Recommendation systems and user representations

Seznam Advertising Systems

SEZNAM.CZ

Agenda

Theoretical part

- About us
- About Seznam
- Introduction to recommender systems
- Recommender system at Seznam
- Recommender system infrastructure
- Ranking models
- Cold start problem

Practical part

- Tools setup and intro
- MIND dataset preparation
- Exploratory data analysis
- Embeddings for user representations
- Train ranking model with user representation

90min **break** **90min**

Practical part

- visit following link: https://tinyurl.com/48se6ze5
- open notebook according to instructions in README.md
- run notebook with prefix 001*

Authors





Václav Blahut

Recommender Systems

Radek Tomšů

Recommender Systems

Tomáš Nováčik

Targeting & model personalization

Adam Jurčík

Targeting & model personalization

Vít Líbal

Relevance in display advertising

Agenda

Theoretical part

- About us
- **About Seznam**
- Introduction to recommender systems
- Recommender system at Seznam
- Recommender system infrastructure
- Ranking models
- Cold start problem

Practical part

- Tools setup and intro
- MIND dataset preparation
- Exploratory data analysis
- Embeddings for user representations
- Train ranking model with user representation

About Seznam.cz

- Technological company and media house
- The product portfolio consists of highly popular online services such:

Seznam.cz Mapy.cz Sreality.cz

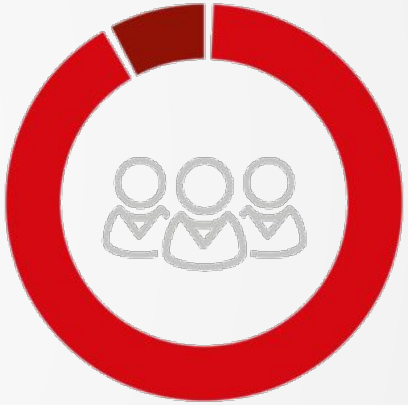
Email Firmy.cz stream

Zboží.cz Sauto.cz Televize Seznam

About Seznam.cz

- **Most visited content websites** on Czech internet
- About **7.3 million unique users** visits Seznam.cz services every month

PROŽENY SUPER.CZ Novinky.cz



ZPRÁVY SPORT.CZ GARÁŽ.CZ

95 %

Monthly reach of the Czech online population.

Agenda

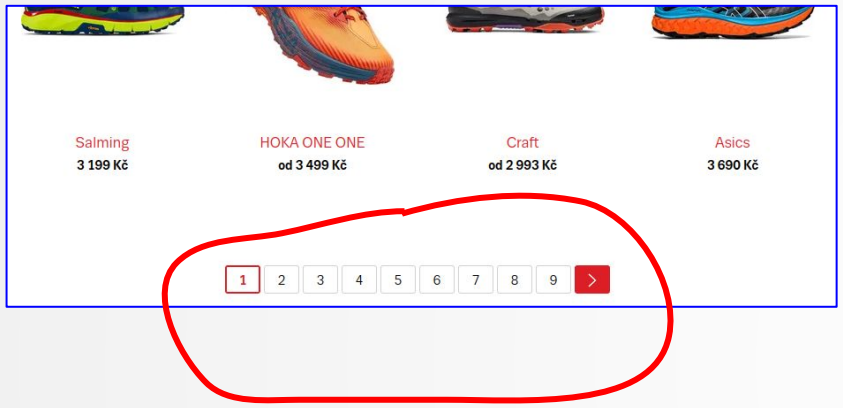
Theoretical part

- About us
- About Seznam
- **Introduction to recommender systems**
- Recommender system at Seznam
- Recommender system infrastructure
- Ranking models
- Cold start problem

Practical part

- Tools setup and intro
- MIND dataset preparation
- Exploratory data analysis
- Embeddings for user representations
- Train ranking model with user representation

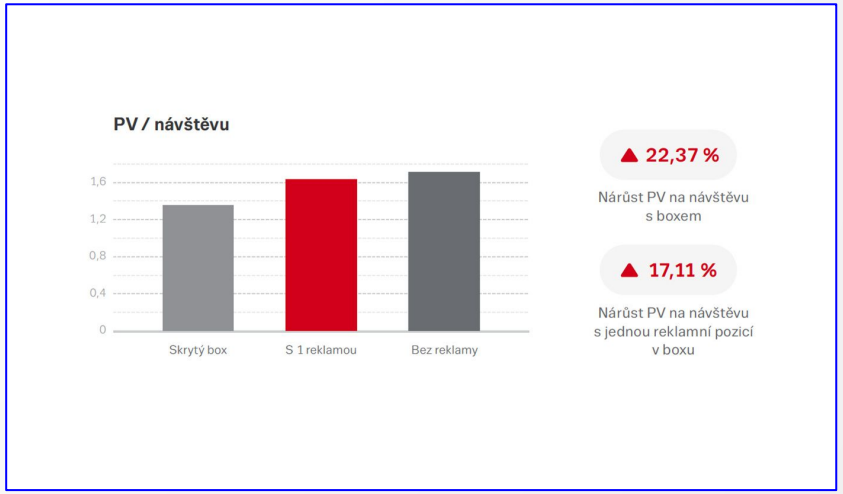
Why are Recommendation Systems important?



“Tyranny of choices”

- “Tyranny of choices”:
  too many options = user’s discomfort
- Ecommerce growth:
  order of magnitude per decade
  (5.7→42 Bln in global sales 2010 to 2020)

- 35% of Amazon purchases from recommendations
- 75% of Netflix watched contents from recommendations
- Seznam 2019 case study: 22% readability increase with RS



Agenda

Theoretical part

- About us
- About Seznam
- Introduction to recommender systems
- **Recommender system at Seznam**
- Recommender system infrastructure
- Ranking models
- Cold start problem

Practical part

- Tools setup and intro
- MIND dataset preparation
- Exploratory data analysis
- Embeddings for user representations
- Train ranking model with user representation

Seznam Zprávy • Pondělí 11. dubna. Svátek má Izabela.

Nákupy v Česku podražily o 12,7 %. Bude ještě hůř, varují ekonomové

Inflace v Česku dosáhla 12,7 procenta. Mohou za to především vysoké ceny energií a pohonných…

Naštvaní stážisti: Nechceme dotovat předsednictví EU z našich brigád

Ukrajinské děti v pasti. Dopoledne česká výuka, potom distanční z Ukrajiny

Concorde: Velký nadzvukový švindl

Sport

Šílenství v cíli. Emoce bouchly po závodě, piloti šli do sebe pěstmi

Emoce po závodě pořádně probublaly. Zatímco vítězství v posledním díle slavného okruhového…

Sedmdesátiny brankářského gentlemana. NHL legendu stále mrzí

Ronaldův zkrat! Hvězda zaútočila na malého fanouška, pak se omlouvala

Fantastický Krejčí! Český basketbalista v NBA vylepšil rekord

Garáž

Za volant od 17, vyšší tresty za hraní s mobilem. Ministerstvo dopravy chystá změny

Rozdávat nižší tresty za malichernosti, ale přísněji trest větší hříšníky. Tak by měl podle ministra…

Podobné auto na silnici nepotkáte: BMW iX má z budoucnosti design, techniku i cenu

Novinky

ŽIVĚ Bude až 20 stupňů, na Velikonoce se ale citelně ochladí

Nevyzpytatelnost dubnového počasí se naplno projeví i o velikonočním týdnu, který právě startu…

Už dva týdny se s námi nikdo nespojil, stěžuje si pluk Azov v Mariupolu

Drahé energie a pohonné hmoty vyhnaly inflaci na 12,7 procenta

Údaje o spotřebě tepla dostanou lidé každý měsíc

Nejstíhlejší mrakodrap na světě je hotov. První nájemníci se začínají stěhovat

Časté známky nízkého sebevědomí a sebeúcty

Zabavené jachty oligarchů polykají miliardy. Kdo to bude platit?

Uzavřenou Šanghaj ovládl hlad. Z oken zní zoufalý nářek obyvatel

Stream

Takhle jste ji ve Tváři ještě neviděli. Z vystoupení Nesvačilové vám bude běhat mráz po zádech Tvoje tvář má známý hlas

Denisa Nesvačilová měla v show Tvoje tvář má známý hlas štěstí zejména na sexy popové divy.…

Nejlepší velikonoční nádivka, která se vám zaručeně povede Lukáš Mozek

Jak ulevit od bolesti v bedrech: pět jednoduchých cviků, které byste měli cvičit pravidelně Proženy

I pejsci umí žárlit. Huskymu se nelíbí „příchod miminka“ Tady virál

Válka na Ukrajině • Українські новини

09:16 Banka Société Générale chystá odchod z Ruska. Dohodla se na prodeji svého podílu v Rosbank a dceřiných pojišťovacích firmách tohoto…

08:54 Německá armáda vypravila v pondělí speciální letoun pro přepravu

Super

Prsa jí vypadávala z dekoltu: S výstřihem do pasu se Aneta Vignerová nebála ani tančit

Rovnou z módního mola na Fashion Weeku přišla v modelu Michaela Kováčika na Český ples…

Bez make-upu, filtrů i dobrého nasvícení: Takhle vypadá Jennifer Lopez po ránu

Leoš Mareš se pochlubil rozkošnou dcerou: Malá Alex oslavila první měsíc na světě a je celý táta

Proženy

Jak ulevit od bolesti v bedrech: pět jednoduchých cviků, které byste měli cvičit pravidelně

Jak si ulevit od bolesti zad? Odložte mobil a počítač, natáhněte se na podložku a poctivě…

Pro velká prsa i drobnější postavu: nejhezčí jarní šaty, které teď koupíte v kolekcích

Nalepená prsa i vyšší čelo: Jak se Lily James proměnila v Pamelu Anderson

Týdenní horoskop: Střelce popadne „lenora“, Panny, pozor na sňatkového podvodníka

Koronavirus

Pozitivní případy	V nemocnici	Úmrtí	
+2 648	-181	+8	• Aktuální opatření
Reinfekce: 397	1 258	39 880	• Cestování

Díváte se na včerejší data, dnešní vydá MZČR okolo 08:30

HR Vše Videa Podcasty Nejkomentovanější

Formule

Newgarden se poprvé dočkal vítězství v Long Beach

Před 4 hodinami

Josef Newgarden si k nadcházejícímu rodičovství nadělil perfektní dárek. V neděli ovládl závod IndyCar na městsk…

Líbí se 0 Komentáře

AutoForum

FIA začala najednou tvrdě uplatňovat 17 let staré, zapomenuté pravidlo, potrápí jen Lewise Hamiltona

Před 3 dny

Je to zvláštní krok, když se najednou stane zásadním něco, na co si léta nikdo ani nevzpomněl. Důvody, proč FIA k …

Líbí se 16 Komentáře 8

Aktuálně

Woods zahrál nejhorší kolo na Masters v kariéře. V Augustě vede Sheffler

Před 1 dnem

Americký golfista Scottie Scheffler se po třetím kole Masters udržel v čele úvodního majoru sezony.

Líbí se 1 Komentáře

Mall.cz Reklama

Slevové kupony na Mall.cz

Největší nákupní svátek českého internetu. Nejlepší slevové kupóny a slevy.

Líbí se Komentáře

Prima Cool

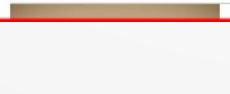
Rychle a zběsile 10 naverbovalo velkou hvězdu Avengers. Hlavním záporákem pak bude Aquaman

Před 1 hodinou

Slavná akční série přidává do svých řad další slavná jména. Její desátý díl půjde do kin v květnu 2023 a stále to …

Líbí se 0 Komentáře

fZone



Finsko má propracovaný plán pro případ ruské invaze

TOMÁŠ TRNĚNÝ

Příslušníci finských aktivních záloh během cvičení v jihovýchodním Finsku v březnu 2022.

10:05

Po desítky let se Finsko, které má s Ruskem hranici dlouhou přes 1 300 kilometrů, připravuje na konflikt se svým sousedem. Ruská invaze na Ukrajinu pak obavy Finů z útoku ještě zvýšila. Země ale hlásí, že je připravena.

Zásoby, evakuační prostory, bojeschopná a početná armáda. Na těchto třech pilířích stojí finská obrana před možnou ruskou agresí. Země má ze svého východního souseda, nejprve Sovětského svazu a později Ruska, obavy už přes 80 let a po celou dobu se připravuje na nejhorší. Může být Finsko inspirací i pro ostatní evropské země?

STALO SE

10:24
Ukrajina varuje před ruskými provokacemi v Podněstří

10:01
Africe hrozí, že se stane světovou popelnicí na plastový odpad

10:00
Nový žebříček Top 50 v českém fotbale: Za Křetínským došlo k povstání trenérů

DALŠÍ ČLÁNKY

DOPORUČOVANÉ

Concorde: Velký nadzvukový švindl

VČERA 19:21

Naštvaní stážisti: Nechceme dotovat předsednictví EU z našich brigád

7:45

Glosa: Čeká nás biblických 7 hubených let. Není na výběr

VČERA 12:00

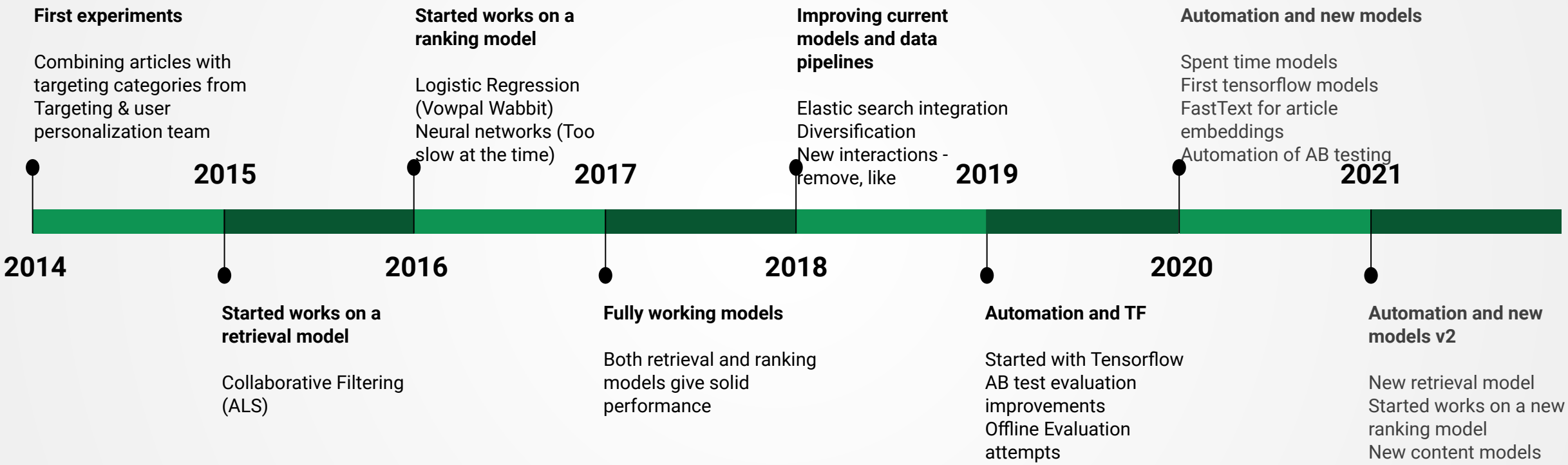
„Sex je jako komunismus.“ Objevili jsme soubor erotických povídek Petra Fialy

8:00

Seznam Native

Proč z české krajiny mizí ovocné stromy?

Recommender system at Seznam



Recommender system at Seznam

~10M

Clicks per day

~10K

Requests per second

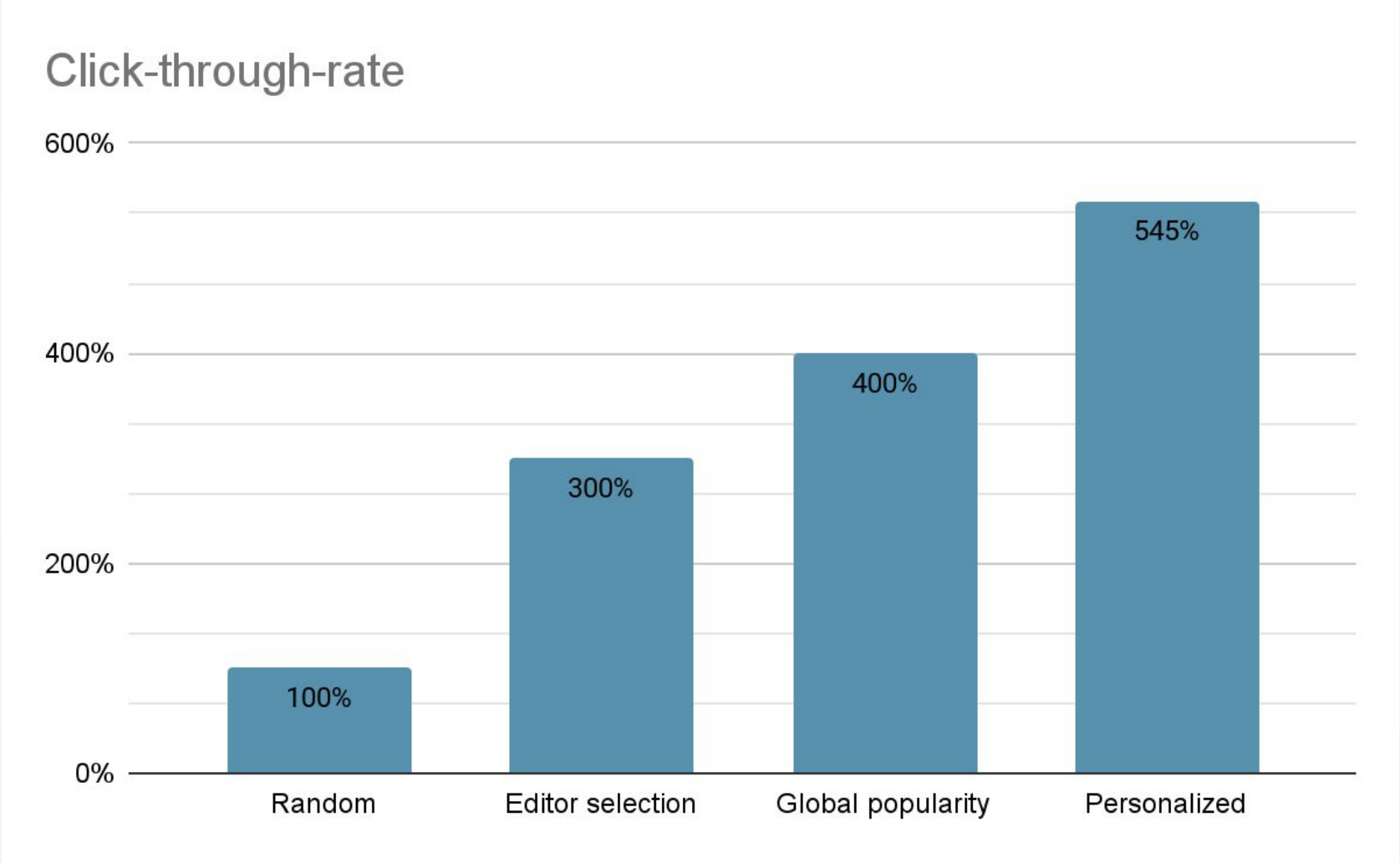
~1K

New items per day

~1K

Experiments per year

Algorithm performance



Agenda

Theoretical part

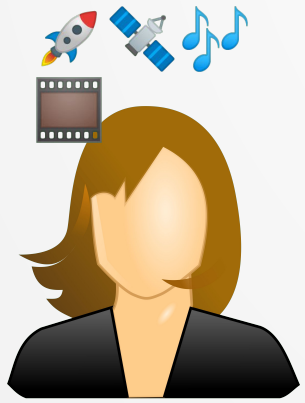
- About us
- About Seznam
- Introduction to recommender systems
- Recommender system at Seznam
- **Recommender system infrastructure**
- Ranking models
- Cold start problem

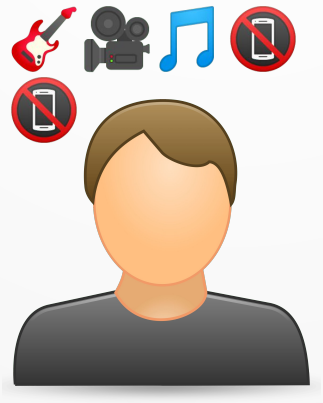
Practical part

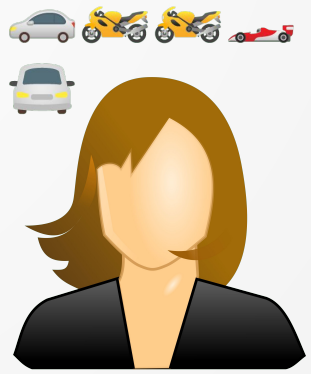
- Tools setup and intro
- MIND dataset preparation
- Exploratory data analysis
- Embeddings for user representations
- Train ranking model with user representation

Recommender systems

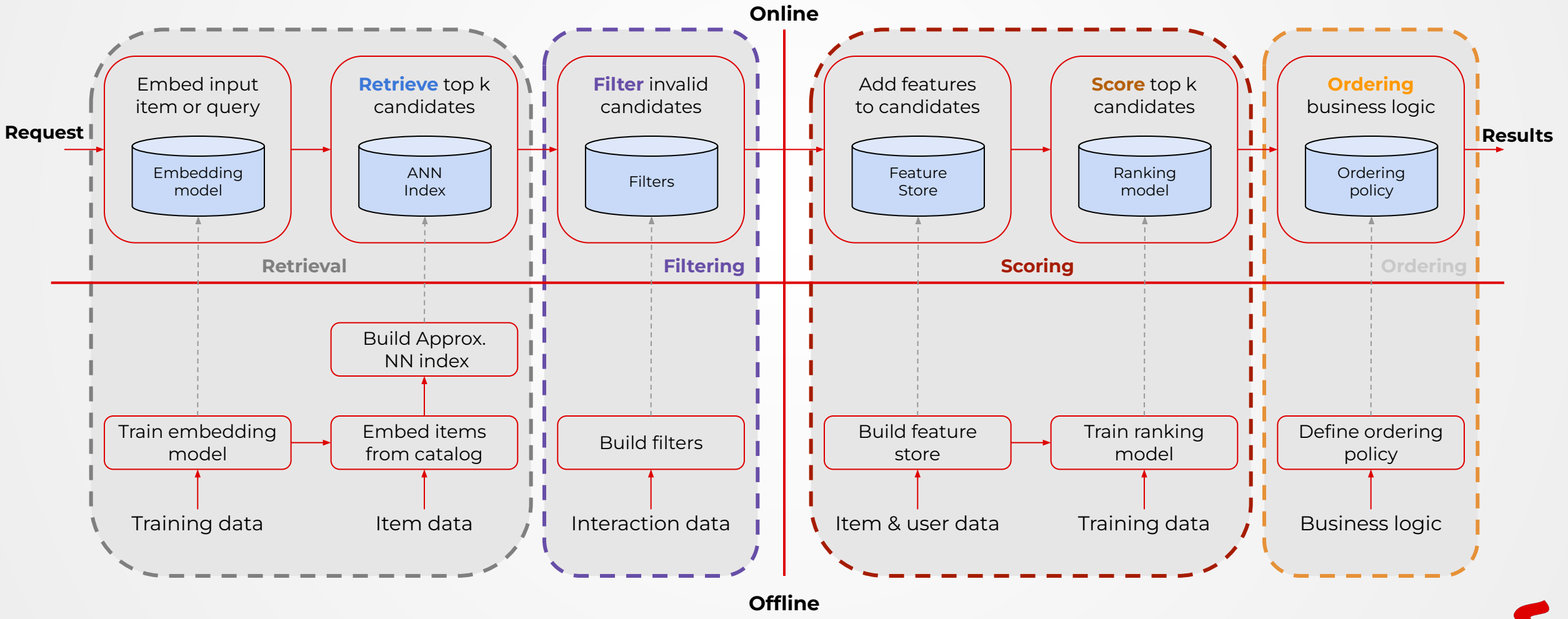
- Set of users, Set of items
- # of items >> # of items a user is able to read through
- The majority of items might be irrelevant for a user
- The goal: recommend only the items relevant to a user







Recommender systems infrastructure



Based on "**Moving Beyond Recommender Models**" by Even Oldridge and Karl Byleen-Higley (NVIDIA), https://www.youtube.com/watch?v=5qjiY-kLwFY

Agenda

Theoretical part

- About us
- About Seznam
- Introduction to recommender systems
- Recommender system at Seznam
- Recommender system infrastructure
- **Ranking models**
- Cold start problem

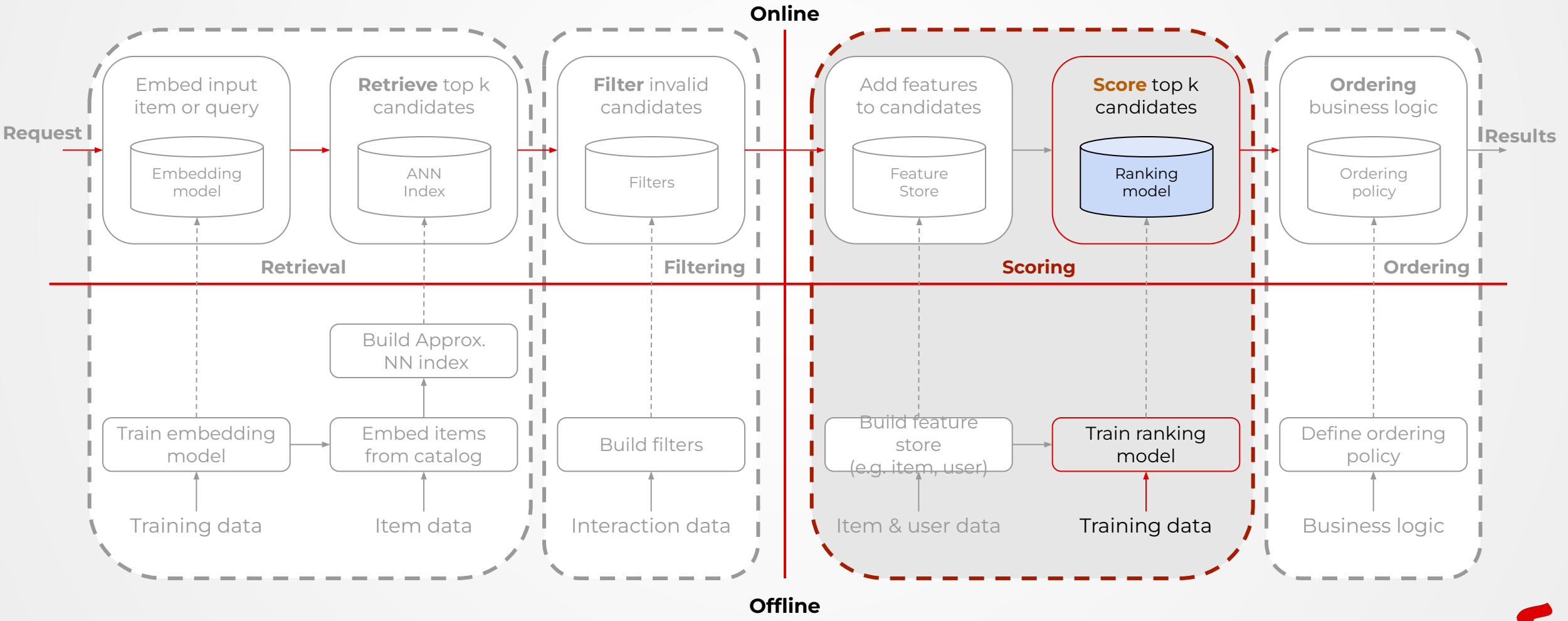
Practical part

- Tools setup and intro
- MIND dataset preparation
- Exploratory data analysis
- Embeddings for user representations
- Train ranking model with user representation

Practical part

- visit following link: https://tinyurl.com/48se6ze5
- open notebook according to instructions in README.md
- run notebook with prefix 003*

Ranking models



Based on "**Moving Beyond Recommender Models**" by Even Oldridge and Karl Byleen-Higley (NVIDIA), https://www.youtube.com/watch?v=5qjiY-kLwFY

Ranking models input - user features

Gender: Male

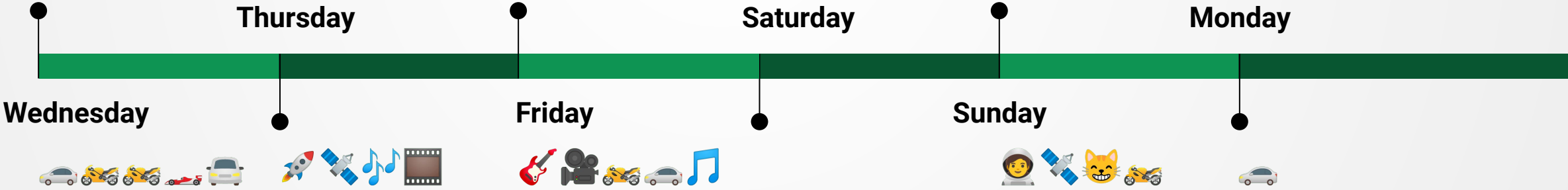
Age: 30s

Location: Prague

Device: Android Smartphone

…

Ranking models input - user interaction history



Ranking models input - item features

Title: Jeden z posledních Wartburgů: Svezli jsme se v autě z roku 1990. Pořád je to starý pohodlný Hans, jen už tak „neprdí“

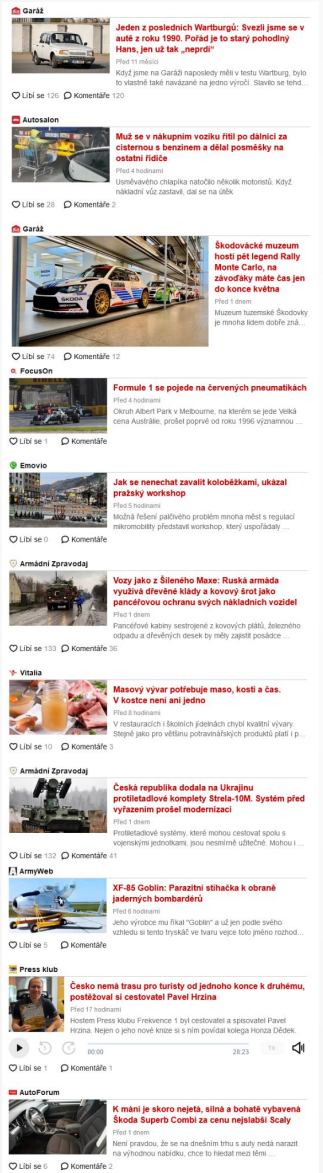
Publisher: Garáž.cz

Published: 15 hours ago

Tags: Cars, Oldtimers, Germany

…

Ranking models output - label



👈 ⏱️7 min

👈 ⏱️2 min

👈 ⏱️10 min = 35 %

Recommender systems evolution

Collaborative filtering methods

- Based only on user-item rating matrix
- Examples
  - User-based, Item-based
  - Matrix Factorization
  - …

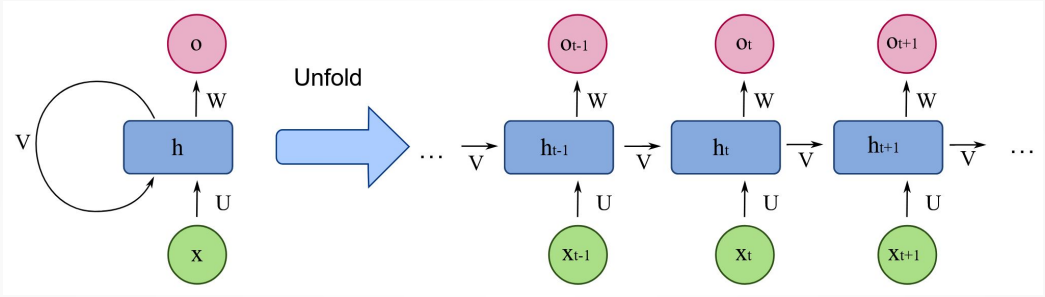
Recommendation as binary classification

- Based on any features available
- Incl. content-based
- Examples
  - Logistic regression (Vowpal Wabbit)
  - Simple NN
  - Factorization Machines
  - …

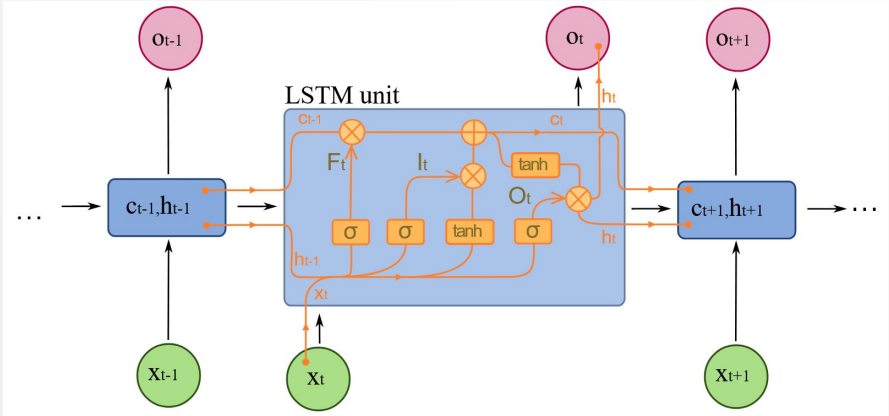
Advanced neural network models

Ranking models - preliminaries

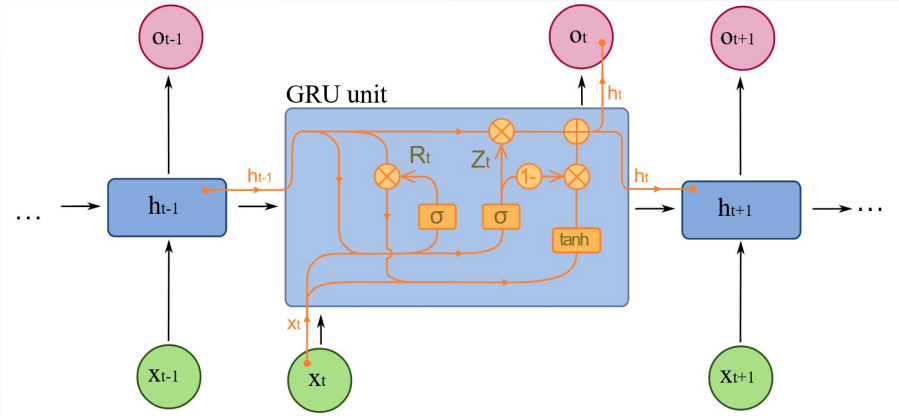
RNN - Recurrent Neural Network



LSTM - Long Short-Term Memory



GRU - Gated Recurrent Unit



New ranking model selection method

Gathering SOTA, reading & pre-selection

Experimental implementation

Offline metrics reality-check

Internal user testing I & II

Production implementation

Online AB tests

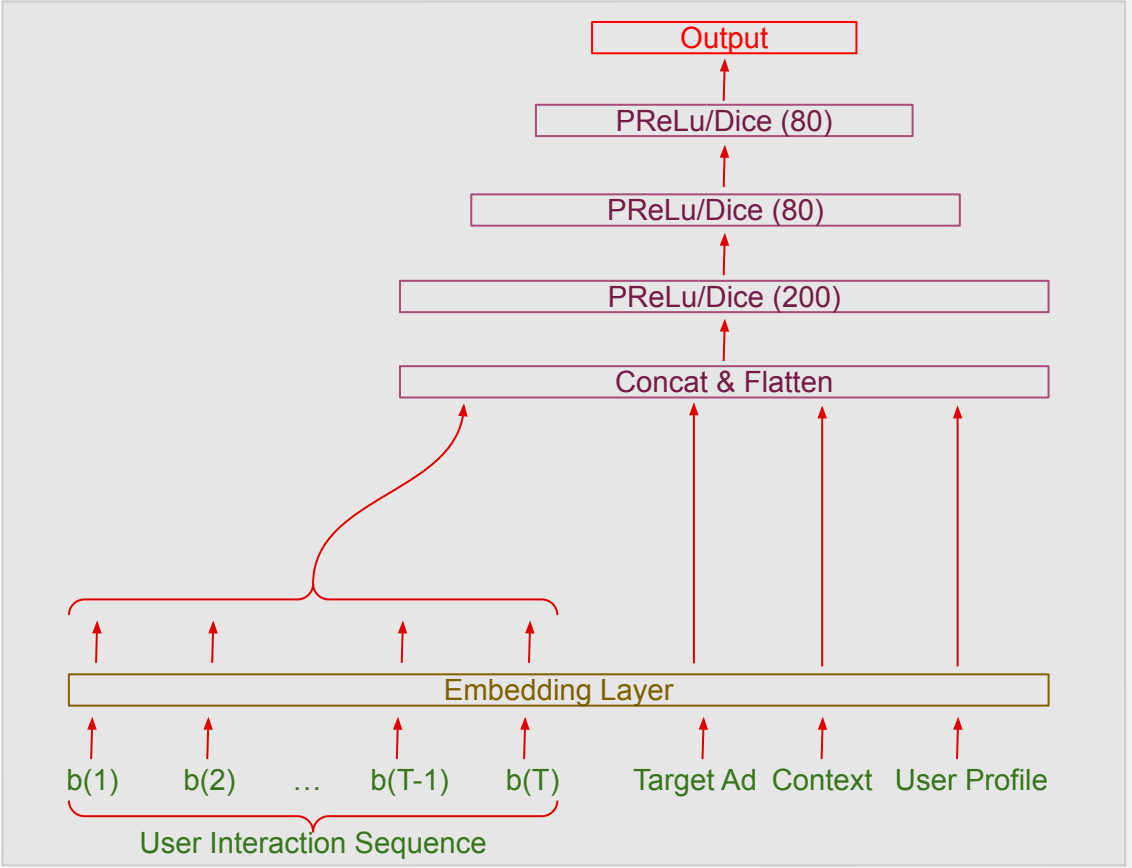
?

Ranking models - SOTA

- Mainly focused on the core, usually accompanied by standard DNN
- Ordered from least to most successful attempts:
  - DIEN
  - SLi-Rec
  - SUM
  - **DCN**

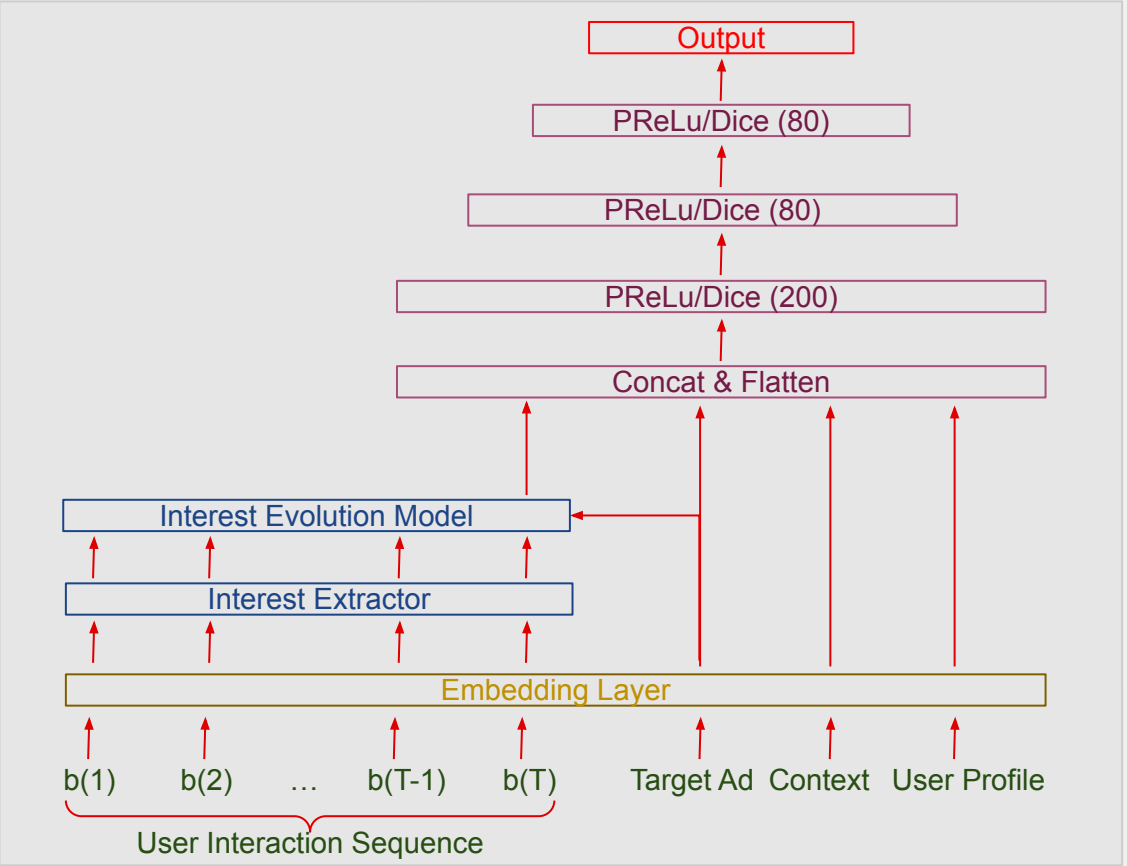
Deep Interest Evolution Network (DIEN)

- Embeddings
  + Interest Extractor
  + Interest Evolution Model
  + MLP



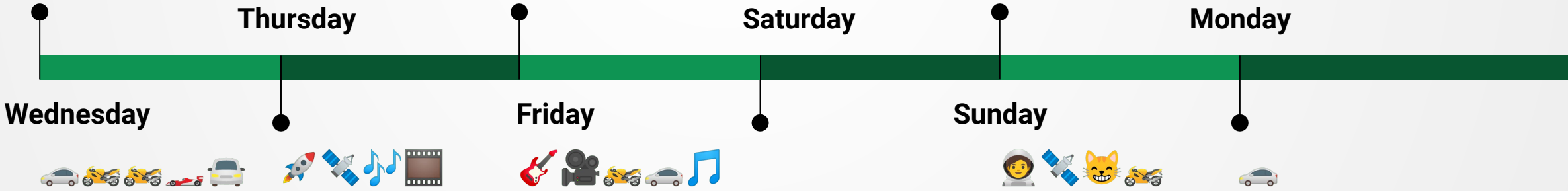
Deep Interest Evolution Network (DIEN)

- Embeddings
  + Interest Extractor
  + Interest Evolution Model
  + MLP
- Interest Extraction + Interest Evolution Model:
  - GRU based (Gate Recurrent Units) representation of latent temporal interest + sequence model
  - 20% online improvement over baseline model
  - Did not show good results @ Seznam RS



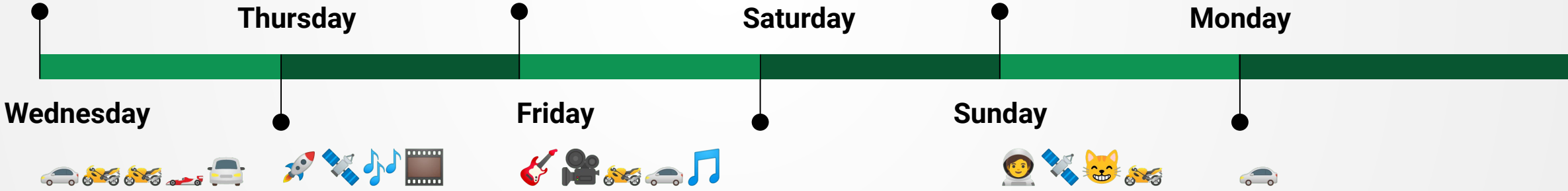
SLi-Rec - **S**hort-term and **L**ong-term preference **I**ntegrated **REC**ommender system

- Advanced sequential RNN for user preference modelling
- Short-term preference modelling
  - Upgraded LSTM
  - Time-aware controller - capture temporal distance between interactions
  - Intent-aware controller - contextual attentive mechanism to suppress deviations



SLi-Rec - Short-term and Long-term preference Integrated RECommender system

- Long-term preference modelling
- Attentive mechanism to adaptively combine short-term and long-term components based on context



Pros:

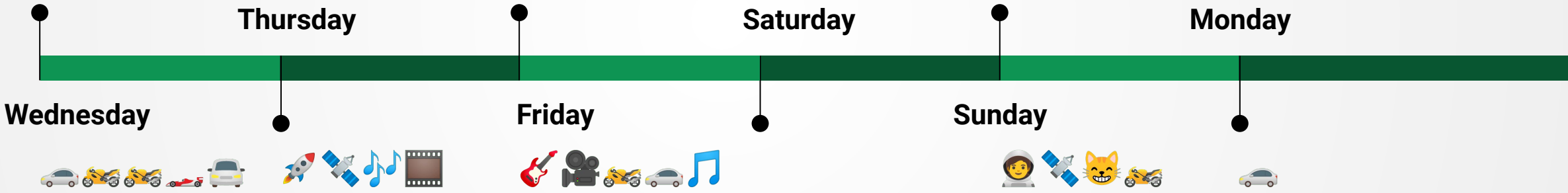
- Clear intuition of what model does
- SOTA performance

Cons:

- No real-world results presented
- Too slow and expensive in SZN practice
- Requires time-stamped data

SUM - Sequential User Matrix

- Multi-channel memory network for NRT large-scale RS
- User representations can be stored and updated incrementally



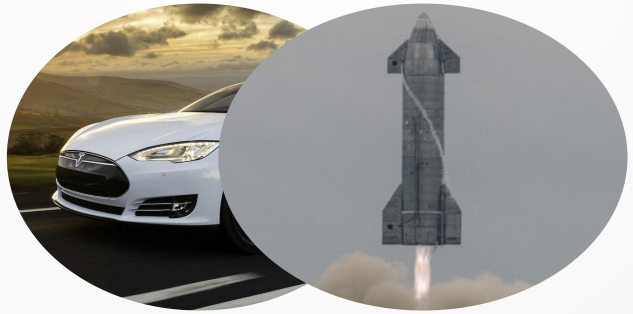
Pros:

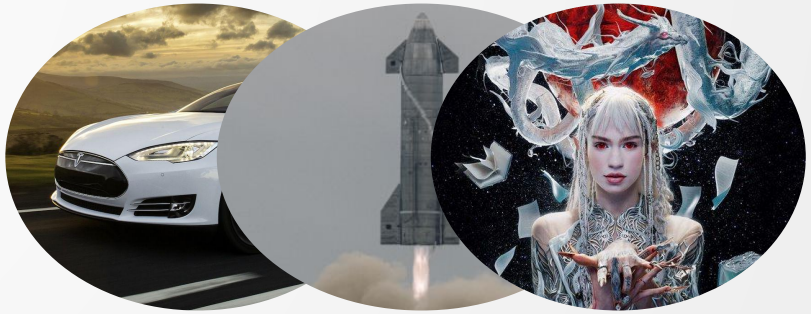
- Scalable, industry-level approach
- Online experiment results available

Cons:

- Less intuitive architecture
- Unable to make it work well enough at SZN (yet)

DCN - Feature interactions (crosses)





Single features

Quadratic interactions

Cubic interactions

DCN - Deep & Cross Network v2

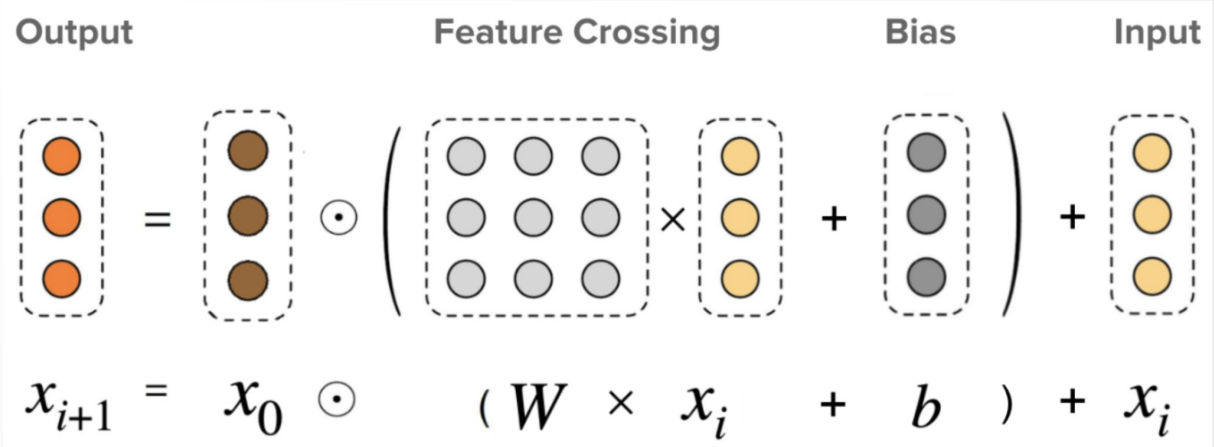
- Implicit f.i. modelled by DNN are not efficient
- Explicit f.i. were formerly hand-designed, now created by cross-layers
- Each cross-layer adds an order of f.i.

Pros:

- Avoids need of manual feature crossing
- Simple, elegant, general solution
- Further optimizable, scaleable, industry-level approach
- Tested at SZN, looking good

Cons:

- Not sequential



Agenda

Theoretical part

- About us
- About Seznam
- Introduction to recommender systems
- Recommender system at Seznam
- Recommender system infrastructure
- Ranking models
- Cold start problem

Practical part

- Tools setup and intro
- MIND dataset preparation
- Exploratory data analysis
- Embeddings for user representations
- Train ranking model with user representation

Cold-start

Missing interaction history?

Cold-start

Missing interaction history = likely poor recommendation

- Not optimal performance yet

Cold-start

Missing interaction history

= likely poor recommendation

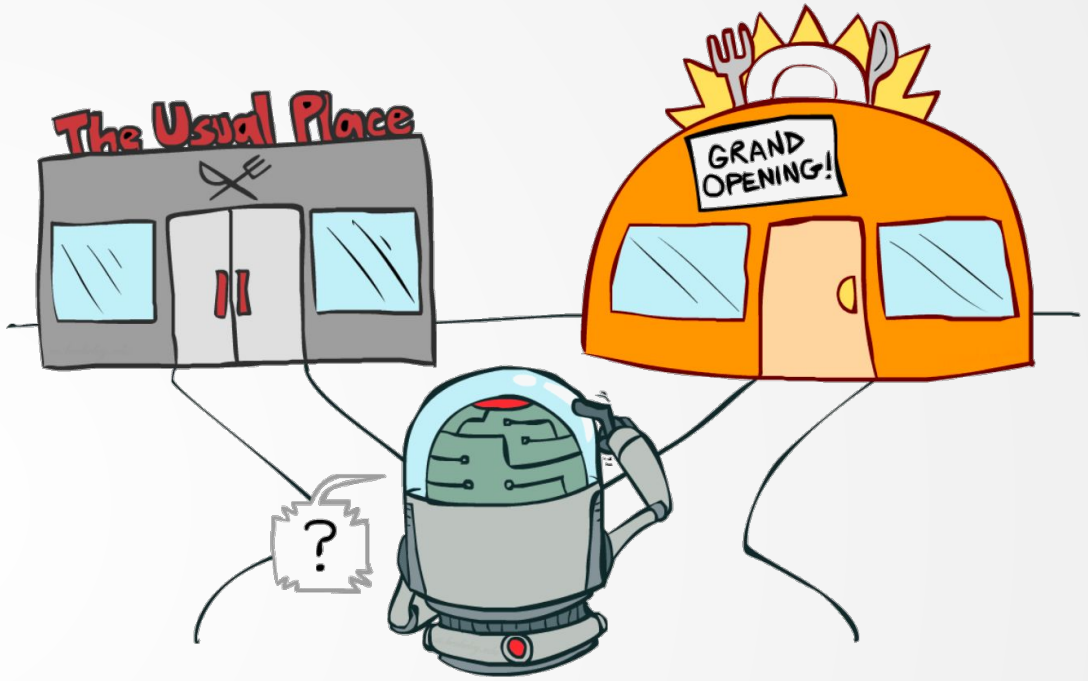
- Not optimal performance yet
- Item cold-start
  - New article, product = no or little interaction with users
  - Unable to kick-off promising items, user engagement loss
- User cold-start
  - New user, weak user identity (3p cookies) = no or little interaction with items
  - Unable to acquire new users

Cold-start mitigation

How to cope with cold-start?

Cold-start mitigation

- User cold-start
  - Popularity model - already known users
  - Metadata - cohort/segment
  - External behavior - social media
- Item cold-start
  - Exploration - multi-armed bandit
  - Metadata - category/topic
  - Content-based features - hybrid approach



Zdroj: https://lilianweng.github.io/posts/2018-01-23-multi-armed-bandit/

Extra user features

- User profile - provided metadata
- Other domain behaviour
  - User history - search queries
  - Classification, segmentation - interests
- Encoding
  - One-hot - sparse, unsuitable for NNs
  - Histogram - denser, but not optimal
  - Embedding
    - Dense - 10s-100s dimensions
    - Items retain similarity
    - Automatic compression/transformation

One-hot

																		
	I	II	⋯	VIII	IX	I	II	III	I	II	⋯	VI	VII	I	II	⋯	VIII	IX
A	✔	✔	⋯	✔	✔	✔	✔	✔	✔		⋯			✔	✔	⋯	✔	✔
B	✔		⋯			✔	✔	✔	✔	✔	⋯	✔	✔	✔		⋯		
C			⋯	✔	✔	✔	✔	✔			⋯			✔	✔	⋯	✔	✔

Histogram

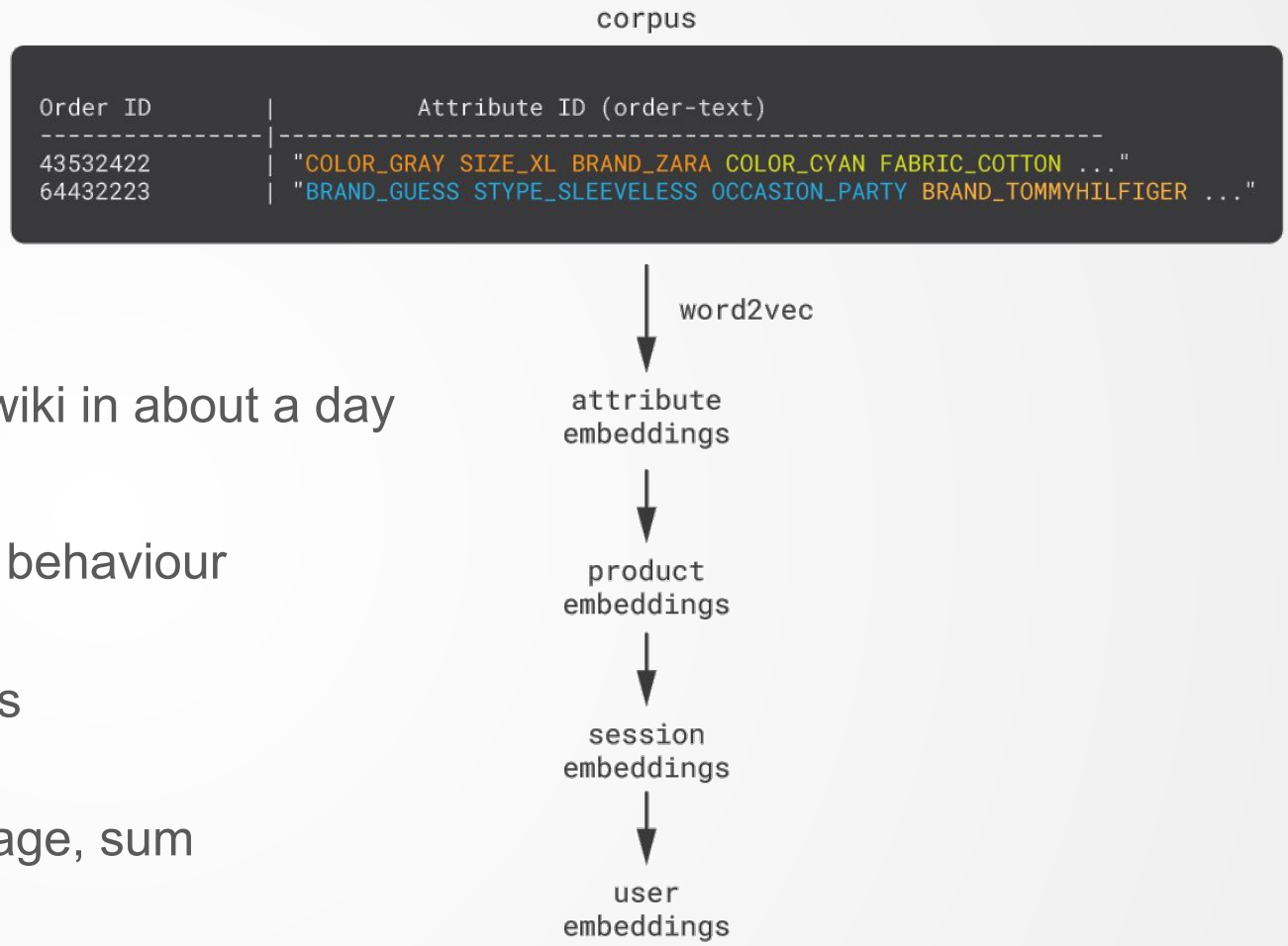
				
A	8	3	1	7
B	1	3	7	1
C	3	3	0	9

Embedding

A	(0.3, -0.7, 0.1)
B	(1.1, 0.2, -0.3)
C	(0.2, -0.4, 0.6)

User embedding

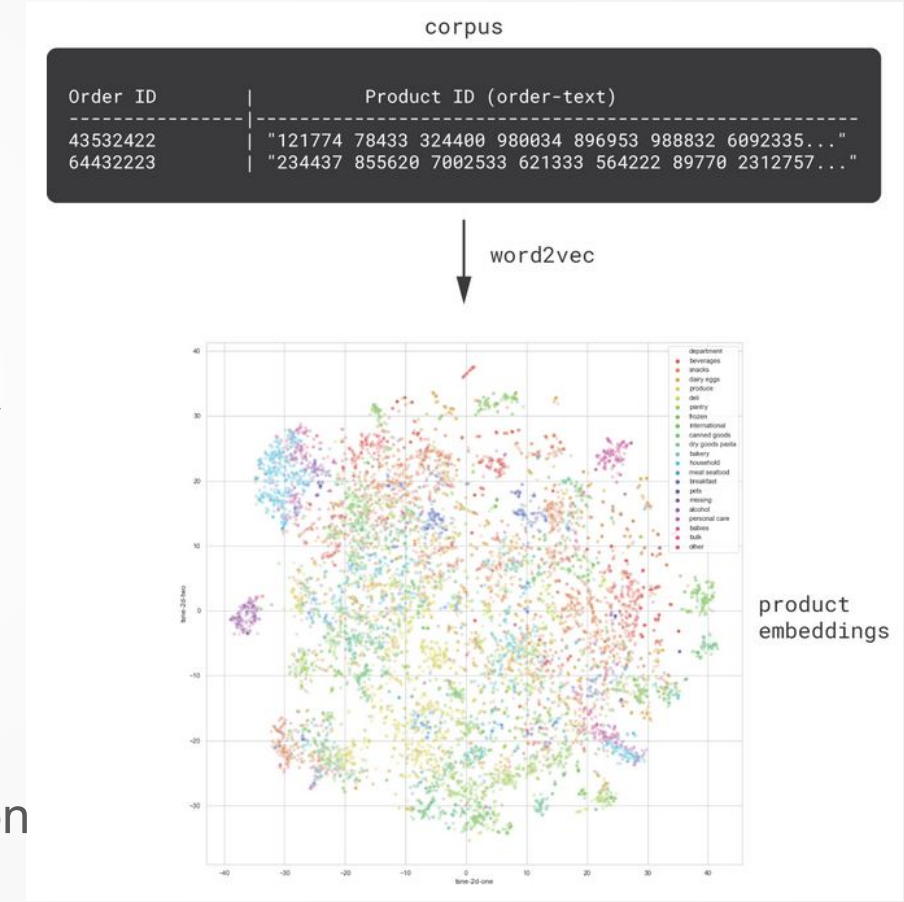
- Embedding (NLP)
  - Classification, translation
  - Word similarity ⇔ vector similarity
  - Very efficient - can process english wiki in about a day
- User history ⇔ Text document
  - Model user similarity using common behaviour
    - User actions ⇔ words
    - Users / sessions ⇔ Documents
  - User vector
    - Combine action vectors - average, sum
    - word2vec, fastText



Zdroj:
https://blog.griddynamics.com/customer2vec-representation-learning-and-automl-for-customer-analytics-and-personalization/

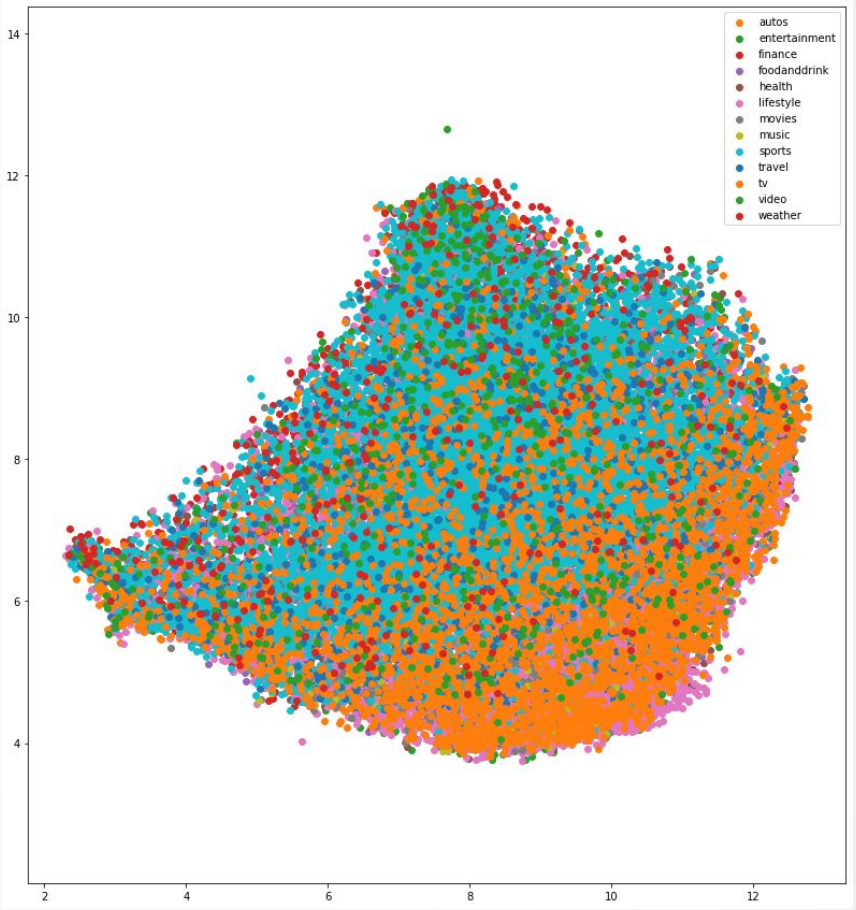
Embedding computation

- Word2vec
  - Embed history
    - Documents ⇔ users / sessions
    - Words ⇔ item clicks
  - User vector = combined item vectors (average, sum)
- Doc2vec - computes document vectors directly
- FastText
  - Employs subword information (char. n-grams)
  - Out of vocabulary inference
- StarSpace
  - Diverse problems - text classification, recommendation
  - Embeds different types of entities



Embedding evaluation

- Auxiliary
  - Vector space organization = debugging
  - Visualization
    - 2D/3D projection (UMAP, t-SNE)
    - Annotations - metadata (age, gender), segments
  - Metadata classification
- Downstream
  - Task performance - segmentation, recommendation
  - Hyperparameter optimization



Agenda

Theoretical part

- About us
- About Seznam
- Introduction to recommender systems
- Recommender system at Seznam
- Recommender system infrastructure
- Ranking models
- Cold start problem
- Offline evaluation challenges

Practical part

- Tools setup and intro
- MIND dataset preparation
- Train and evaluate baseline models
- Embeddings for user representations
- Train ranking model with user representation

A/B tests vs off-policy evaluation

A/B tests

- Experimental and control group
- Runs on live traffic
- Run for week(s) and evaluate

Off-policy evaluation

- Use historical data - offline
- Time << week(s)
- Conduct A/B test only if there is a potential

Popular metrics

- Precision@k: TP[:k] / (TP[:k] + FP[:k])
- Recall@k : TP[:k] / (TP + FN)
- MAP (mean average precision):
  - AP: $\frac{\sum_k P@k \times rel(k)}{TP + FN}$ - MAP: $\frac{1}{N}\sum_i^N AP_i$
- nDCG
  - DCG: $\sum_k \frac{2^{rel(k)} - 1}{log_2(k+1)}$ nDCG: DCG / ideal DCG (if all relevant items were at top)
- Perplexity: $2^{Entropy}$
- Novelty: $-log(p_i)$
- Coverage: # recommended items / all items

Historical data drawbacks

- Only implicit interactions
- Not uniformly distributed data
  - Biased towards the previous model (Exposure bias)
- Uniformly distributed data are expensive
- Biases in data
  - Selection bias
  - Position bias

Debiasing historical data

- Debias exposure bias
- IPS (inverse propensity score)
- Propensity : probability of observing user/item interactions $P_{u,i} = P(O_{u,i} = 1)$
  - Known x estimated
- Mean squared error with Propensity:

$$\frac{1}{U \cdot I} \sum_{(u,i):O_{u,i}=1} \frac{(Y_{u,i} - \hat{Y}_{u,i})^2}{P_{u,i}}$$

- Model the bias explicitly
  - Click models (position bias)
  - Add bias to the model

IPS limitations

- The unbiasedness of the IPS-based estimator is guaranteed only when the true propensities are available
- IPS has high variance
- The problem gets worse with large item collections
- More advanced approaches needed for slate recommendations

Summary

- Introduction to recommender systems
- Complexity/performance tradeoff in SOTA models
- Cold-start problem can be tackled by using additional data
- Feature embedding is the way to go

Thank you!

Questions & inquiries	→	vit.libal@firma.seznam.cz
We are hiring!	→	https://kariera.seznam.cz

- **Relevance in display advertising:**
  - https://kariera.seznam.cz/402701-vyzkumnik-v-oblasti-machine-learning/
  - ML models to predict clicks in display advertising, exploration & exploitation, bandits.
- **Online auction Bidding Automation:**
  - https://kariera.seznam.cz/403956-machine-learning-vyzkumnik-automatizace-bidovani/ Reinforcement learning a control theory for bidding strategy automatizaci, ML models to predict ad conversions.
- **Relevance in search advertising:**
  - https://kariera.seznam.cz/403972-machine-learning-vyzkumnik-relevance-reklamy-ve-vyhledavani/
  - ML modely predikce prokliku reklamy ve vyhledávání včetně NLP technologií.
- **Targeting and personalization:**
  - https://kariera.seznam.cz/400105-vyzkumnik-strojoveho-uceni-pro-cileni-reklamy-a-personalizaci/
  - ML models to estimate user profile and interests.
- **Recommendation systems:**
  - https://kariera.seznam.cz/395074-machine-learning-vyzkumnik-pro-doporucovaci-systemy/
  - ML models to recommend content
- **Operational research:**
  - https://kariera.seznam.cz/405314-operacni-vyzkum-data-scientist/
  - Discrete optimization, ML modeling of online auctions, game theory for design of ad systems and its logic.

References

Motivation:

- https://redstagfulfillment.com/2010s-ecommerce-growth-decade/
- https://faculty.washington.edu/jdb/345/345%20Articles/Iyengar%20%26%20Lepper%20(2000).pdf
- https://blog.seznam.cz/2020/07/pripadova-studie-zvyste-pocet-zhlednutych-stranek-a-sve-vynosy-diky-seznam-doporucuje/
- https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers

DIEN:

- https://arxiv.org/abs/1809.03672
- https://github.com/mouna99/dien

SLi-Rec:

- https://www.microsoft.com/en-us/research/uploads/prod/2019/07/IJCAI19-ready_v1.pdf

SUM:

- https://arxiv.org/abs/2102.09211

DCN:

- https://arxiv.org/abs/2008.13535

User Embedding:

- [1] https://jalammar.github.io/illustrated-word2vec/
- [2] https://fasttext.cc/
- [3] https://medium.com/wisio/a-gentle-introduction-to-doc2vec-db3e8c0cce5e
- [4] https://ai.facebook.com/tools/starspace/

Please leave us your feedback:

https://forms.gle/s2F3qzPr5WS3mMQy9