reikšmės stebimos laikui bėgant. Tokio atsitiktinio dydžio reikšmių seka $(X_1, X_2, \dots, X_t)$ vadinama **laiko eilute** (*time series*).
- Įprastai laikoma, kad yra 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

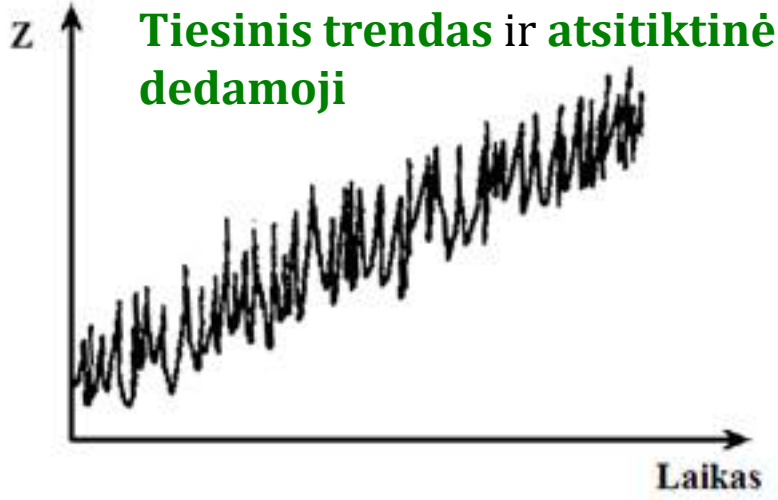
- Dažnai laiko eilutėse stebimos dvi dedamosios: **atsitiktinė** ir **apibrėžtoji**.
- **Apibrėžtosios** dedamosios dalys:
  - **trendas** (atspindi pagrindines bei ilgalaikes laiko eilutės tendencijas, esminius tiriamo proceso bruožus; trendas gali būti tiesinis, eksponentinis ir kt.),
  - **sezoniniai 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

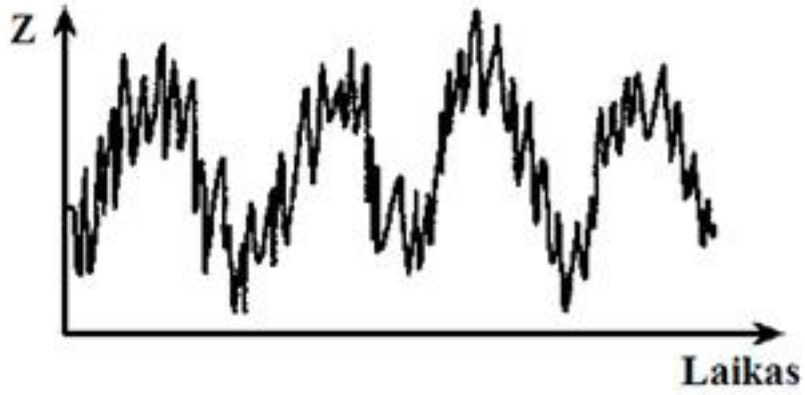
Vien tik **atsitiktinė dedamoji**



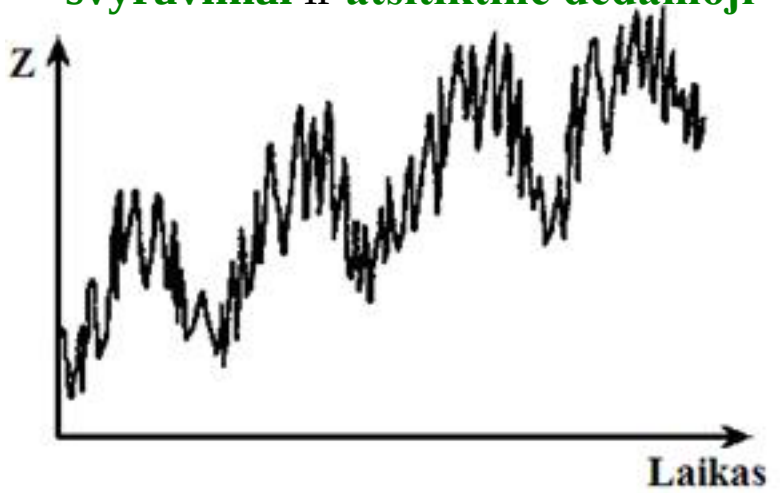
Tiesinis trendas ir **atsitiktinė dedamoji**



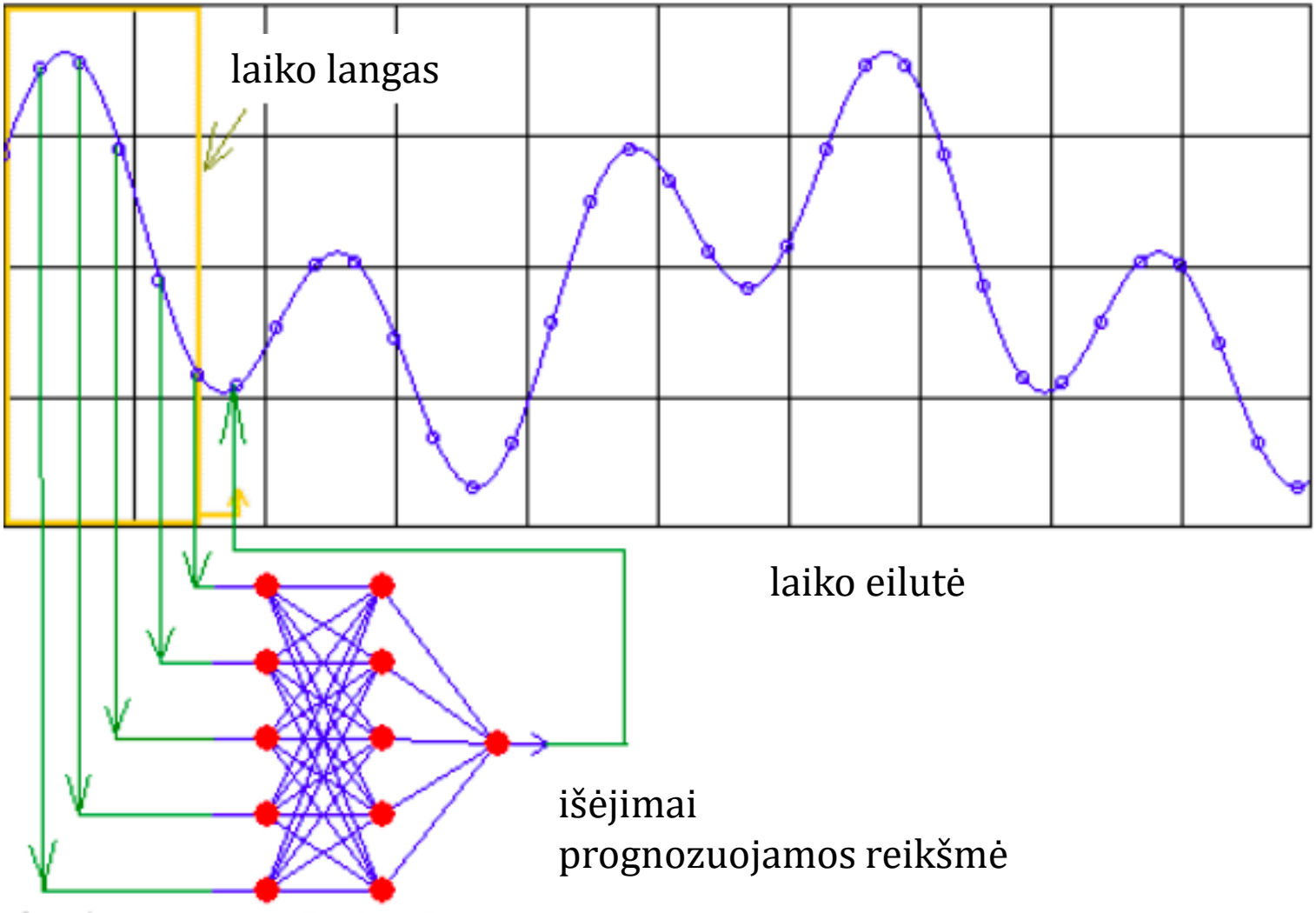
Sezoniniai svyravimai ir **atsitiktinė dedamoji**



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



DNT prognozavimui

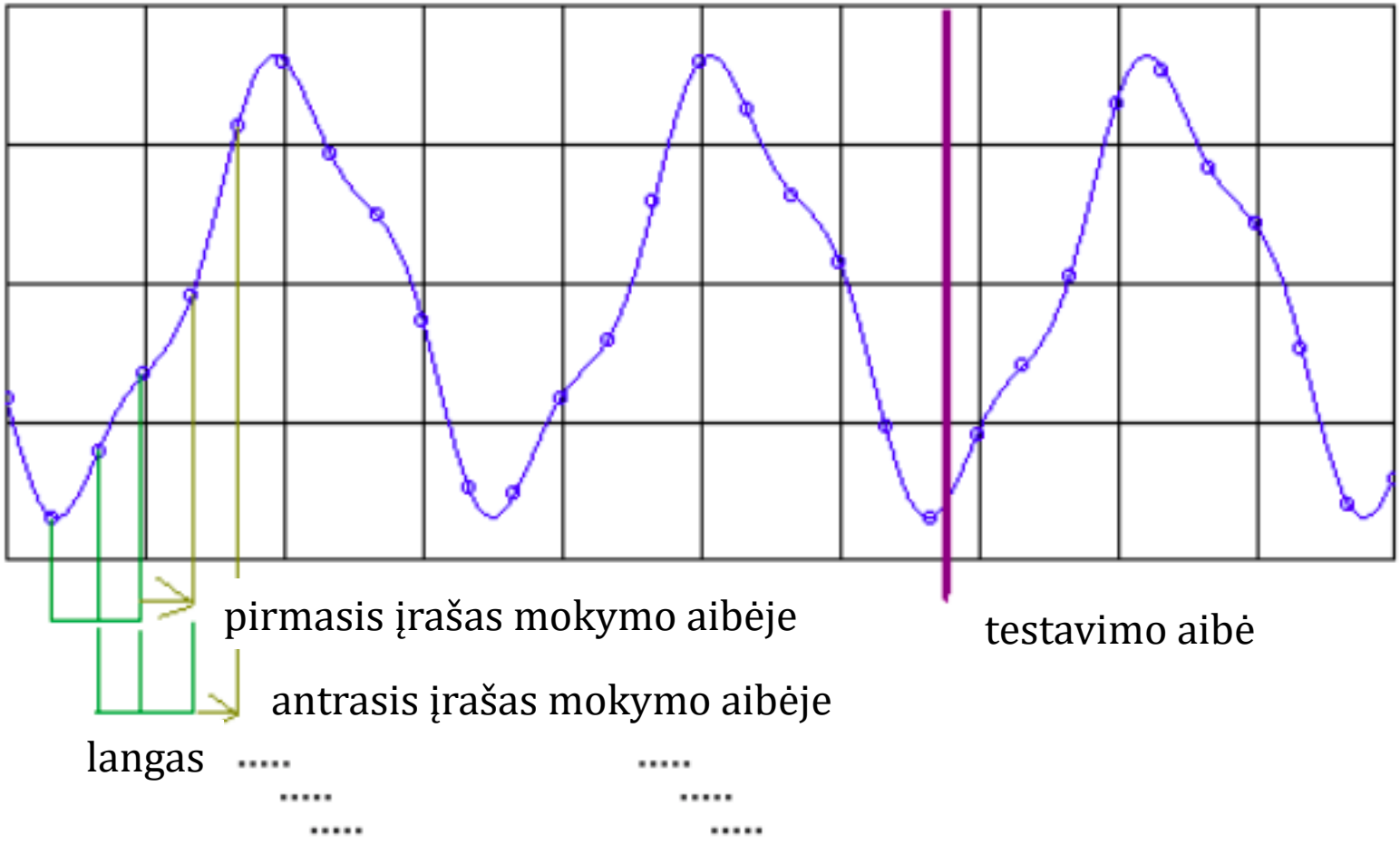


Duomenų suformavimas

- **Žali įrašai** – duomenys įvestims (*inputs*);
- **Raudoni langeliai** – išėjimų trokštamos reikšmės (*targets*);
- **Mėlynas įrašas** – nauji duomenys, kuriems reikia prognozuoti reikšmę **oranžiniame langelyje**.

2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
858	459	389	711	243	467	441	989	495	706	479	
250	476	800	494	722	811	945	841	264	711	957	
388	415	192	672	800	155	361	888	975	505	888	
633	451	146	526	572	640	228	377	109	146	165	
102	239	324	823	228	900	603	906	135	394	354	
728	890	305	149	621	650	191	197	350	310	795	
763	473	468	882	443	467	192	856	136	363	427	???

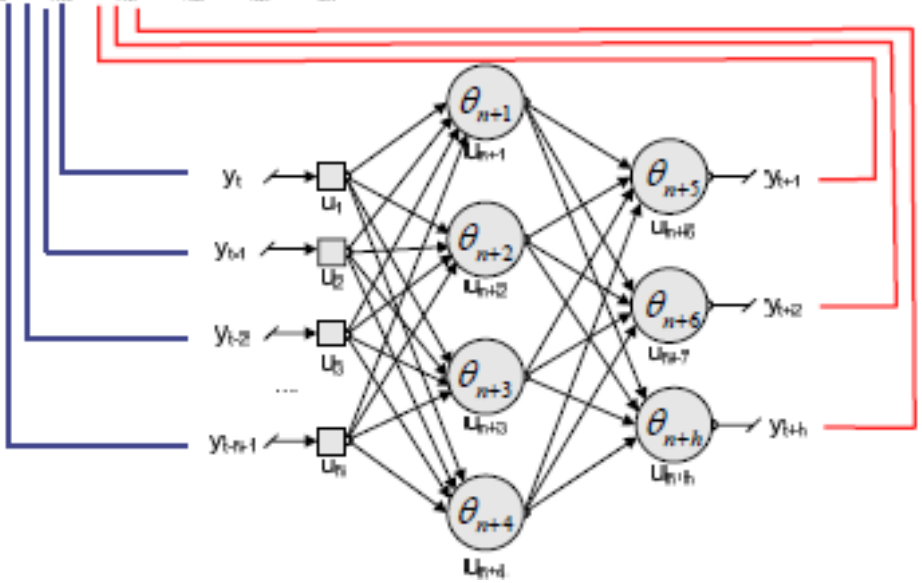
Mokymo ir testavimo duomenys



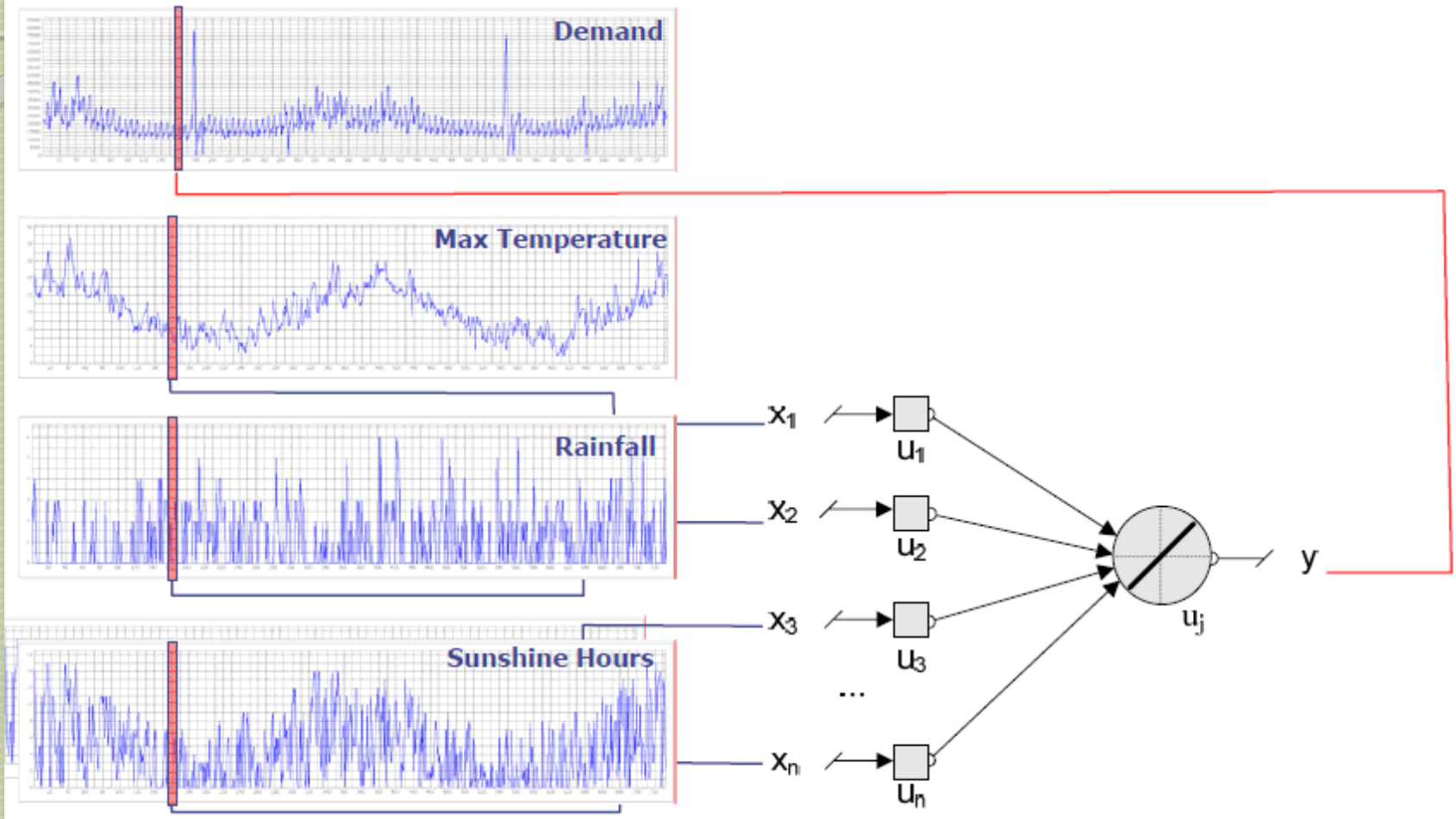
DNT prognozuojantis kelis išėjimus

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 (trokštama) 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*)

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

MAE are RMSE?

CASE 1: Evenly distributed errors

ID	Error	|Error|	Error^2
1	2	2	4
2	2	2	4
3	2	2	4
4	2	2	4
5	2	2	4
6	2	2	4
7	2	2	4
8	2	2	4
9	2	2	4
10	2	2	4

MAE	RMSE
2.000	2.000

CASE 2: Small variance in errors

ID	Error	|Error|	Error^2
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
6	3	3	9
7	3	3	9
8	3	3	9
9	3	3	9
10	3	3	9

MAE	RMSE
2.000	2.236

CASE 3: Large error outlier

ID	Error	|Error|	Error^2
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0
9	0	0	0
10	20	20	400

MAE	RMSE
2.000	6.325

https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

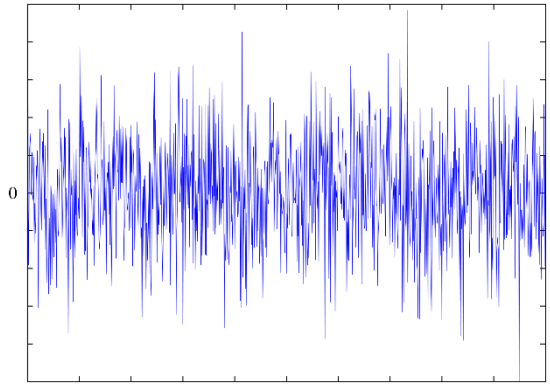
Neuroninių tinklų tipai

Galima išskirti **du pagrindinius neuroninių tinklų tipus**, kurie sėkmingai taikomi prognozavimo uždaviniams spręsti:

- **Daugiasluoksniai perceptronai** (žr. ankstesnes skaidres)
- **Rekurentiniai neuroniniai tinklai** (apie tai kitose paskaitose).

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.



Kodėl DNT?

- **Statistiniai metodai** atsižvelgia į duomenų **statistines charakteristikas** (pasiskirstymą, periodą ir kt.).
- Būtina įvertinti, kokia **matematinė funkcija geriausiai aproksimuoja** analizuojamus duomenis.
- Sprendžiant realius uždavinius, šias sąlygas dažnai sunku (arba visai neįmanoma) įvertinti. **DNT automatiškai adaptuojasi** prie analizuojamų duomenų charakteristikų.