Seismický polygon Frenštát: Čištění dat a aplikace pro strojové učení

Michael Skotnica, Marek Pecha, Jana Rušajová, Bohuslav Růžek, Vít Wandrol

Motivace

Severovýchod ČR – seismicky aktivní region

- přirozená zemětřesení
- důlní indukované seismické jevy
- exploze

Cíle:

- automatické lokace
- automatické rozpoznávání typů zemětřesení
- postupné rozšíření na Českou regionální seismickou síť

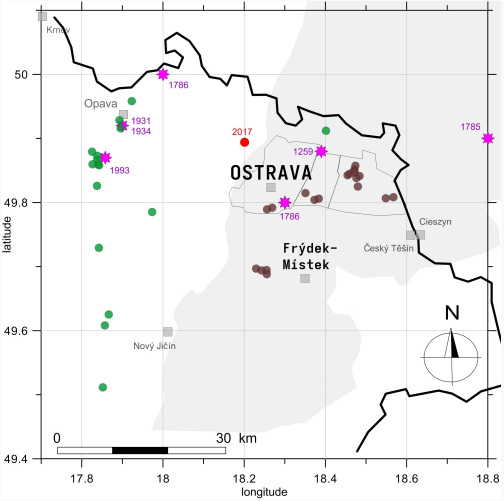


Figure 1: Mapa nejsilnějších zemětřesení.

Seismické stanice – SPF

Seismický polygon Frenštát (1992 – 2002)

- Čeladná, Palkovice, Pstruží, Trojanovice, Vyšní Lhoty
- triggerovaná data (STA/LTA), 125 Hz
- 17 462 jevů se záznamy na 5 stanicích (= 87 310 záznamů)
- obsahuje cenná data (mimo jiné přirozená lokální zemětřesení)
- data je potřeba vyčistit
  - deterministické metody
  - strojové učení bez učitele

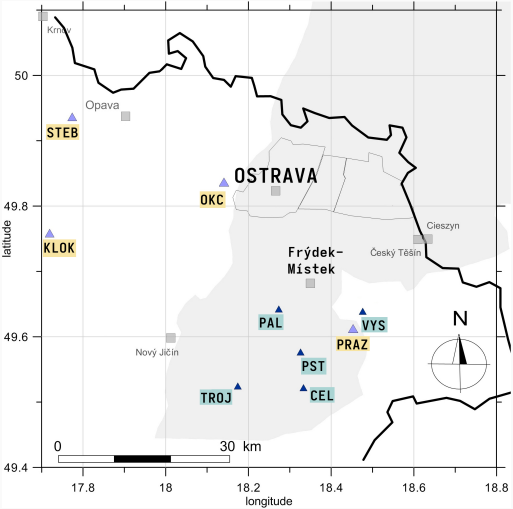


Figure 2: Mapa seismických stanic.

Seismický záznam

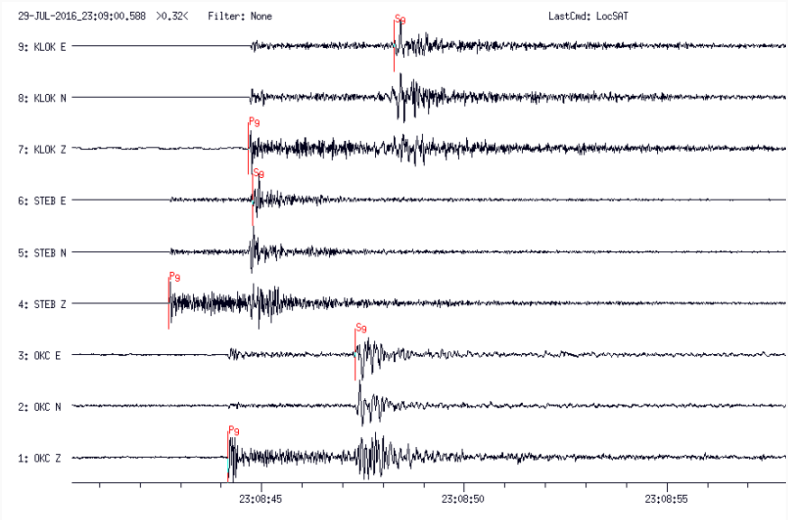


Figure 3: Záznam přirozeného zemětřesení na Opavsku 29. 7. 2016.

Seismický polygon Frenštát – problémová data

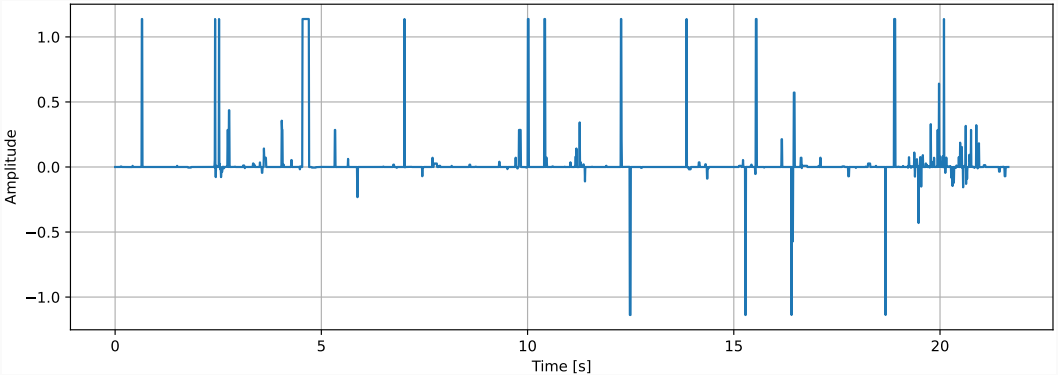


Figure 4: Poškozený záznam.

Seismický polygon Frenštát – problémová data

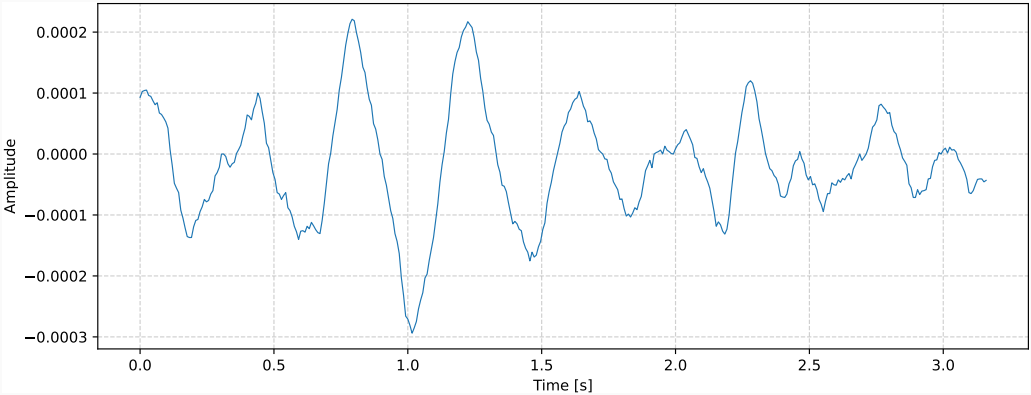


Figure 5: Fragment vzdáleného zemětřesení.

Čištění dat – deterministické metody

Signál dosáhne maximální amplitudy

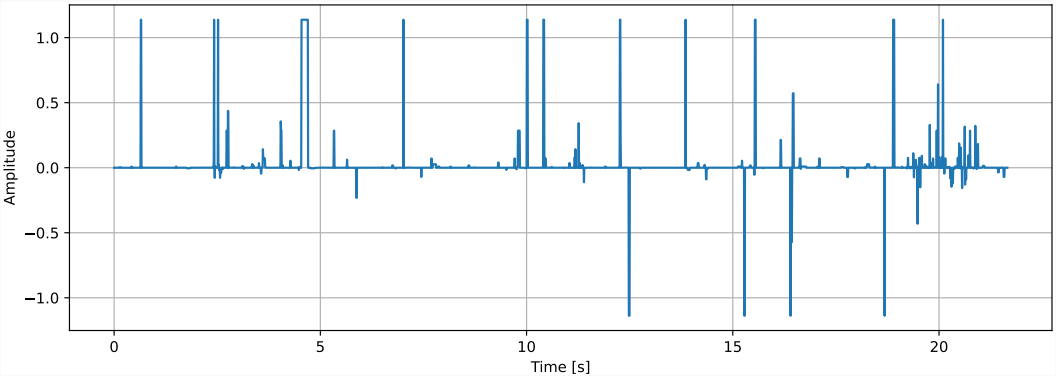


Figure 6: Poškozený záznam.

Čištění dat – deterministické metody

Víc než 30 % hodnot je nad $0.85\times$ maximální hodnoty.

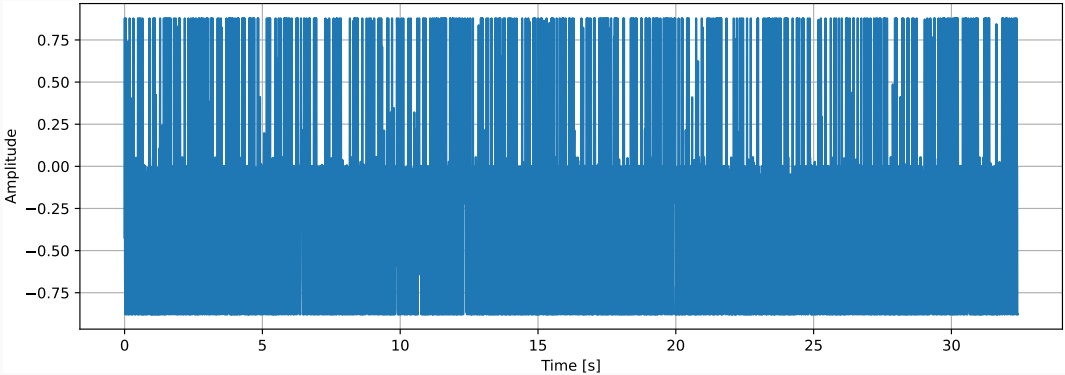


Figure 7: Poškozený záznam.

Čištění dat – deterministické metody

Víc než 95 % hodnot je pod průměrem (v absolutní hodnotě).

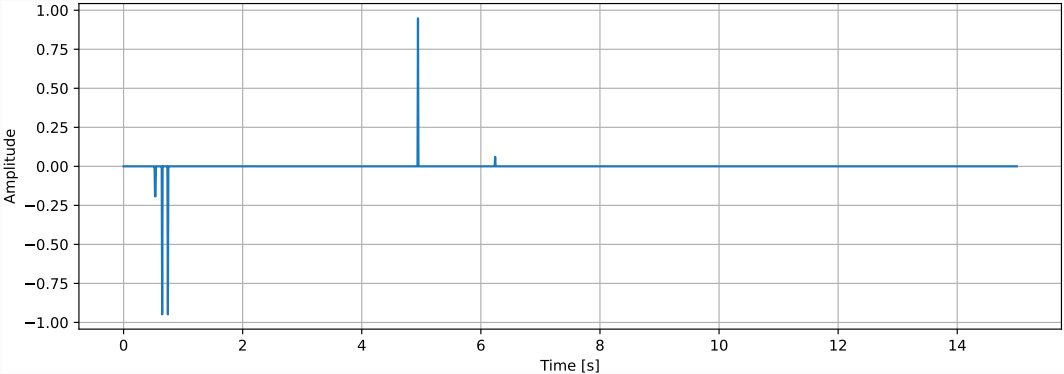


Figure 8: Poškozený záznam.

Čištění dat – deterministické metody

Víc než 40 % hodnot je 0.

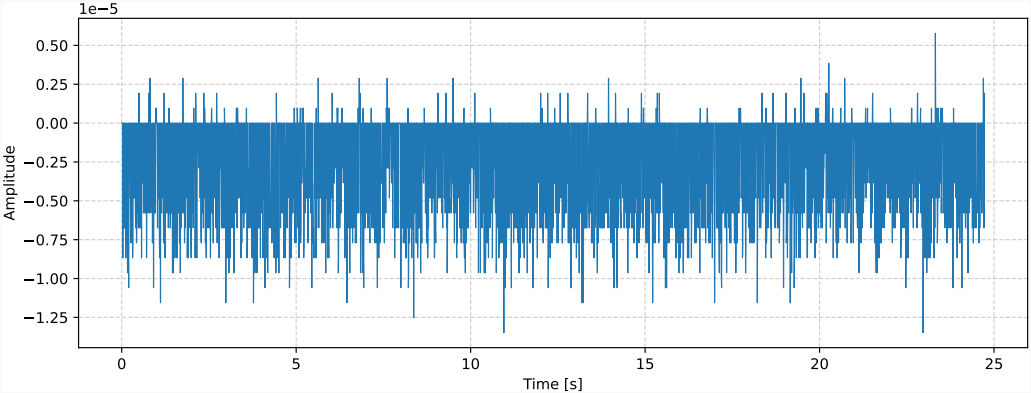


Figure 9: Poškozený záznam.

Předzpracování dat

Pro každou ze tří složek zvlášť:

- Standardizované score
  - lineární transformace signálu
  - $z = \frac{x-\mu}{\sigma}$
- **Zarovnání záznamu na fixní délku n**
  - oříznutí (záznam je delší)
  - doplnění 0 (záznam je kratší)
- Fourierova transformace

Pro každý záznam jevu tedy máme:

- 3 vektory z $\mathbb{R}^n$,
- nebo vektor z $\mathbb{R}^{3\times n}$

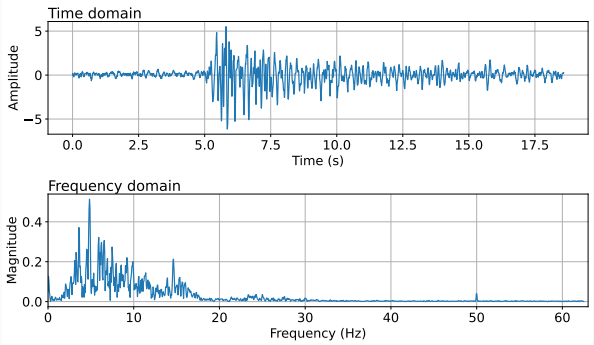


Figure 10: Fourierova transformace jedné složky záznamu důlního indukovaného jevu.

Čištění dat – Strojové učení bez učitele

Na rozdíl od strojového učení s učitelem:

- nemáme ohodnocená trénovací data
- nebo máme příklady jen pro pozitivní třídu.

Čištění dat – K-means clustering

K-means

- Chceme rozdělit data ($x_1, x_2, \ldots, x_N \in \mathbb{R}^D$) do K tříd.

$$z_{i,t} = \begin{cases} 1 & \text{pokud } x_i \text{ je ve třídě } t \\ 0 & \text{jinak} \end{cases}$$

- Pro každou třídu chceme její střed.
  - $\mu_1, \mu_2, \ldots, \mu_K$
- Minimalizujeme součet čtverců vzdáleností bodů od středů tříd, do nichž náleží.

$$\sum_{i=1}^{N} \sum_{t=1}^{K} z_{i,t} ||x_i - \mu_t||^2$$

- V našem případě $K = 2$ a menší třída reprezentuje chyby.

Čištění dat – K-means clustering

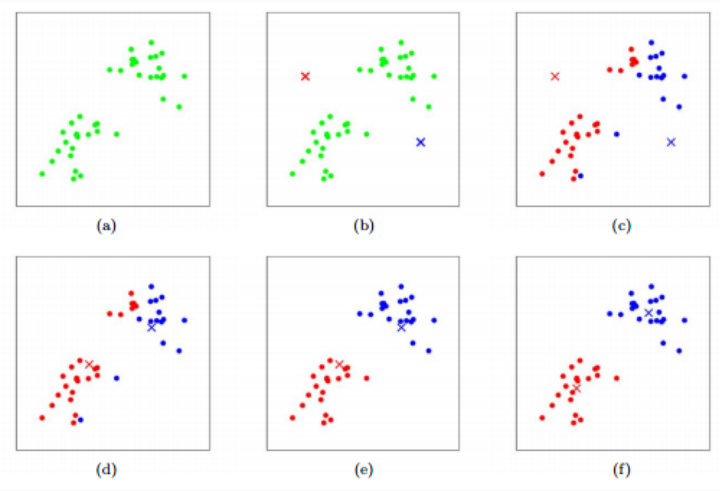


Figure 11: Zdroj:
https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

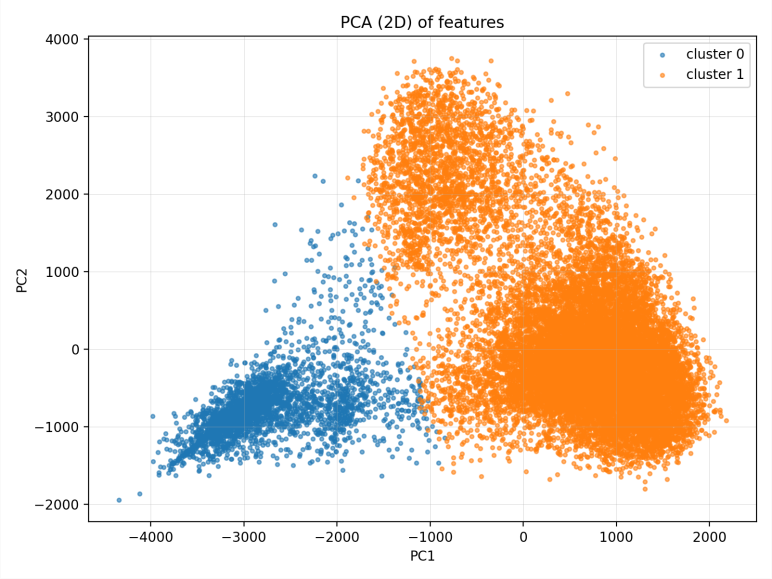
Čištění dat – K-means clustering (volba středů)

Jak volit v prvním kroku K středů? (K-means++)

- První střed náhodně z bodů.
- i-tý náhodně z bodů tak, že pravděpodobnost zvolení bodu je úměrná vzdálenosti k nejbližšímu z $i-1$ již zvolených středů

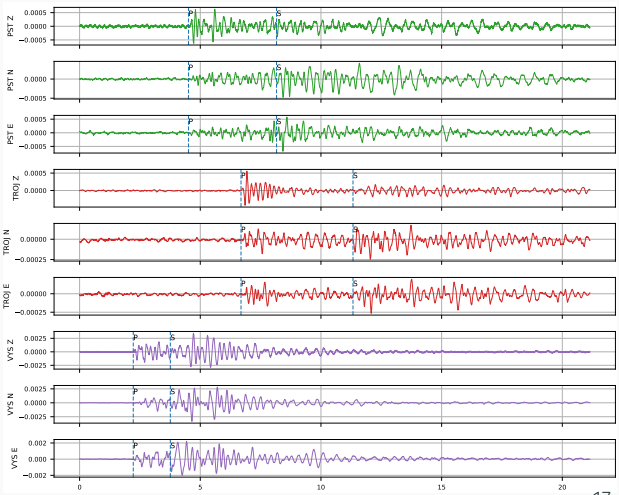
Algoritmus se obvykle pustí vícekrát (např. $10\times$) a vybere se rozdělení s nejmenší hodnotou účelové funkce.

Čištění dat – K-means clustering



Čištění dat s využitím příkladů kvalitních dat

- Pro 4 568 záznamů máme k dispozici časy nasazení P-vlny
- Pokud není čas nasazení u jevu na všech 5 stanicích, zbytek bude pravděpodobně poškozený/neúplný
- Dají se kvalitní záznamy využít i jinak?



Neuronová síť – připomenutí

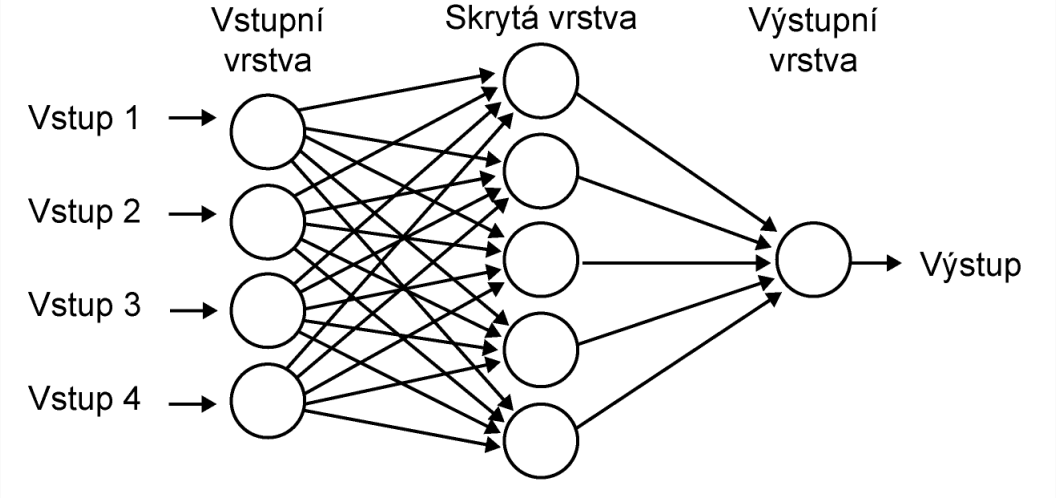


Figure 12: Zdroj: https://portal.matematickabiologie.cz

Čištění dat – Autoenkodér

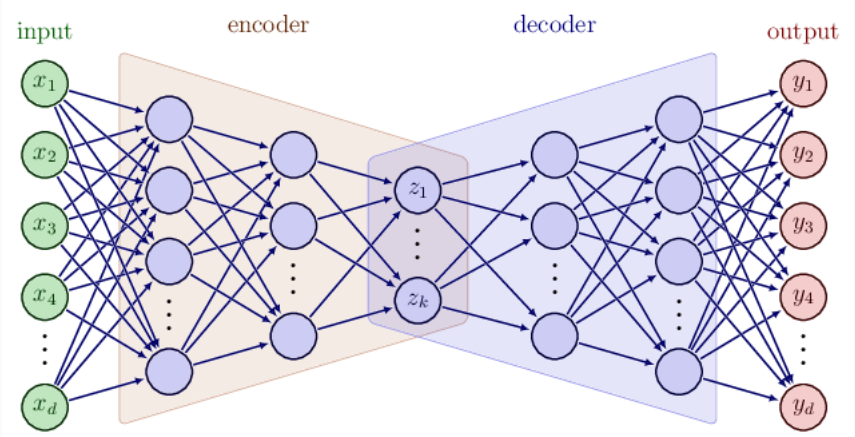


Figure 13: Autoři: Wei Zhang, Christof Schütte,
zdroj: DOI: 10.48550/arXiv.2307.00365, licence: CC BY 4.0.

Čištění dat – Autoenkodér

Myšlenka:

- Natrénujeme autoenkodér pouze na hezkých datech.
- Autoenkodér se učí rekonstruovat vstupní signál na výstupu.
- Při aplikaci na nová data bude rekonstrukční chyba:
  - malá pro hezká (normální) data,
  - výrazně větší pro poškozená (anomální) data.
- Jako míru chyby lze použít např.:
  - Mean Absolute Error (MAE),
  - Mean Squared Error (MSE)

Jak nastavit threshold pro chybu?

- Zvolíme tak, aby např. 99 % validačních hezkých dat mělo rekonstrukční chybu menší než threshold.

Čištění dat – Autoenkodér

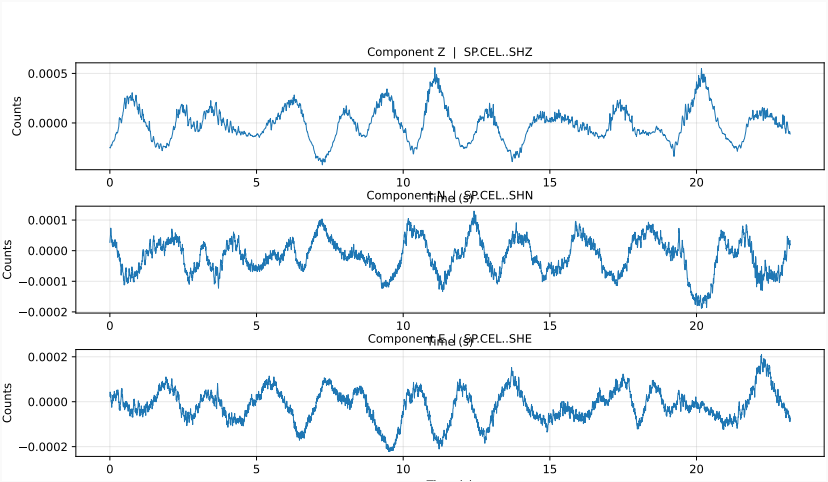


Figure 14: Příklad špatného záznamu detekovaného MLP autoenkodérem ($\rightarrow 512 \rightarrow 128 \rightarrow \mathbf{32} \rightarrow 128 \rightarrow 512 \rightarrow$).

Principal component analysis (PCA) – redukce dimenze

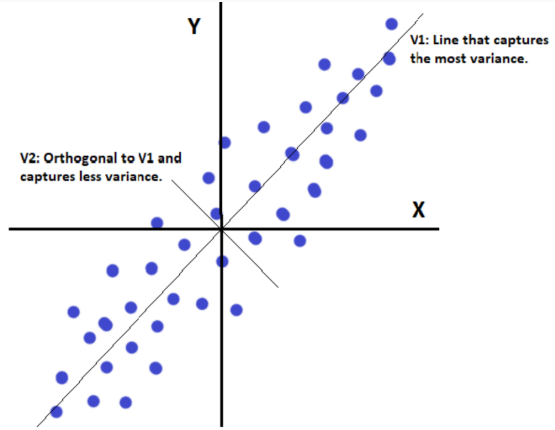


Figure 15: Zdroj: https://statisticsbyjim.com/basics/principal-component-analysis/