PA081: Programování numerických výpočtů

9. Metódy redukcií dimenzionality

Aleš Křenek a Jana Hozzová

jaro 2019

Dáta vyzerajú aj takto ...

Cvs	Alcohol	Malic acid	Ash	Alcality of ash	Magnesium	Total phenols	Flavanoids	Nonflavanoid phenols	Proanthocyanins	Color intensity	Hue	0D280/0D315 of diluted wines	Proline
2	13.05	5.8	2.13	21.5	86	2.62	2.65	0.3	2.01	2.6	0.73	3.1	380
3	13.71	5.65	2.45	20.5	95	1.68	0.61	0.52	1.06	7.7	0.64	1.74	740
3	12.53	5.51	2.64	25	96	1.79	0.6	0.63	1.1	5	0.82	1.69	515
3	13.17	5.19	2.32	22	93	1.74	0.63	0.61	1.55	7.9	0.6	1.48	725
3	13.88	5.04	2.23	20	80	0.98	0.34	0.4	0.68	4.9	0.58	1.33	415
3	13.62	4.95	2.35	20	92	2	0.8	0.47	1.02	4.4	0.91	2.05	550
3	12.25	4.72	2.54	21	89	1.38	0.47	0.53	0.8	3.85	0.75	1.27	720
3	12.87	4.61	2.48	21.5	86	1.7	0.65	0.47	0.86	7.65	0.54	1.86	625
3	13.4	4.6	2.86	25	112	1.98	0.96	0.27	1.11	8.5	0.67	1.92	630
2	12.42	4.43	2.73	26.5	102	2.2	2.13	0.43	1.71	2.08	0.92	3.12	365
3	13.73	4.36	2.26	22.5	88	1.28	0.47	0.52	1.15	6.62	0.78	1.75	520
2	11.87	4.31	2.39	21	82	2.86	3.03	0.21	2.91	2.8	0.75	3.64	380
2	12.04	4.3	2.38	22	80	2.1	1.75	0.42	1.35	2.6	0.79	2.57	580
3	13.27	4.28	2.26	20	120	1.59	0.69	0.43	1.35	10.2	0.59	1.56	835
3	13.84	4.12	2.38	19.5	89	1.8	0.83	0.48	1.56	9.01	0.57	1.64	480
3	14.13	4.1	2.74	24.5	96	2.05	0.76	0.56	1.35	9.2	0.61	1.6	560
1	14.21	4.04	2.44	18.9	111	2.85	2.65	0.3	1.25	5.24	0.87	3.33	1080
1	14.22	3.99	2.51	13.2	128	3	3.04	0.2	2.08	5.1	0.89	3.53	760
1	13.24	3.98	2.29	17.5	103	2.64	2.63	0.32	1.66	4.36	0.82	3	680
3	13.4	3.91	2.48	23	102	1.8	0.75	0.43	1.41	7.3	0.7	1.56	750
3	13.08	3.9	2.36	21.5	113	1.41	1.39	0.34	1.14	9.4	0.57	1.33	550
3	12.25	3.88	2.2	18.5	112	1.38	0.78	0.29	1.14	8.21	0.65	2	855
2	12.7	3.87	2.4	23	101	2.83	2.55	0.43	1.95	2.57	1.19	3.13	463
2	13.05	3.86	2.32	22.5	85	1.65	1.59	0.61	1.62	4.8	0.84	2.01	515
1	13.41	3.84	2.12	18.8	90	2.45	2.68	0.27	1.48	4.28	0.91	3	1035
3	12.36	3.83	2.38	21	88	2.3	0.92	0.5	1.04	7.65	0.56	1.58	520
1	12.93	3.8	2.65	18.6	102	2.41	2.41	0.25	1.98	4.5	1.03	3.52	770
2	11.46	3.74	1.82	19.5	107	3.18	2.58	0.24	3.58	2.9	0.75	2.81	562
3	13.45	3.7	2.6	23	111	1.7	0.92	0.43	1.46	10.68	0.85	1.56	695
1	14.38	3.59	2.28	16	102	3.25	3.17	0.27	2.19	4.9	1.04	3.44	1065

Dimenzionalita dát

- dimenzie ako
  - viacero premenných, ktoré patria spolu
  - stĺpce v tabuľke
  - koordináty N-dimenzionálneho priestoru
- dáta ako
  - hodnoty premenných, ktoré patria spolu (jedno meranie, vlastnosti jedného objektu)
  - riadky v tabuľke
  - body v N-dimenzionálnom priestore

Prekliatie dimenzionality

- dáta zbierame, aby sme z nich niečo zistili
- to ide blbo, keď má tabuľka 50 stĺpcov
- prekliatie dimenzionality
  - čím viac dimenzií, tým exponenciálne viac dát potrebujeme
  - pri vysokom počte dimenzií sú dáta riedke
  - štatisticky významné závery s riedkymi dátami neurobíte

  Z toho vyplýva, že potrebujeme
- znížiť počet dimenzií
- ale zachovať podstatu dát (znalosti, vzory, vývodené dôsledky)
- sa zbaviť časti informácií

Redukcie dimenzionality dát

Redukcia dimenzií ako

- zbavenie sa premenných, ktoré sú redundanté a nepodstatné
- odstránenie niektoré stĺpce z tabuľky
- projekcia z originálneho, N-dimenzionálneho priestoru do redukovaného, M-dimenzionálneho priestoru, kde $M \ll N$

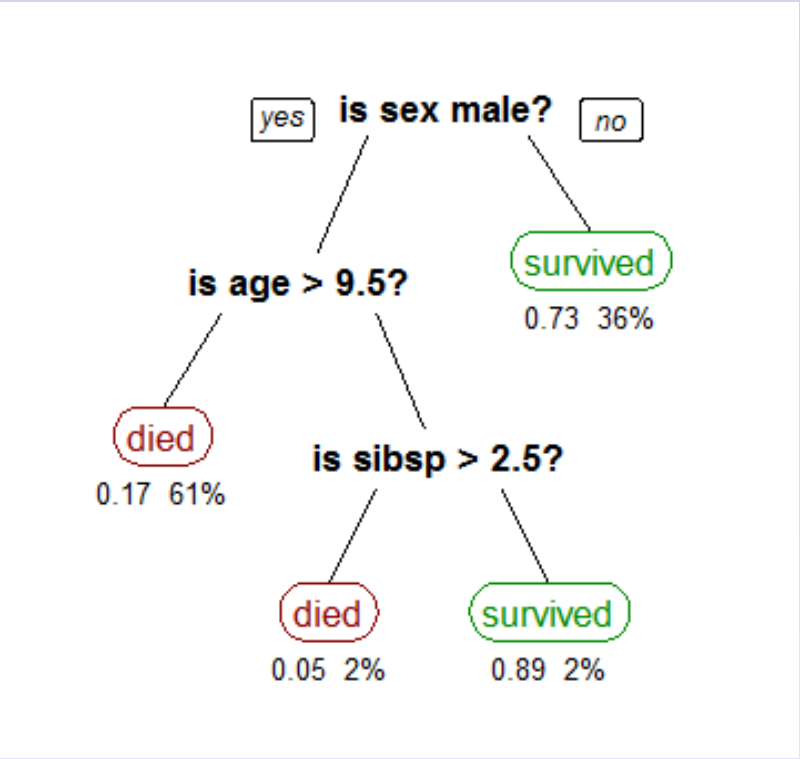
Na čo je to dobré?

- aproximuje model dát
- zjednodušuje pochopenie/interpretácia dát

Survived	Pclass	Sex	Age	SibSp
0	3	male	22	1
1	1	female	38	1
1	3	female	26	0
1	1	female	35	1
0	3	male	35	0
0	3	male		0
0	1	male	54	0
0	3	male	2	3
1	3	female	27	0
1	2	female	14	1
1	3	female	4	1
1	1	female	58	0
0	3	male	20	0
0	3	male	39	1
0	3	female	14	0
1	2	female	55	0

Na čo je to dobré?

- aproximuje model dát
- zjednodušuje pochopenie/interpretácia dát



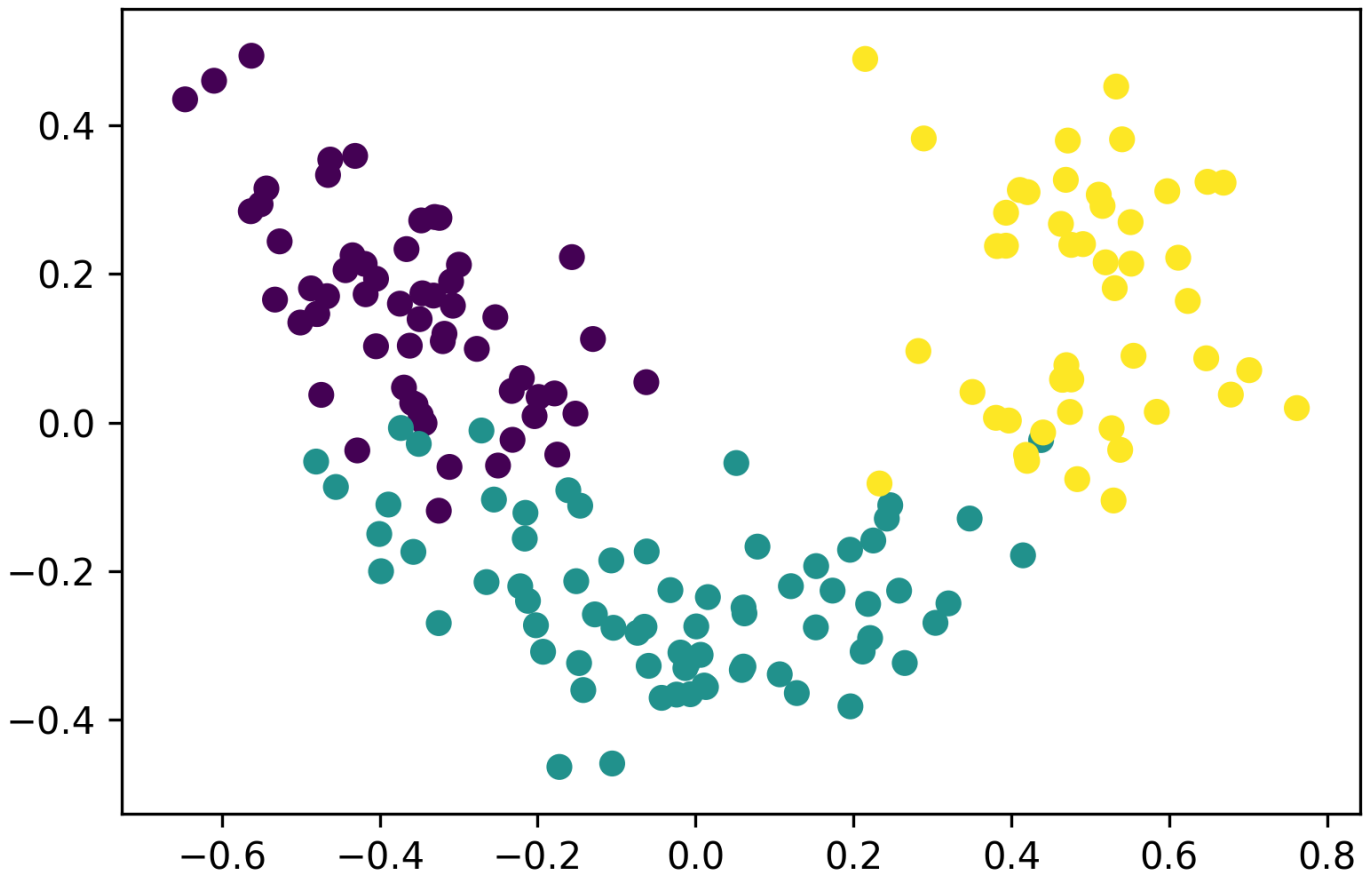
Na čo je to dobré?

- umožňuje dáta vizualizovať

Cvs	Alcohol	Malic acid	Ash	Alcality of ash	Magnesium	Total phenols	Flavanoids	Nonflavanoid phenols	Proanthocyanins	Color intensity	Hue	0D280/0D315 of diluted wines	Proline
2	13.05	5.8	2.13	21.5	86	2.62	2.65	0.3	2.01	2.6	0.73	3.1	380
3	13.71	5.65	2.45	20.5	95	1.68	0.61	0.52	1.06	7.7	0.64	1.74	740
3	12.53	5.51	2.64	25	96	1.79	0.6	0.63	1.1	5	0.82	1.69	515
3	13.17	5.19	2.32	22	93	1.74	0.63	0.61	1.55	7.9	0.6	1.48	725
3	13.88	5.04	2.23	20	80	0.98	0.34	0.4	0.68	4.9	0.58	1.33	415
3	13.62	4.95	2.35	20	92	2	0.8	0.47	1.02	4.4	0.91	2.05	550
3	12.25	4.72	2.54	21	89	1.38	0.47	0.53	0.8	3.85	0.75	1.27	720
3	12.87	4.61	2.48	21.5	86	1.7	0.65	0.47	0.86	7.65	0.54	1.86	625
3	13.4	4.6	2.86	25	112	1.98	0.96	0.27	1.11	8.5	0.67	1.92	630
2	12.42	4.43	2.73	26.5	102	2.2	2.13	0.43	1.71	2.08	0.92	3.12	365
3	13.73	4.36	2.26	22.5	88	1.28	0.47	0.52	1.15	6.62	0.78	1.75	520
2	11.87	4.31	2.39	21	82	2.86	3.03	0.21	2.91	2.8	0.75	3.64	380
2	12.04	4.3	2.38	22	80	2.1	1.75	0.42	1.35	2.6	0.79	2.57	580
3	13.27	4.28	2.26	20	120	1.59	0.69	0.43	1.35	10.2	0.59	1.56	835
3	13.84	4.12	2.38	19.5	89	1.8	0.83	0.48	1.56	9.01	0.57	1.64	480
3	14.13	4.1	2.74	24.5	96	2.05	0.76	0.56	1.35	9.2	0.61	1.6	560
1	14.21	4.04	2.44	18.9	111	2.85	2.65	0.3	1.25	5.24	0.87	3.33	1080
1	14.22	3.99	2.51	13.2	128	3	3.04	0.2	2.08	5.1	0.89	3.53	760
1	13.24	3.98	2.29	17.5	103	2.64	2.63	0.32	1.66	4.36	0.82	3	680
3	13.4	3.91	2.48	23	102	1.8	0.75	0.43	1.41	7.3	0.7	1.56	750
3	13.08	3.9	2.36	21.5	113	1.41	1.39	0.34	1.14	9.4	0.57	1.33	550
3	12.25	3.88	2.2	18.5	112	1.38	0.78	0.29	1.14	8.21	0.65	2	855
2	12.7	3.87	2.4	23	101	2.83	2.55	0.43	1.95	2.57	1.19	3.13	463
2	13.05	3.86	2.32	22.5	85	1.65	1.59	0.61	1.62	4.8	0.84	2.01	515
1	13.41	3.84	2.12	18.8	90	2.45	2.68	0.27	1.48	4.28	0.91	3	1035
3	12.36	3.83	2.38	21	88	2.3	0.92	0.5	1.04	7.65	0.56	1.58	520
1	12.93	3.8	2.65	18.6	102	2.41	2.41	0.25	1.98	4.5	1.03	3.52	770
2	11.46	3.74	1.82	19.5	107	3.18	2.58	0.24	3.58	2.9	0.75	2.81	562
3	13.45	3.7	2.6	23	111	1.7	0.92	0.43	1.46	10.68	0.85	1.56	695
1	14.38	3.59	2.28	16	102	3.25	3.17	0.27	2.19	4.9	1.04	3.44	1065

Na čo je to dobré?

- umožňuje dáta vizualizovať



Na čo je to dobré?

- znižuje výpočetný čas: metadynamika
- znižuje množstvo dát na ukladanie/prenos
- zjednodušuje/umožňuje vyvodenie štatisticky významných záverov z dát: autotuning, RNA-sekvencovanie
- umožňuje použiť algoritmy, ktoré nefungujú vôbec/dobre s vysokým počtom dimenzií
- znižuje riziko overfittingu
- znižuje množstvo šumu v dátach: RNA-sekvencovanie
- zvyšuje presnosť výsledkov

Základná štatistika

- rozptyl (variance)

$$\sigma^2 = mean((x - \mu)^2)$$

- štandardná odchýlka σ
- pomáha nám táto premenná odlíšiť dáta od seba?
- pri redukcii dimenzií: nízky alebo vysoký rozptyl?

Základná štatistika

- kovariancia (covariance)

$$cov(X,Y) = mean((x-\mu_x)(y-\mu_y))$$

- korelácia (correlation)

$$corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

- dá sa táto premenná odvodiť z nejakej inej?
- pri redukcii dimenzií: nízka alebo vysoká kovariancia?

Dva základné prístupy

Výber premenných/rysov (feature selection)

- vyberá z existujúcich premenných
- orezáva N-dimenzionálny proestor rušením niektorých koordinátov

Extrakcia rysov (feature extraction)

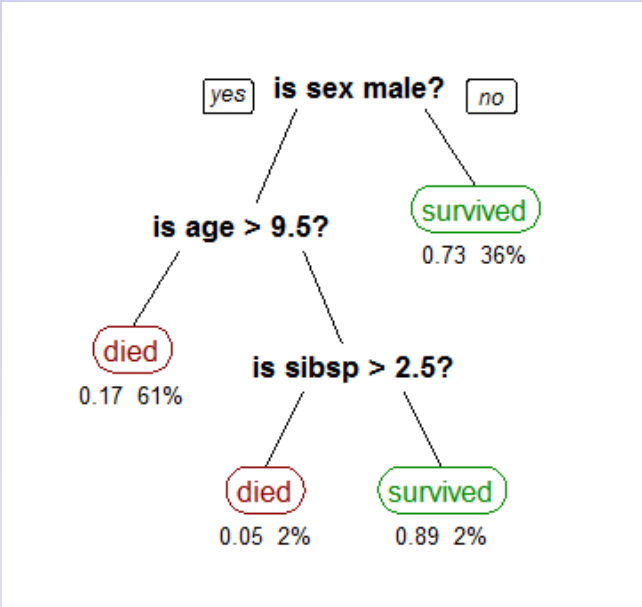
- vytvára nové premenné (ktoré viem spočítať z existujúcich)
- vytváva nový M-dimenzionálny priestor, ktorý má nový systém koordinátov

Výber premenných

- podmnožina premenných
- irelevantné: nezáleží na nich (nízky rozptyl)
- redundantné: dajú sa odvodiť z inej premennej (vysoká kovariancia)
- metódy
  - hľadajú najlepšiu podmnožinu dimenzií
  - postupne vytvárajú nový model dát

Náhodný les

- rozhodovací strom vytvára vetvy, aby čo najlepšie rozdelil dáta

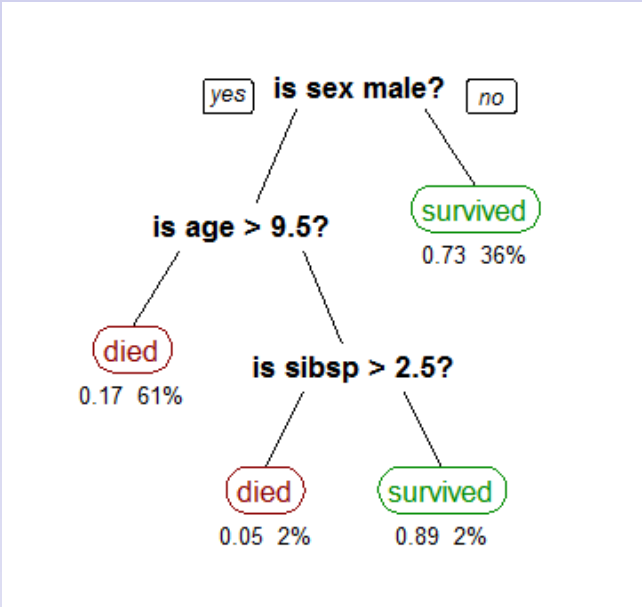


[1]

- málo robustné, náchylné na overfitting

[1] Stephen Milborrow, https://commons.wikimedia.org/wiki/File:CART_tree_titanic_survivors.png

Náhodný les

- rozhodovací strom vytvára vetvy, aby čo najlepšie rozdelil dáta 

[1]
- málo robustné, náchylné na overfitting
- viacero stromov: les
- jeden strom = podmnožina dát s podmnožinou premenných
- pri predikcii sa rozhodujú všetky stromy a ich výsledky sa "spriemerujú"

[1] Stephen Milborrow, `https://commons.wikimedia.org/wiki/File:CART_tree_titanic_survivors.png`

Jednoduchý príklad

prežitie na Titanicu

Komplexný príklad - Autotuning

- portabilita výkonu, napr. pri prechode na iné GPU
- jeden kód, viacero parametrov
- hodnoty parametrov na danom hw určujú výkon
- je ich veľa a navzájom sa ovplyvňujú

Niekedy stromy nestačia

- veľmi riedke dáta
- iné než kolmé hrany závislostí znamenajú veľké stromy
- časové rady, dáta vo forme grafov, obrázky
- niekedy sú podstatné premenné ukryté medzi originálnymi dimenziami

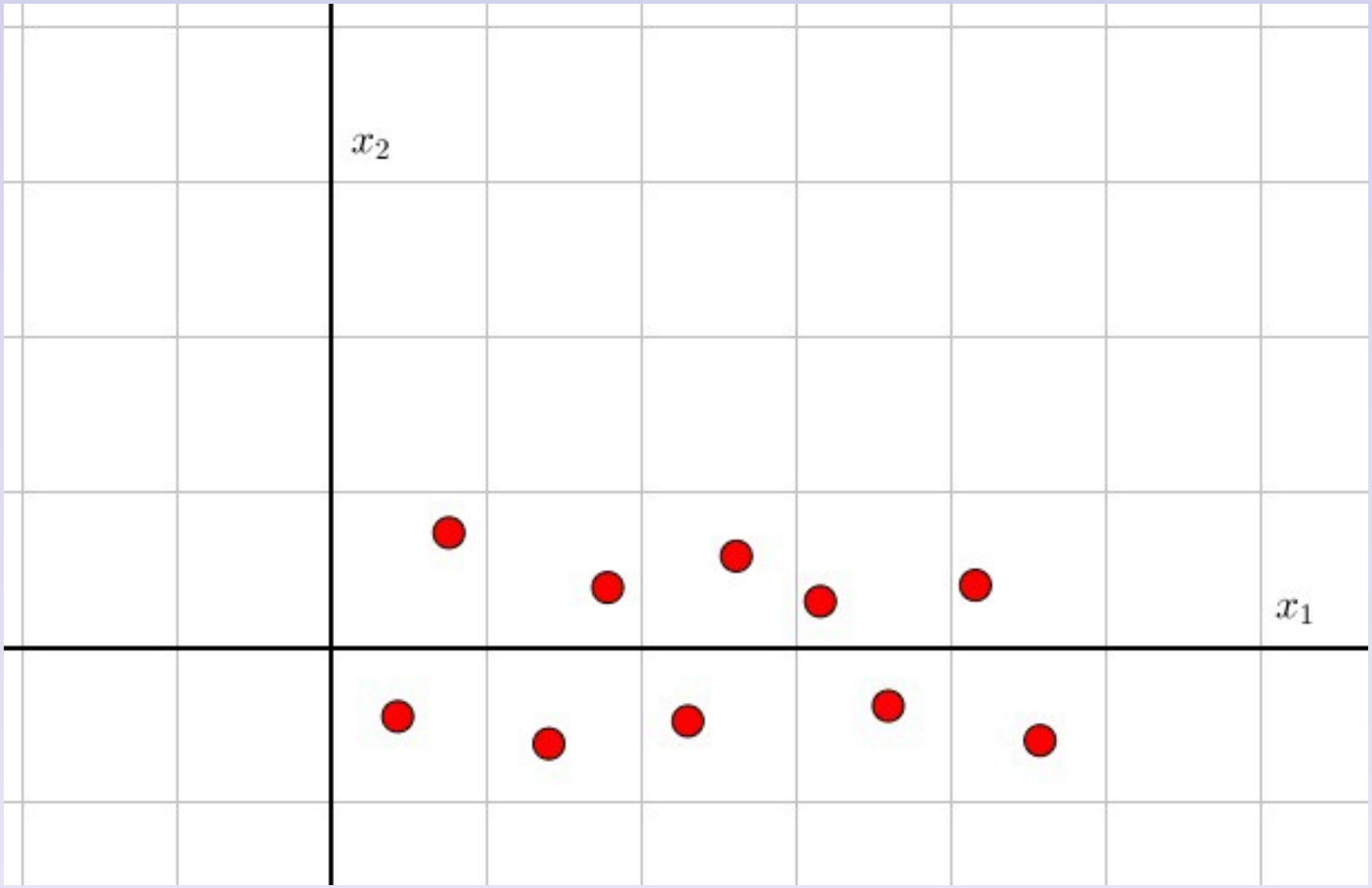
Extrakcia redukovaných dimenzií

- nové premenné/rysy/koordináty v redukovanom priestore
- pomerne presne popisujú dátovú sadu
- dajú sa spočítať z pôvodných premenných
- majú vysoký rozptyl, nízka kovariancia
- (niekedy) zachovávajú vlastnosti pôvodného priestoru (napr. vzdialenosti medzi bodmi)

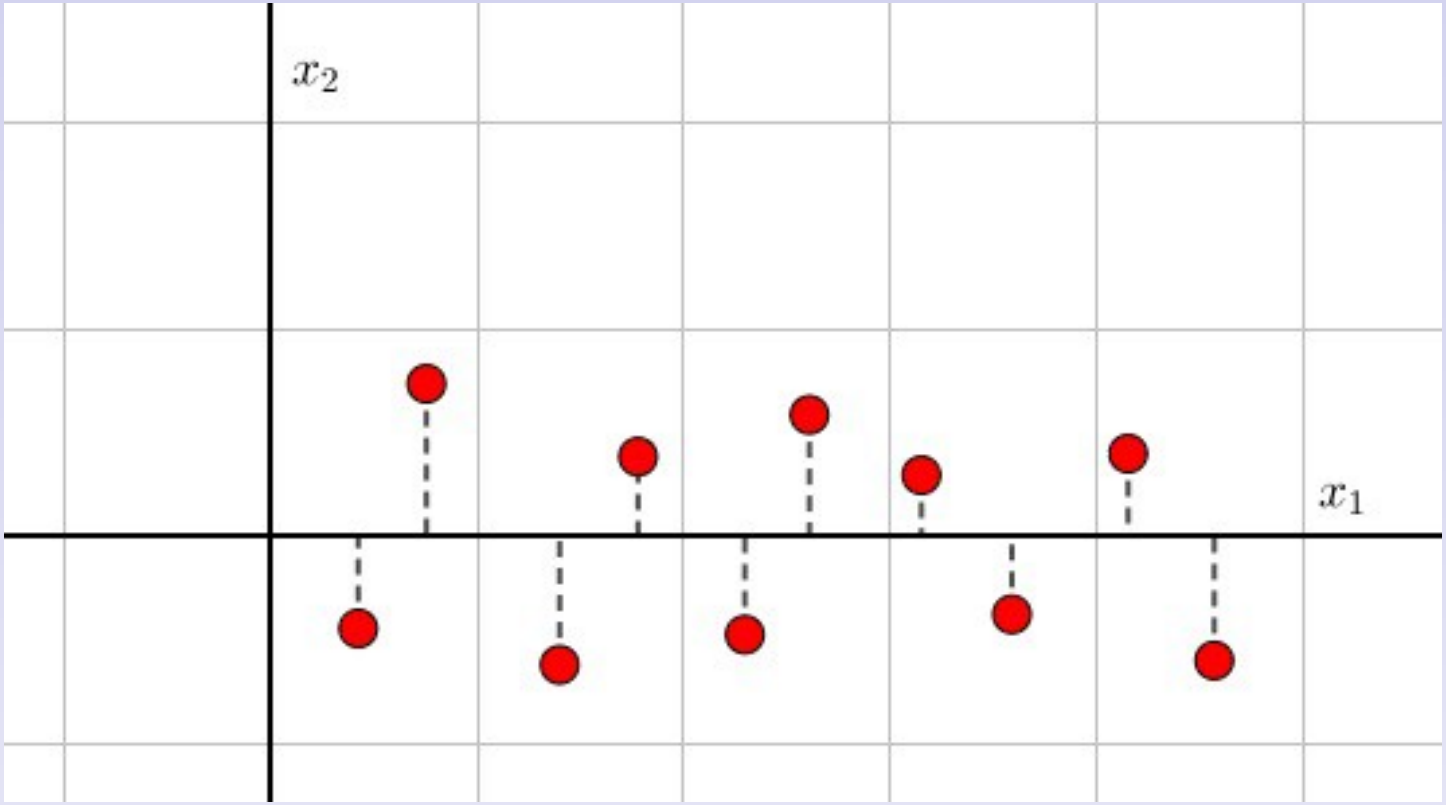
Druhy metód

- využívajú lineárne závislosti medzi premennými (PCA)
- zachovávajú vzdialenosti medzi bodmi v priestore (MDS, Isomap)
- aproximujú nelineárne závislosti ako vážený súčet lineárnych fitovaní (LLE)
- využívajú lokalitu dát pomocou spektrálnych metód (Spectral Embedding)
- využívajú rozloženie pravdepodobnosti medzi bodmi (t-SNE)
- využívajú skryté závislosti pomocou strojového učenia

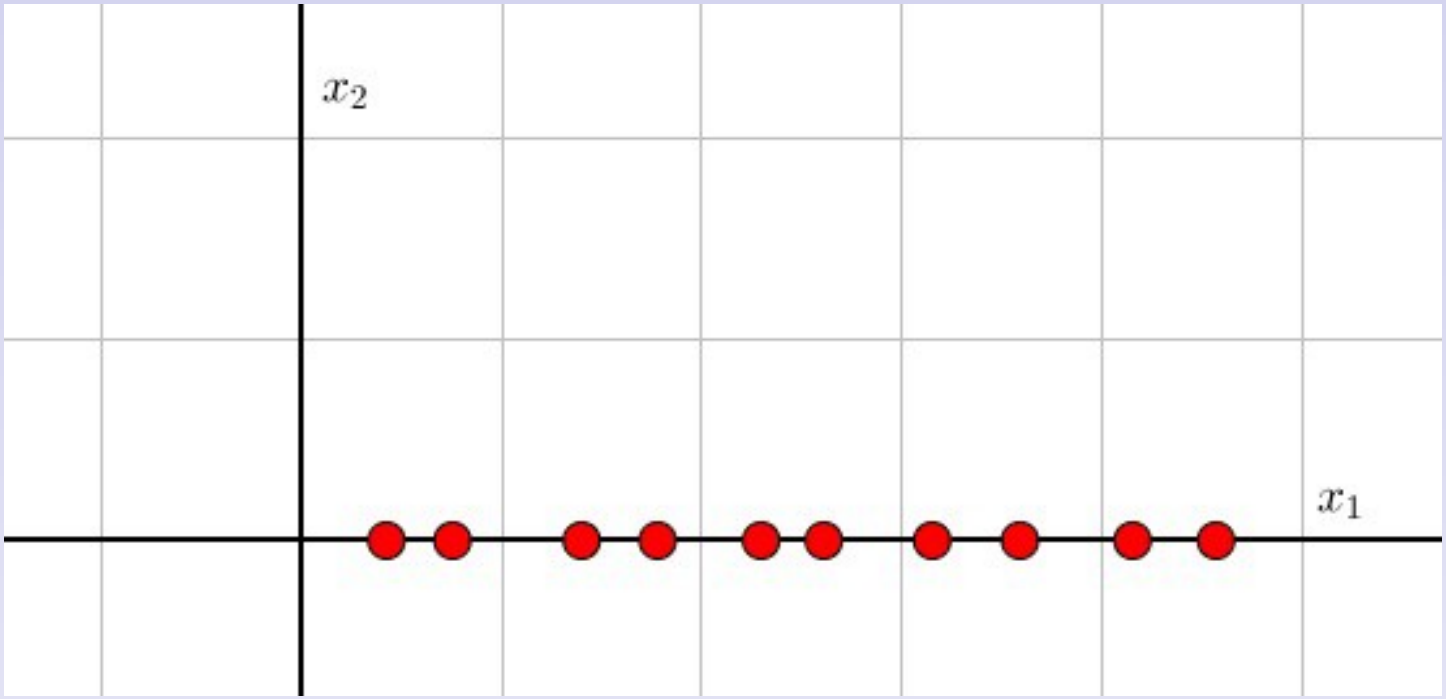
PCA - princíp



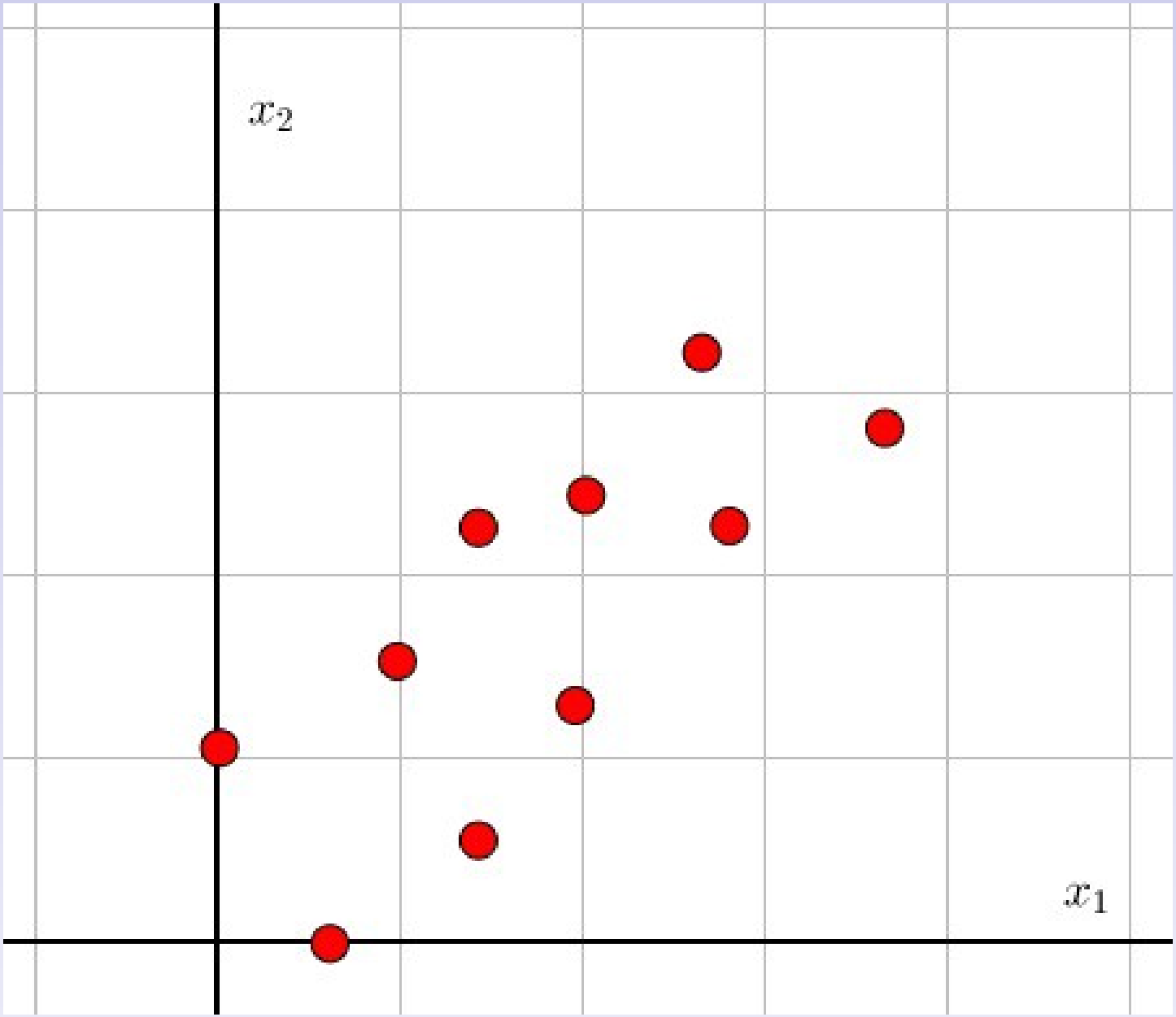
PCA - princíp



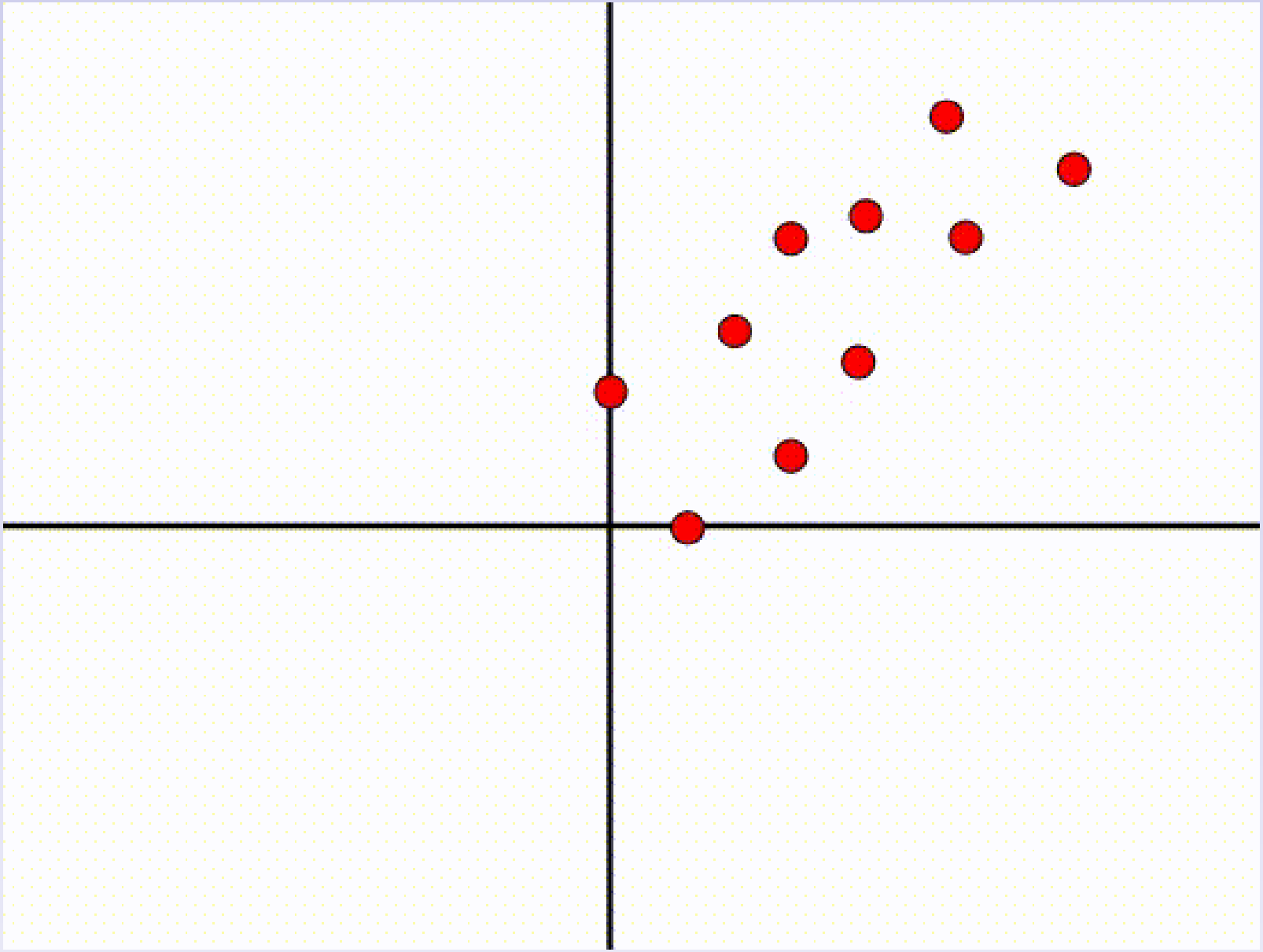
PCA - princíp



PCA - princíp



PCA - princíp



PCA - Ako sa to spočíta?

- spočítame kovariačnú maticu
- nájdeme jej vlastné čísla a vlastné vektory
- vlastné vektory najvyšších vlastných čísel ukazujú najvýznamnejšie smery rozptylu premenných
- projekcia z originálneho priestoru do redukovaného priestoru
- $$X'_m = XW_m$$

  kde W_m je matica so stĺpcami vlastných vektorov kovariačnej matice

PCA - Čo to znamená?

- hľadáme smer najvyššieho rozptylu

$$Variance\ along\ the\ line =$$
$$\sigma_v^2 = p^T \cdot p = (Av)^T(Av) = v^T(A^TA)v = v^T\ C\ v$$

- kde A sú naše dáta, ν je hľadaný smer, p je projekcia na hľadaný smer
- hľadáme ν také, aby σ^2, rozptyl, bolo maximálne
- kovariačná matica C zachytáva rozptyl aj kovarianciu dimenzií [2]

$$Covariance\ Matrix = X^TX$$
$$= \begin{pmatrix} — & x & — \\ — & y & — \end{pmatrix} \begin{pmatrix} | & | \\ x & y \\ | & | \end{pmatrix} = \begin{pmatrix} x^Tx & x^Ty \\ y^Tx & y^Ty \end{pmatrix} =$$
$$\begin{pmatrix} cov(x,x) & cov(x,y) \\ cov(y,x) & cov(y,y) \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & cov(x,y) \\ cov(y,x) & \sigma_y^2 \end{pmatrix}$$

- najvyššie vlastné číslo ~ veľkosť vektora ν
- smer korešpondujúceho vlastného vektora ~ smer vektora ν

[2] https://towardsdatascience.com/dimensionality-reduction-for-dummies-part-2-3b1e3490bdc9

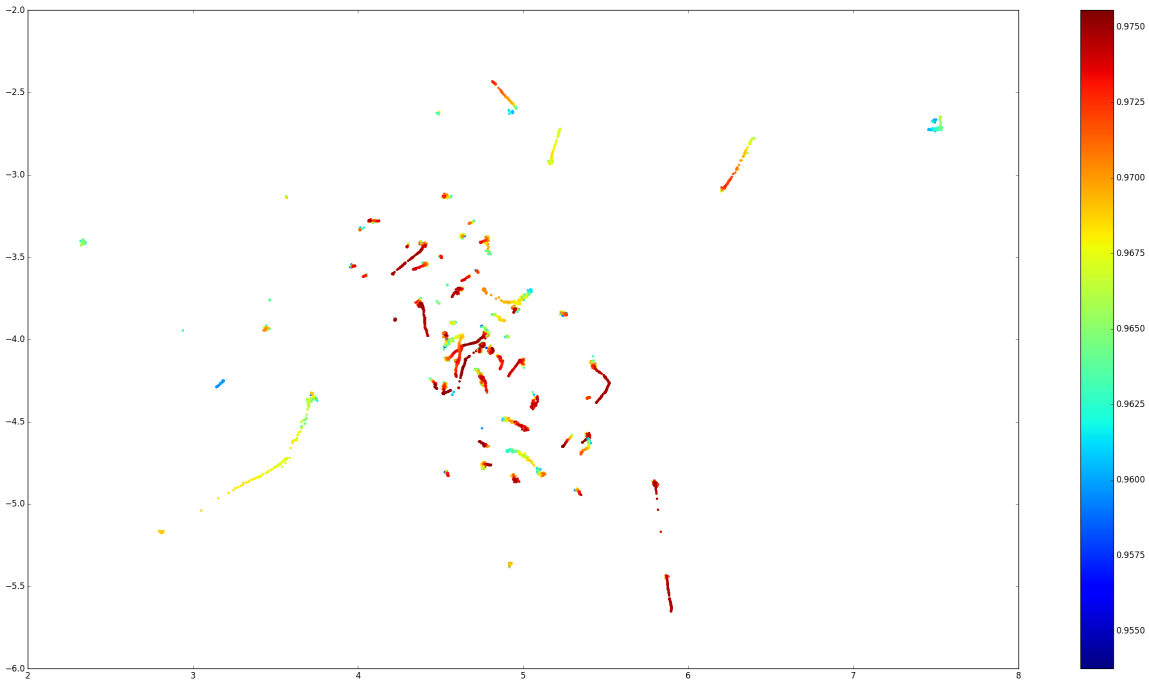
Jednoduchý príklad

odrody vína

Komplexný príklad - Atómové náboje

- výpočet parciálnych atómových nábojov (prednáška Optimalizace)
- presné ale výpočetne náročne kvantové metódy
- aproximatívne ale rýchle empirické metódy
- systém lineárnych rovníc s parametrami
- trénovacia sada nábojov spočítaných kvantovkov
- hľadáme také hodnoty parametrov, ktoré vyústia v také isté hodnoty nábojov
- optimalizačný problém na maximalizáciu korelácie
- ako vyzerá optimalizovaný priestor?
- redukcia dimenzií za účelom vizualizácie

Komplexný príklad - Atómové náboje



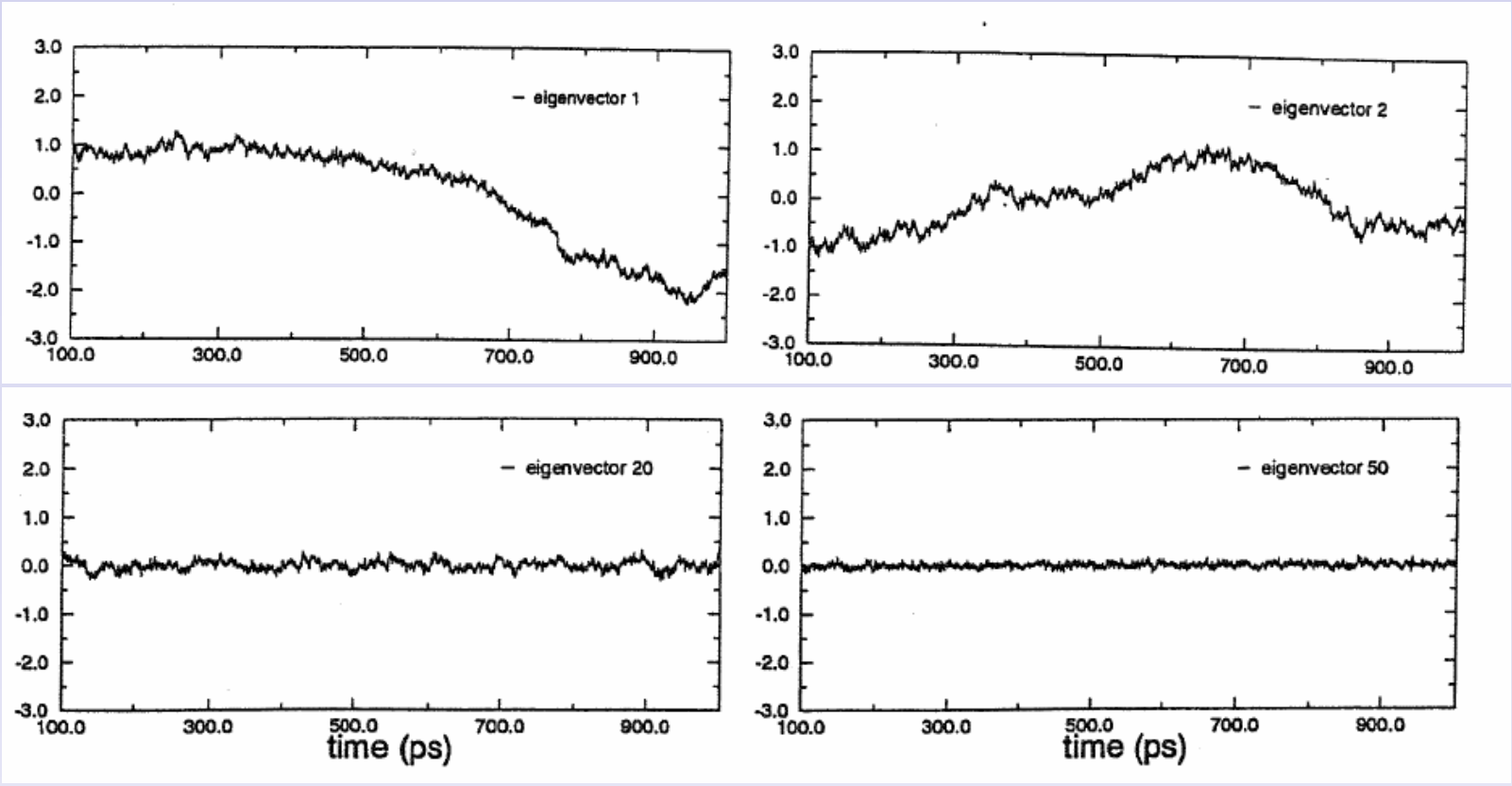
Komplexný príklad - MD trajektórie

- trajektórie molekulovej dynamiky
- v každom časovom bode koordináty každého atómu
- chaotický systém, ako porovnám dve trajektórie?
- esenciálne koordináty [3]
- vlastné hodnoty a vektory kovariačnej matice: PCA
- porovnávali sme vektory vlastných čísel

[3] Amadei, A., Linssen, A. B., Berendsen, H. J. (1993). Essential dynamics of proteins. Proteins, 17(4), 412–425. https://doi.org/10.1002/prot.340170408

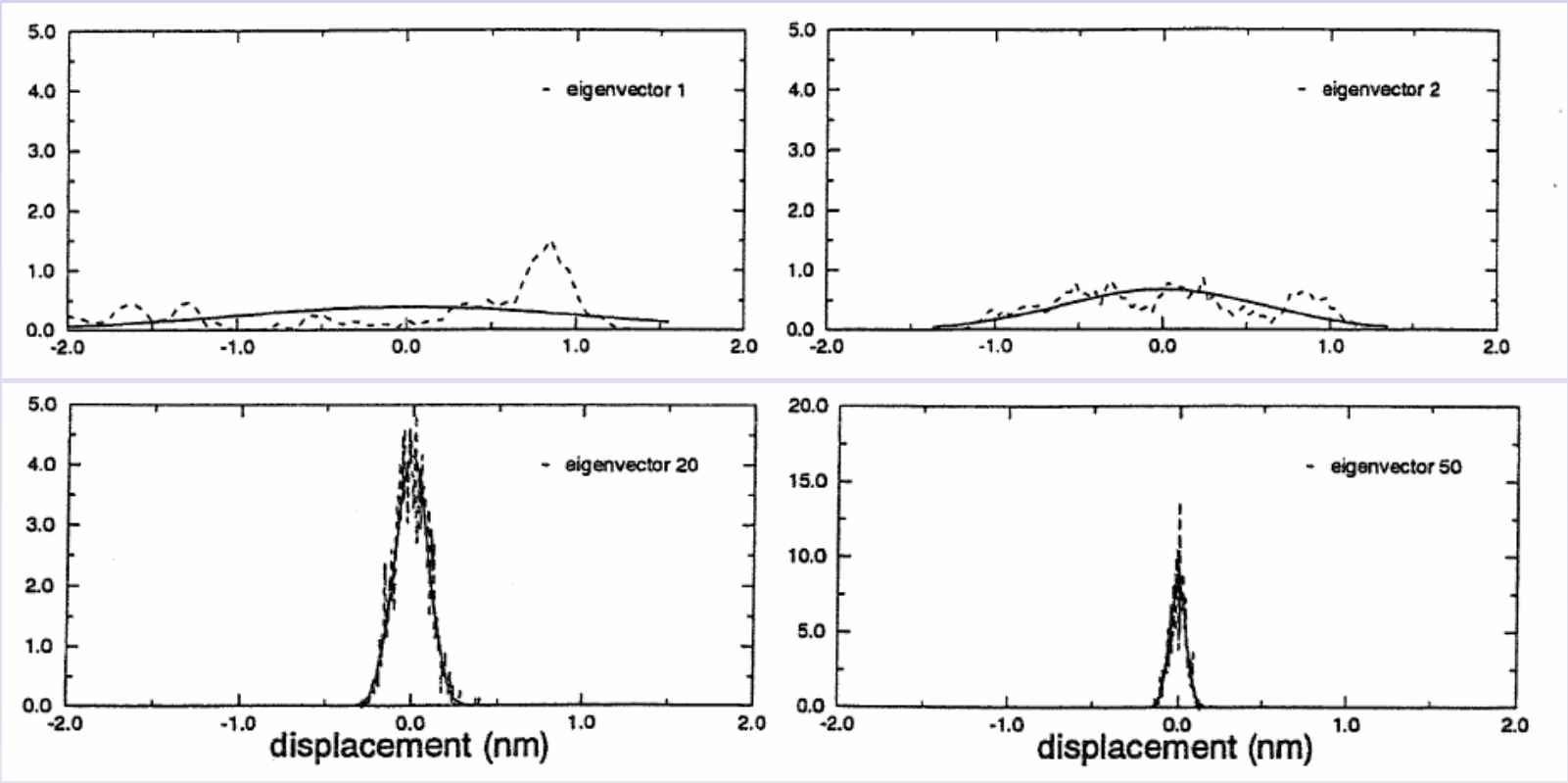
Komplexný príklad - MD trajektórie

čas vs. fluktuácie v pohybe (displacement)



Komplexný príklad - MD trajektórie

pohyb vs. pravdepodobnostné rozloženie v čase



Multidimensional scaling

- dáta s nelineárnymi vzťahmi medzi premennými
- zachováva vzdialenosti medzi bodmi
- na vstupe je matica skutočných vzdialeností medzi bodmi v originálnom priestore
- na výstupe je matica vzdialeností medzi bodmi v redukovanom priestore
- algoritmus minimalizuje stresovú maticu, ktorá reprezentuje rozdiel medzi vstupnou a výstupnou maticou

Isomap

- rozšírenie MDS
- poznáme len vzdialenosti medzi susednými bodmi
- ostatné sa dopočítajú ako geodetická vzdialenosť
- euklidovská: vzdialenosť vzdušnou čiarou
- geodetická: dĺžka najkratšej cesty

Isomap algoritmus

- nájdi susedov pre každý bod (K-najbližších, pevný polomer)
- vytvor graf susedov (spojené uzly sú susedia, dĺžka hrany je euklidovská vzdialenosť medzi nimi)
- pomocou algoritmu Floyd-Warshall spočítaj geodetické vzdialenosti medzi všetkými bodmi
- pomocou MDS spočítaj projekcie do redukovaných dimenzií

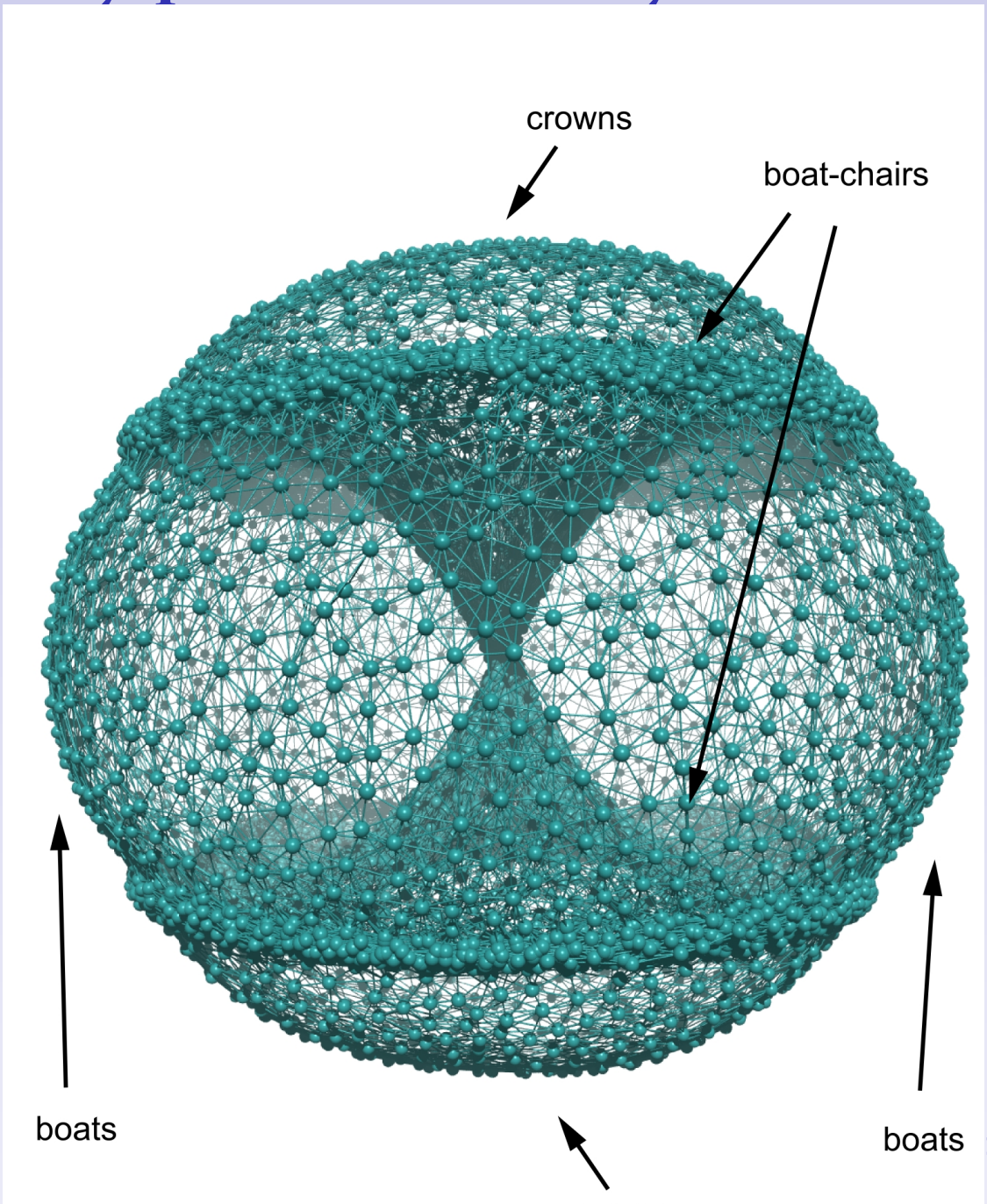
Jednoduchý príklad

S-krivka

Komplexný príklad - Metadynamika

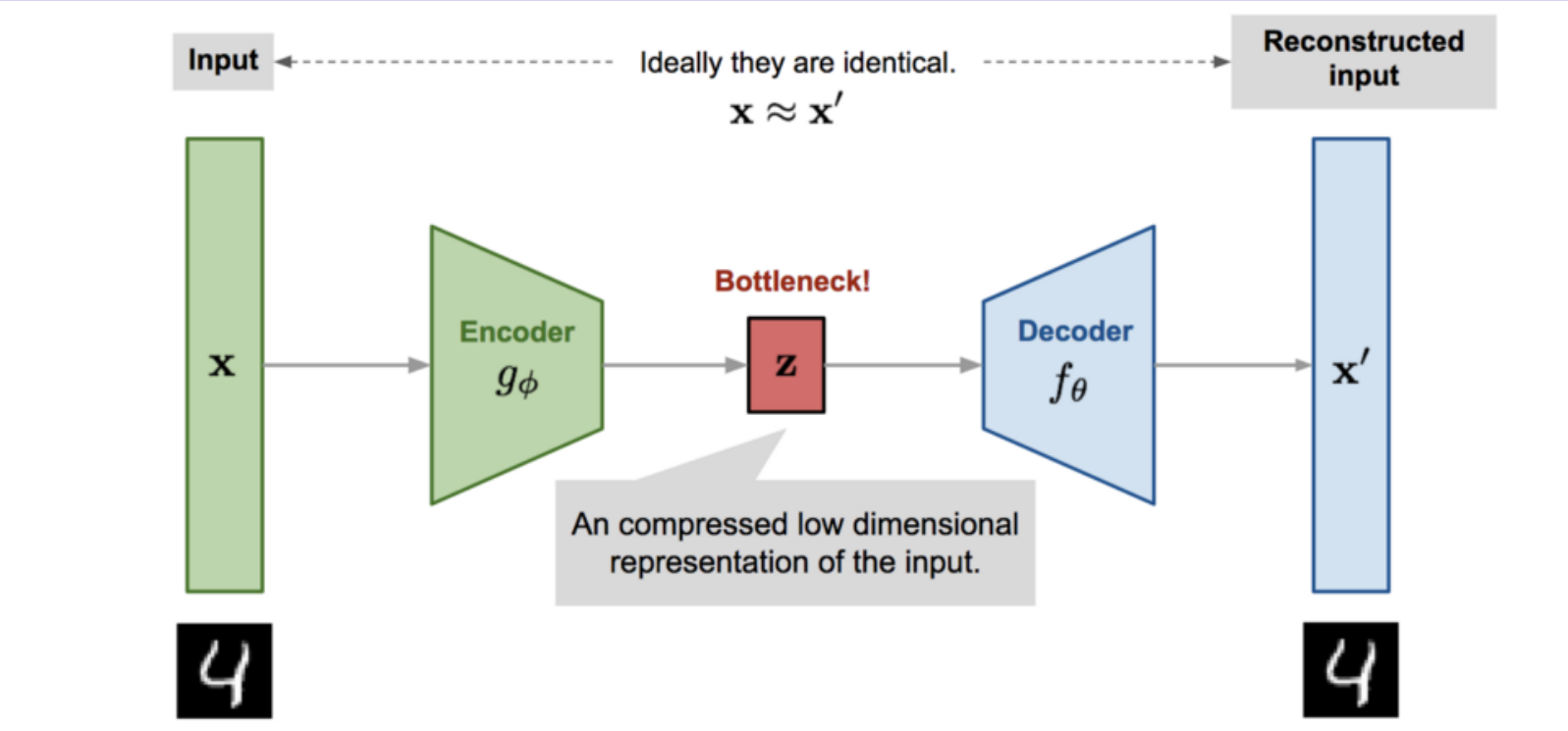
- molekulová dynamika, kde molekuly pošťuchujeme správnym smerom
- ktorý smer je ten správny?
- supermarket, vozíky a smrad
- v priestore s veľkým počtom dimenzií ľahko smradu utečiem
- redukcie dimenzionality s cieľom nájsť správny smer pohybu

Komplexný príklad - Metadynamika



Variačné enkódery

4



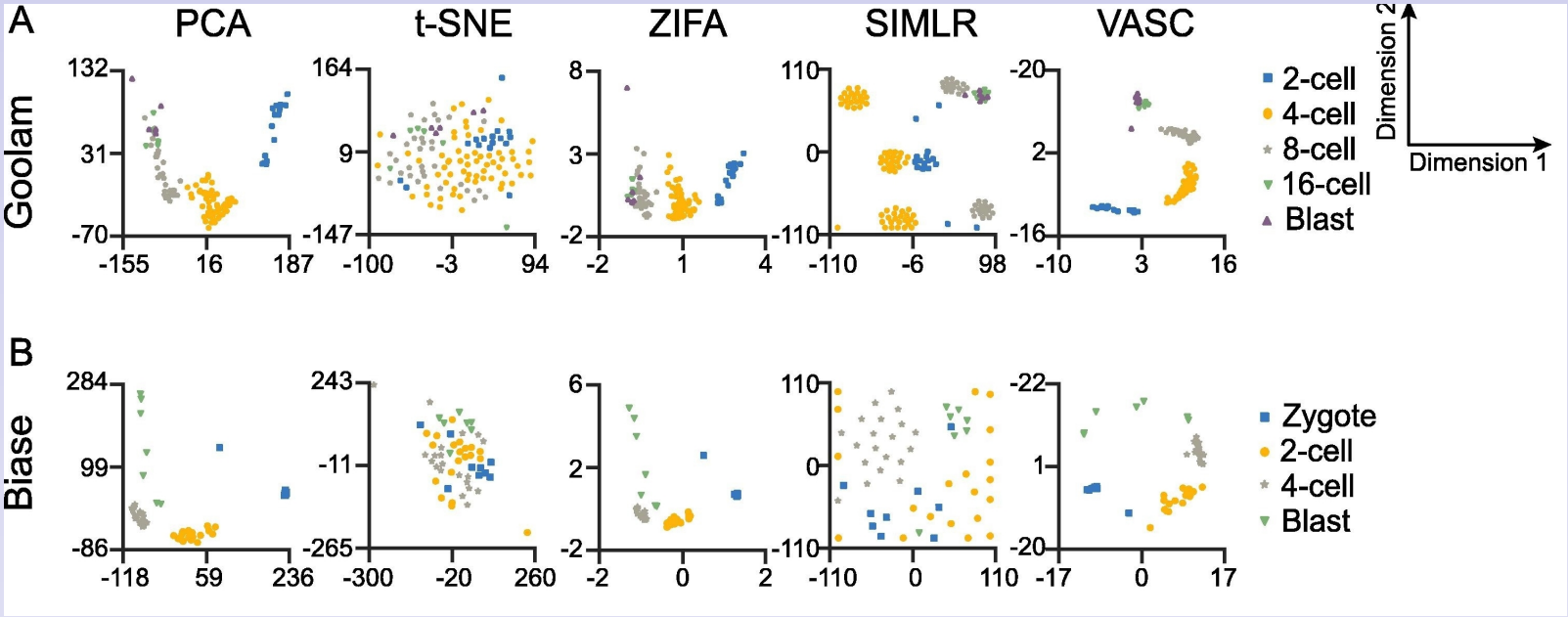
[4] https://towardsdatascience.com/dimensionality-reduction-for-machine-learning-80a46c2ebb7e

Komplexný príklad - RNA sekvencie

- RNA sekvencie vždy z jedinej bunky: veľa šumu
- sekvenciu genómu: čo by bunka mohla robiť
- expresie genómu: ako aktívny je každý gén, čo bunka v daný moment robí
- rôzne typy buniek (neurón, koža, pečeň) majú iné expresie
- v rôznych štádiách vývoja má bunka iné expresie genómu
- chceme vedieť rôzne typy/štádia/stavy bunky rozlišovať
- potrebujeme redukovať veľké množstvo dimenzií (tisíce) na málo typov buniek (jednotky)
- a ideálne to vizualizovať [5]

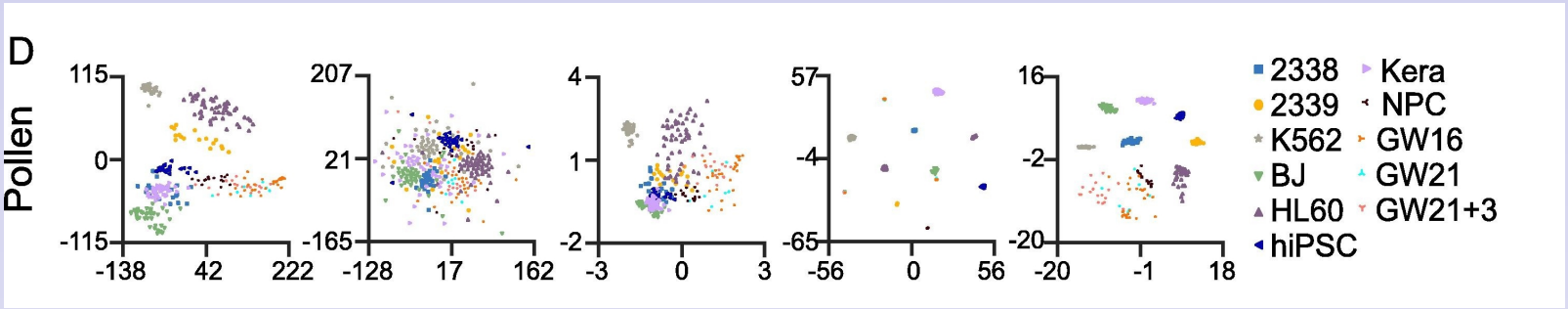
[5] Wang, D., & Gu, J. (2018). VASC: Dimension Reduction and Visualization of Single-cell RNA-seq Data by Deep Variational Autoencoder. Genomics, Proteomics & Bioinformatics, 16(5), 320–331. https://doi.org/10.1016/J.GPB.2018.08.003

Komplexný príklad - RNA sekvencie



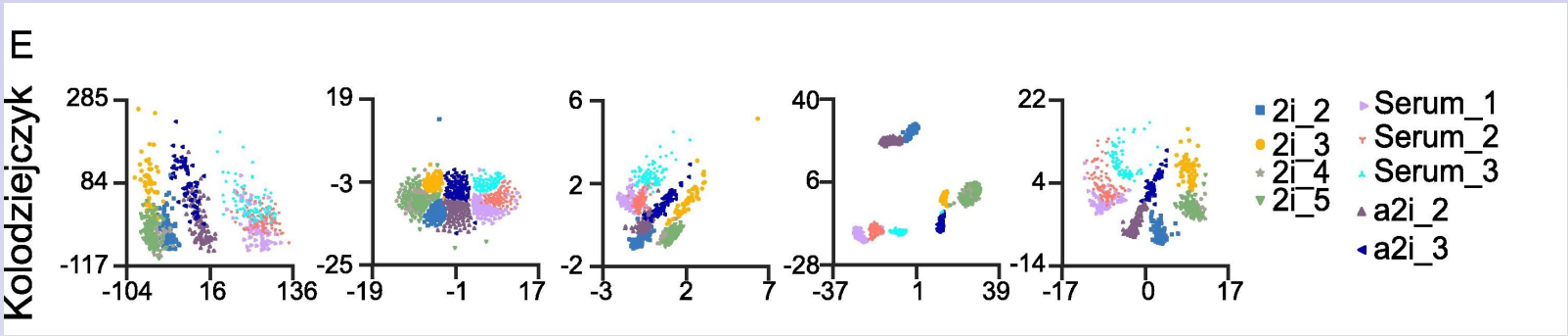
- bunky rôzneho štádia vývinu
- PCA, ZIFA, VASC ok
- t-SNE, SIMLR nok

Komplexný príklad - RNA sekvencie



- 11 rôznych druhov buniek
- PCA, ZIFA nok
- SIMLR klustery obsahujú aj bunky iného typu
- VASC osem klusterov, neurónové bunky (GW*) nerozlíšené

Komplexný príklad - RNA sekvencie



- kmeňové bunky v troch rôznych prostrediach (serum, *2i)
- tri rôzne zmesi
- PCA odlíšila prostredia ale nerozlíšila zmesi
- ZIFA ok, ale zmiešala 2i_2 s a2i
- SIMLR ok, ale zmiešala niektoré 2i a a2i
- t-SNE, VASC ok