



TECHNICKÁ UNIVERZITA V LIBERCI  
Fakulta mechatroniky, informatiky  
a mezioborových studií ■

# Doporučovací systémy

*František Kynych*  
16. 12. 2021 | MVD





TECHNICKÁ UNIVERZITA V LIBERCI  
Fakulta mechatroniky, informatiky  
a mezioborových studií ■

# Část I.: Úvod do problematiky



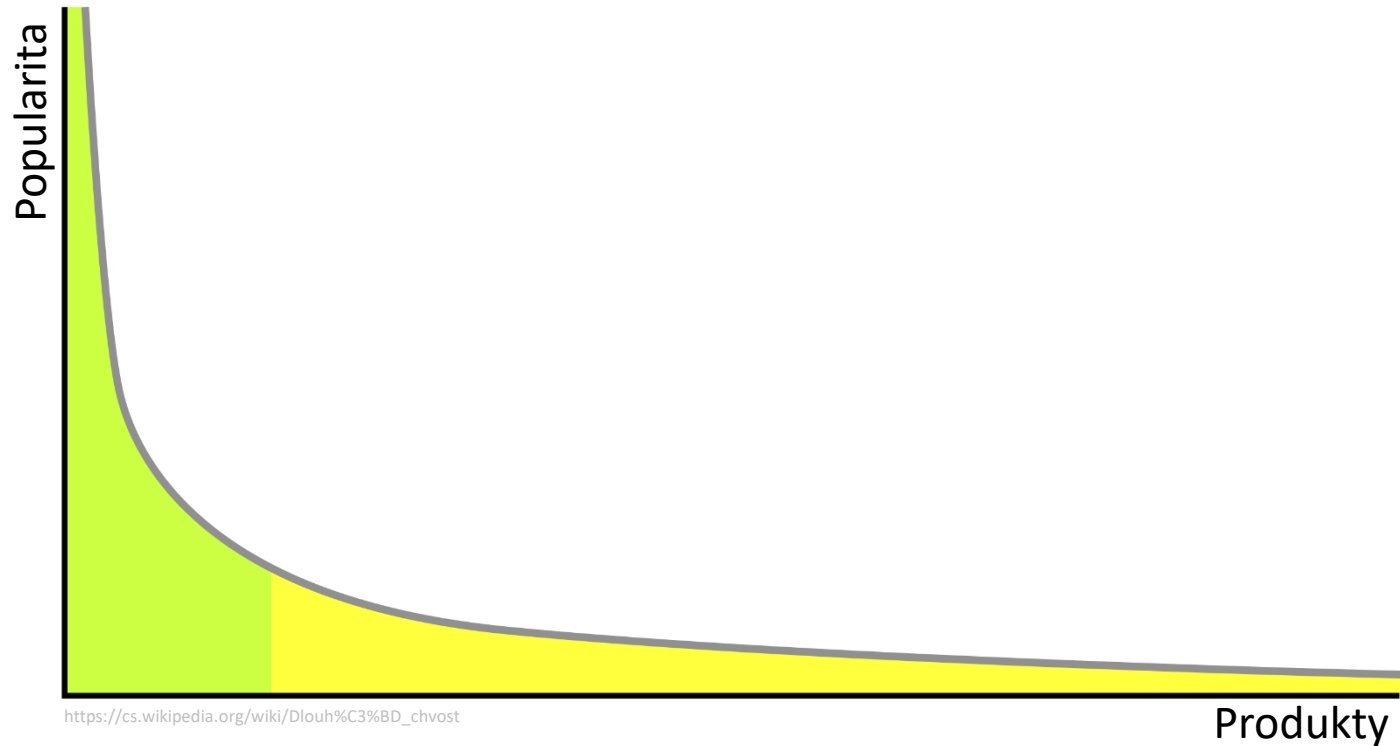
# Doporučovací systémy

- Systém pro filtrování informací
- Snaží se predikovat hodnocení nebo preferenci uživatele
- Přirozený přechod od globálně populárních produktů k více personalizovaným doporučením
  - Založení na popularitě nebylo dostatečné
    - U některých služeb se ale jedná o nejlepší způsob doporučení
    - Nejprodávanější, Nejlépe hodnocené, ...

# Doporučovací systémy

- Využíváno téměř každou službou
  - Google
  - YouTube
  - Facebook
  - Twitter
  - Spotify
  - Netflix
  - E-shopy
  - Telekomunikace

# Dlouhý chvost (Long Tail)



# Doporučovací systémy

- Dělení:
  - Unimodální
    - Doporučují na základě jednoho druhu interakce s uživatelem
    - V daných intervalech se přepočítává doporučení
    - **Collaborative, content-based filtering**
  - Multimodální (hybridní)
    - Doporučuje na základě více interakcí s uživatelem
    - Používá více modelů a kombinuje je dohromady
    - Poskytování doporučení v reálném čase



## Část II.: Content-based filtering

# Content-based filtering

- Intuice
  - Doporučení podobných produktů, které byly od daného uživatele kladně hodnoceny
- Například:
  - Filmy
    - Stejní herci, žánr, ...
  - Webové stránky
    - Články s podobným obsahem
  - Lidé
    - Doporučení na základě velkého množství společných přátel



# Content-based filtering

- Uživatelské profily
  - V první fázi je získat základní preference uživatele
  - Na základě získaných dat můžeme vytvořit uživatelské profily
  - Poté již můžeme doporučovat další produkty
- Profily produktů
  - Set příznaků pro každou položku
  - Např. film -> autor, název, herci, žánr, ...
  - Vektor (boolean / reálné hodnoty)

# Content-based filtering

	Uživatel
Free Guy	4
Interstellar	5
Eternals	2
Wish Dragon	4
Spider-Man	?
...	?

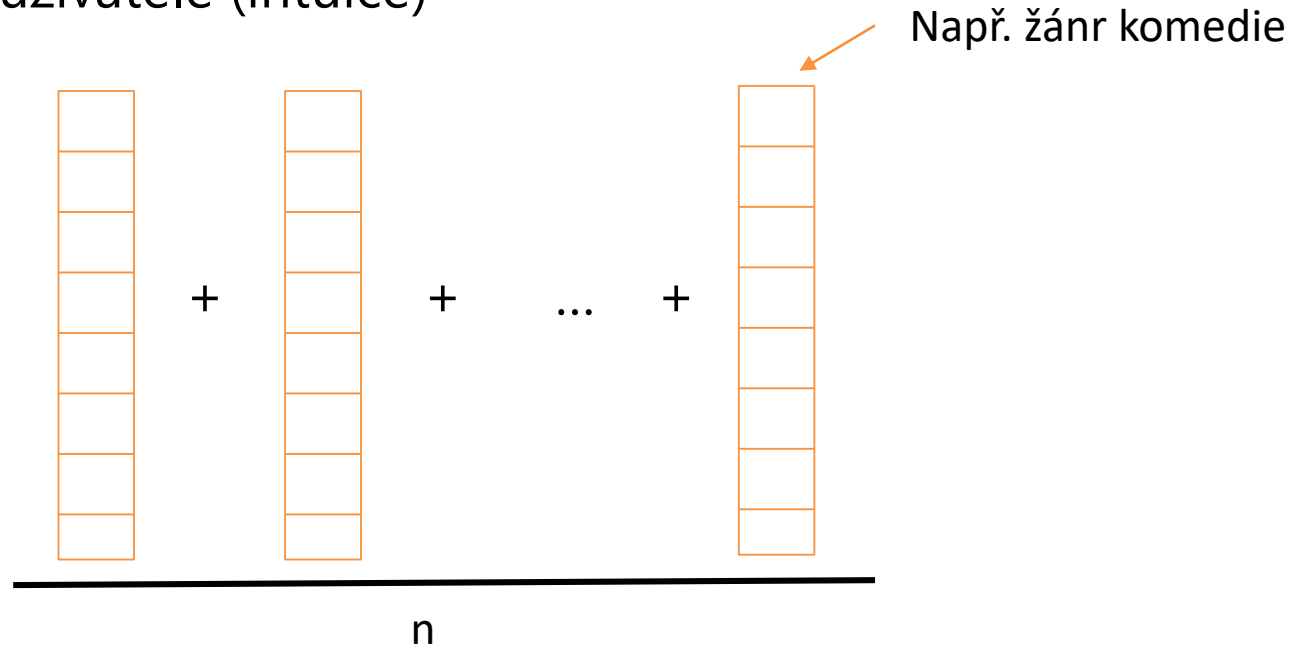
Komedie	Akční	...
0.4	0.5	...
0	0.3	...
0.1	0.8	...
0.7	0.2	...
0.2	0.6	...
...	...	...

$$x_i = [1, 0.4, 0.5, \dots]^T$$

bias

# Content-based filtering

- Získání profilu uživatele (intuice)



- Využíváme hodnot od hodnocených produktů daným uživatelem

# Content-based filtering

- Získání profilu uživatele s využitím preferencí uživatele

$$w_1 \begin{bmatrix} \square \\ \square \\ \square \\ \square \\ \square \\ \square \\ \square \\ \square \\ \square \end{bmatrix} + w_2 \begin{bmatrix} \square \\ \square \\ \square \\ \square \\ \square \\ \square \\ \square \\ \square \\ \square \end{bmatrix} + \dots + w_n \begin{bmatrix} \square \\ \square \\ \square \\ \square \\ \square \\ \square \\ \square \\ \square \\ \square \end{bmatrix}$$

# Content-based filtering

-> úloha regrese (User-centred linear regression)

- Nalezení vah  $W$  pro jednoho uživatele:

$$W = \arg \min_W \frac{1}{2} \sum_{i \in \text{rated\_products}} (W^T x_i - y_i)^2$$

Hodnocení uživatele

- Získání predikce pro neohodnocený film  $i$ :

$$\hat{y}_i = W^T x_i$$

# Content-based filtering

- **Výhody**
  - Nejsou potřeba data ostatních uživatelů
  - Doporučíme přesně to, co má uživatel rád
  - Můžeme doporučit i nové nebo obecně nepopulární produkty
  - Lehce zjistíme, proč byla nějaká položka doporučena
- **Nevýhody**
  - Těžké hledání vhodných příznaků (např. obrázky)
  - Příliš se zaměříme na jeho první hodnocené položky
    - Uživatel může mít více zájmů
  - Cold-start problém u nových uživatelů



# Část III.: Collaborative filtering

# Collaborative filtering

- Intuice
  - K uživateli X nalezneme skupinu N uživatelů, kteří hodnotí produkty podobně
  - Predikujeme hodnocení uživatele X na základě hodnocení skupiny N uživatelů
    - Nalezneme produkty, které uživatel ještě nehodnotil, ale skupina ano
    - Na základě predikce doporučíme uživateli produkty, které by se mu mohly líbit
- Matice interakcí
  - Záznam minulých interakcí uživatele s produkty (např. hodnocení)





# Collaborative filtering


- Memory based
  - Založeno na podobnosti uživatelů nebo předmětů v matici interakcí
- Model based
  - Vytvoření modelu pro user-item nebo item-item interakce
  - Reprezentace uživatele a produktů naučena z matice interakcí

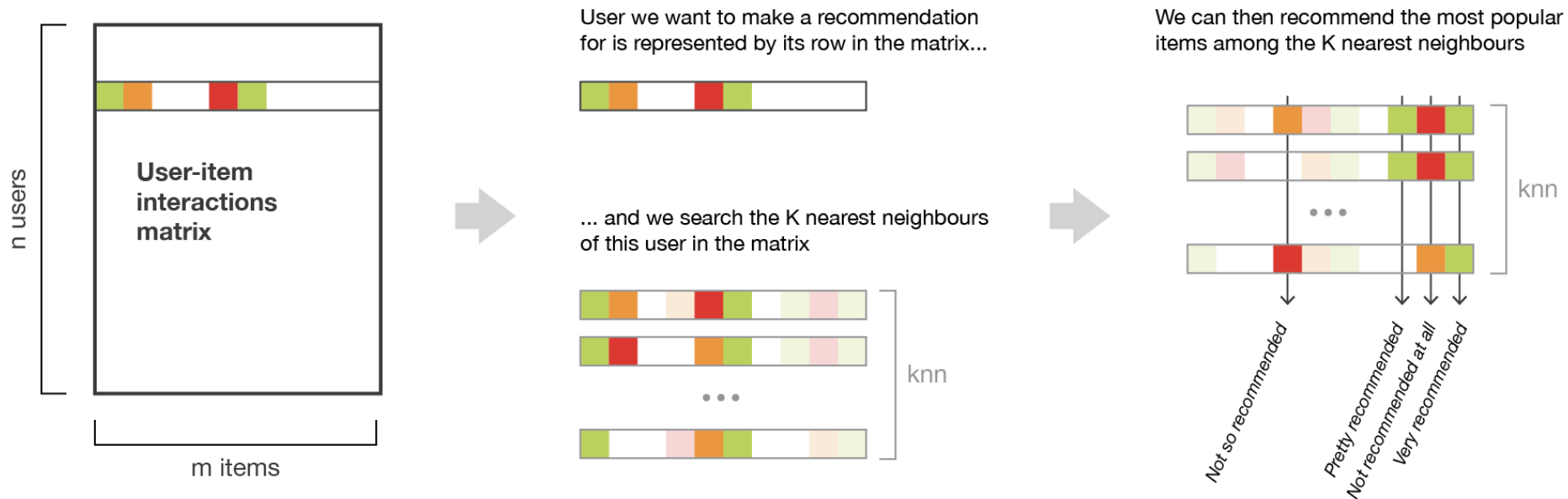
# Collaborative filtering – memory based

- User-user

 positive interactions

 neutral interactions

 negative interactions



<https://towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada>

# Collaborative filtering – memory based

- Jak měřit vzdálenost uživatelů?
  - **Možnost 1: Kosinová podobnost**

	LOTR 1	LOTR 2	LOTR 3	SW 1	SW 2	SW 3
User A	5	5	4	0	0	0
User B	5	0	0	4	3	0
User C	2	0	0	0	5	5

- Neohodnocené filmy vyplníme nulami
- $\text{sim}(A, B) = 0.44$ ,  $\text{sim}(A, C) = 0.17$ ,  $\text{sim}(B, C) = 0.48$
- Problém
  - Neohodnocené filmy  $\rightarrow 0$  (= použité nejhorší hodnocení)

# Collaborative filtering – memory based

- Jak měřit vzdálenost uživatelů?
  - Možnost 2: Centered Cosine similarity**

	LOTR 1	LOTR 2	LOTR 3	SW 1	SW 2	SW 3	$\overline{user}$
User A	5	5	4				14/3
User B	5			4	3		12/3
User C	2				4	5	11/3

- Normalizace odečtením průměrného hodnocení uživatele

	LOTR 1	LOTR 2	LOTR 3	SW 1	SW 2	SW 3
User A	1/3	1/3	-2/3			
User B	1			0	-1	
User C	-5/3				1/3	4/3

# Collaborative filtering – memory based

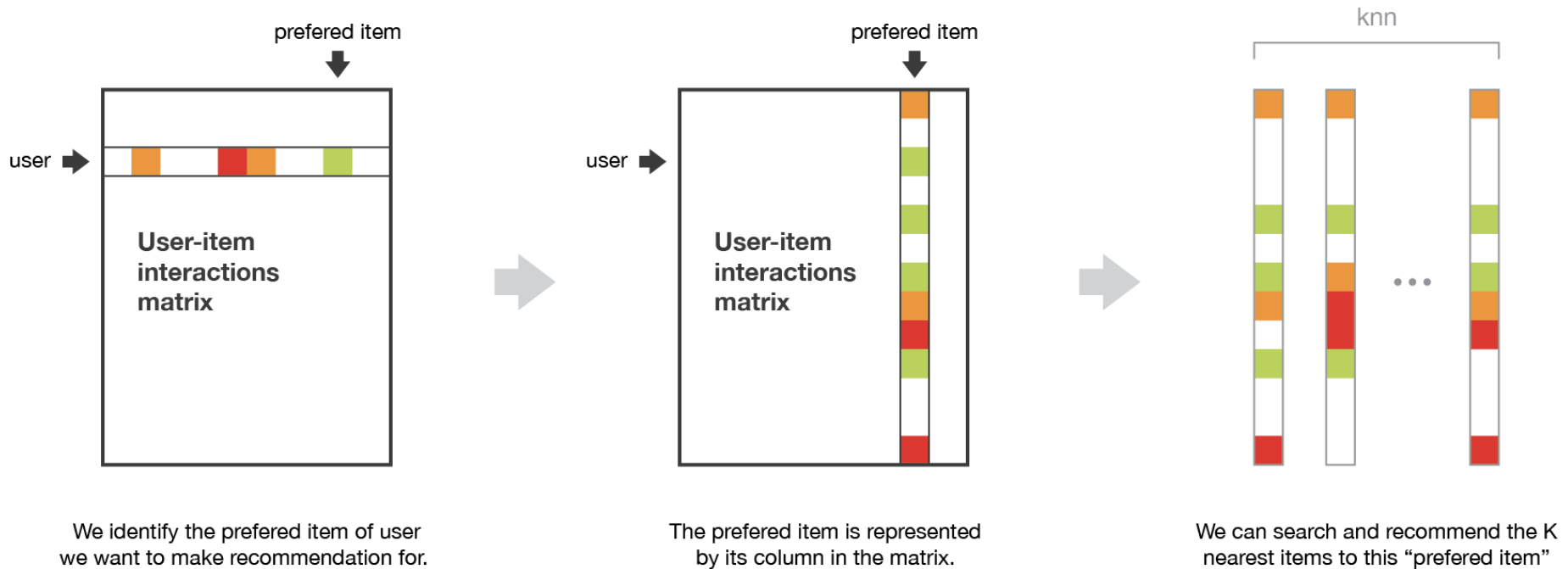
- Jak měřit vzdálenost uživatelů?
  - **Možnost 2: Centered Cosine similarity**
    - Normalizace odečtením průměrného hodnocení uživatele

	LOTR 1	LOTR 2	LOTR 3	SW 1	SW 2	SW 3
User A	1/3	1/3	-2/3	0	0	0
User B	1	0	0	0	-1	0
User C	-5/3	0	0	0	1/3	4/3

- Nula nyní označuje průměrné hodnocení
- $\text{sim}(A, B) = 0.28$ ,  $\text{sim}(A, C) = -0.31$ ,  $\text{sim}(B, C) = -0.65$
- Také známé pod názvem **Pearson Correlation**

# Collaborative filtering – memory based

- Item-item



<https://towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada>

# Collaborative filtering (item-item)

- Jak vypočítat hodnocení produktu **i** uživatelem **x**?
  - Nalezneme **n** nejpodobnějších produktů k produktu **i**
  - Výpočet hodnocení **r<sub>xi</sub>**:

$$r_{xi} = \frac{\sum_{j \in N(i;x)} s_{ij} r_{xj}}{\sum_{j \in N(i;x)} s_{ij}}$$

$s_{ij}$  ... podobnost produktu **i** a **j**

$r_{xj}$  ... hodnocení produktu **j** uživatelem **x**

$N(i;x)$  ... množina produktů podobných produktu **i**  
hodnocených uživatelem **x**

# Collaborative filtering (item-item)

- Pro 2 nejbližší sousedy

		uživatelé												sim(1,m)
produkty		1	2	3	4	<u>5</u>	6	7	8	9	10	11	12	
	1	1		3		?	5			5		4		1.00
	2			5	4			4			2	1	3	-0.18
	<u>3</u>	2	4		1	2		3		4	3	5		<u>0.41</u>
	4		2	4		5			4			2		-0.10
	5			4	3	4	2					2	5	-0.31
	<u>6</u>	1		3		3			2			4		<u>0.59</u>

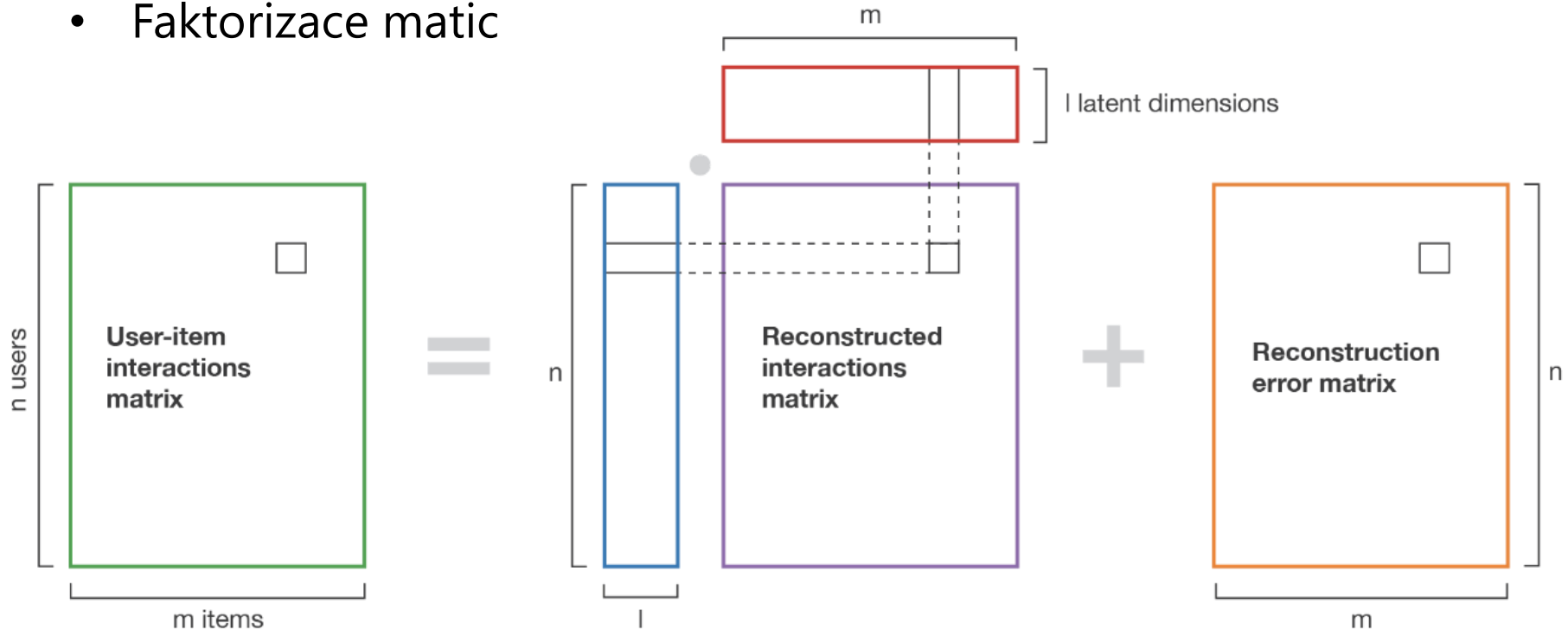
$$r_{15} = \frac{0.41 * 2 + 0.59 * 3}{0.41 + 0.59} = 2.6$$

Pearson correlation + výpočet podobnosti řádků



# Collaborative filtering – model based

- Faktorizace matic



The **user-item interactions matrix** is assumed to be equal to...

... the **dot product** of a **user matrix** and a **transposed item matrix**...

... plus some **reconstruction error**

<https://towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada>

# Collaborative filtering – model based

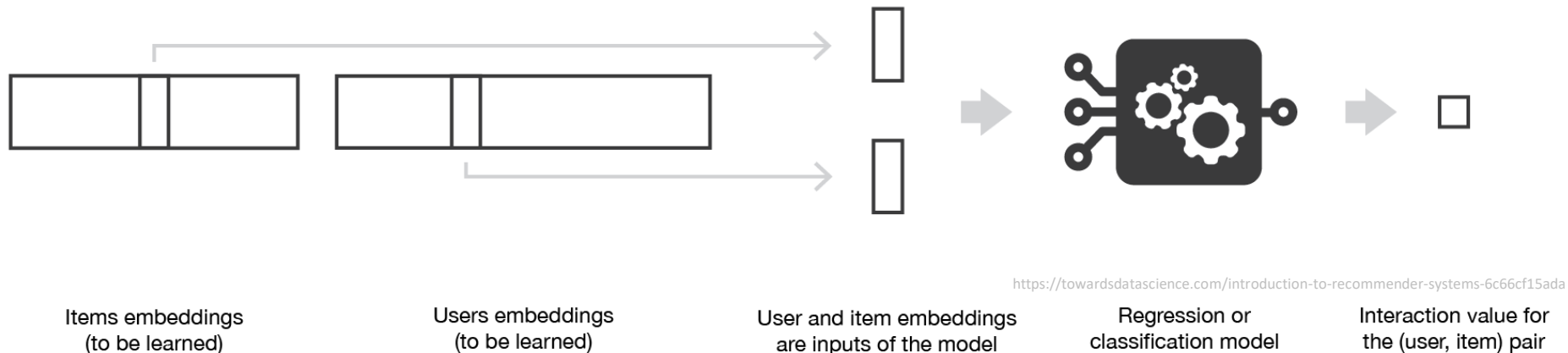
- Podobné jako u content-based filtering
- $Y \approx WTX$ 
  - $W$  ... profily uživatelů,  $X$  ... matice produktů (u filmů označovala kategorie, zde se také učí)
- Iterativní proces

$$W = \arg \min_w \frac{1}{2} \sum_{i,j \in \text{rated\_products}} (W_i^T X_j - Y_{ij})^2$$
$$X = \arg \min_x \frac{1}{2} \sum_{i,j \in \text{rated\_products}} (W_i^T X_j - Y_{ij})^2$$

+ Je vhodné použít L2 regularizaci

# Collaborative filtering – model based

- Je možné i využití jiných metod faktorizace matice (např. SVD)
- Možnosti učení embeddingů pomocí hlubokých neuronových sítí





TECHNICKÁ UNIVERZITA V LIBERCI  
Fakulta mechatroniky, informatiky  
a mezioborových studií ■

# Část IV.: Hybridní systémy



# Hybridní systémy

- Různé přístupy mají své výhody a nevýhody
  - Hybridní systémy se snaží kombinovat jednotlivé přístupy tak, aby byly jejich nevýhody omezeny
- Kombinace systémů:
  - Content-based filtering
  - Collaborative filtering
  - Session-based filtering
  - Demographic filtering
  - Knowledge-Based filtering

# Hybridní systémy

- Přístupy:
  - Weighted
    - Kombinace výstupů různých systémů
  - Switching
    - Přepínání mezi různými systémy
  - Mixed
    - Sjednocení výsledků (listů)
  - Feature Combination
    - Použijeme výstup jednoho systému jako rozšíření vstupu do dalšího systému
  - Feature Augmentation
    - Jeden systém generuje příznaky navíc (např. podobné produkty, autory), které jsou poté použity v dalším systému
  - Cascade
    - Více systémů generuje doporučení za sebou (první vygeneruje širší doporučení a druhý ho více zužuje)
  - Meta-level
    - Vytvoření modelu pro doporučení, který je použit v dalších krocích
    - Např. vytvoření content-based modelu a použití u collaborative modelu



TECHNICKÁ UNIVERZITA V LIBERCI  
Fakulta mechatroniky, informatiky  
a mezioborových studií ■

## Část IV.: Vyhodnocení

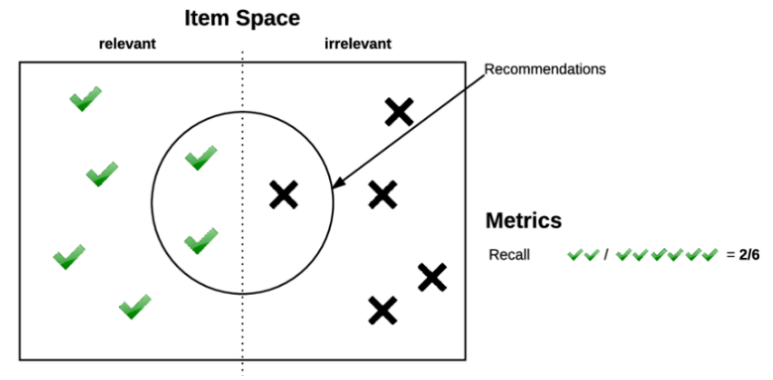
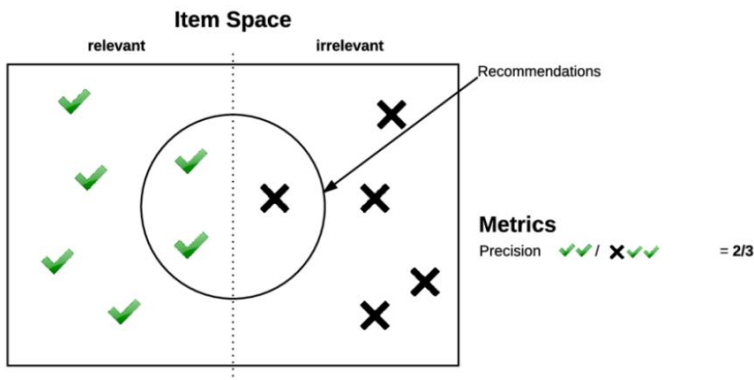


# Vyhodnocení

- Mean Squared Error (MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

- Také se může použít RMSE (Root MSE) nebo Mean Absolute Error
- Precision, Recall





# Vyhodnocení

- Pro seřazené doporučení
  - Normalized discounted cumulative gain (NDCG, viz 4. přednáška - vyhledávání)
  - Average Precision (mean Average Precision)

Correct	Incorrect	Correct	Correct	Incorrect	Incorrect	Correct	Incorrect
Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
1/1	0	2/3	3/4	0	0	4/7	0

Average Precision = 0.427

- Často se používá také mAP@K
  - Omezeno pro K prvních doporučení

# Užitečná literatura / kurzy

- Článek [Hybrid Recommender Systems: A Systematic Literature Review](#)
- Coursera kurz [Machine Learning](#) (9. týden)
- Coursera specializace [Recommender Systems Specialization](#)