

# Recommendation systems and user representations



Seznam Advertising Systems

SEZNAME.CZ

# Agenda

## Theoretical part (90 min. + break)

- *practical part prep: notebook #001*
- About us
- About Seznam
- Introduction to recommender systems
- Recommender system at Seznam
- Recommender system infrastructure
- *practical part prep: notebook #003*
- Ranking models
- Cold start problem

## Practical part (90 min.)

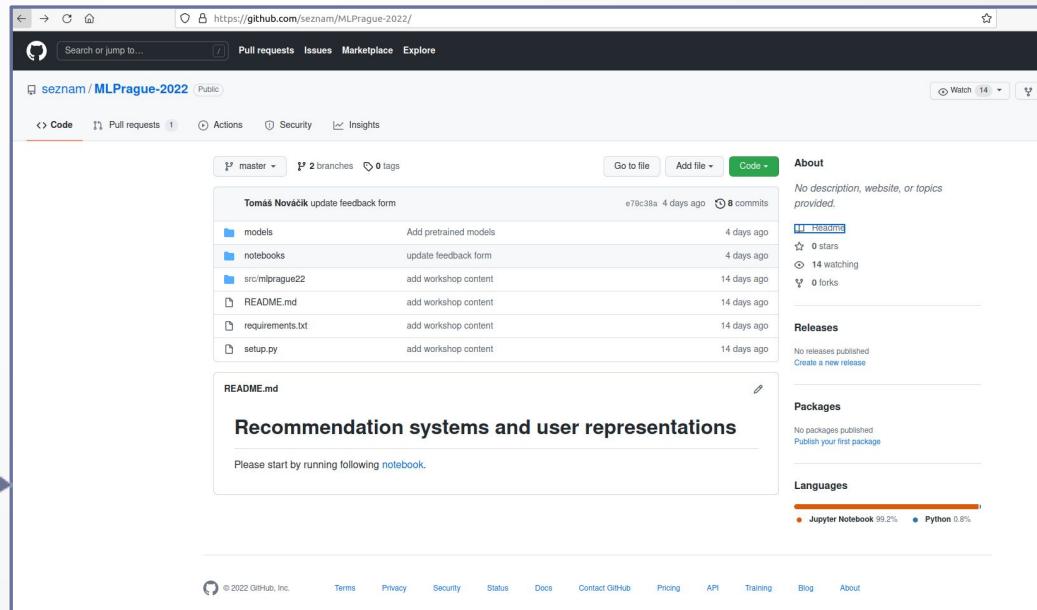
- Tools setup and intro
- MIND dataset preparation
- Exploratory data analysis
- Embeddings for user representations
- Train ranking model with user representation



# practical part prep: notebook #001

- 1) go to: <https://tinyurl.com/48se6ze5>
- 2) open the initial notebook (as in README.md)
- 3) open the “001-prepare-dataset”
- 4) copy
- 5) launch

1.



# practical part prep: notebook #001

1) go to: <https://tinyurl.com/48se6ze5>

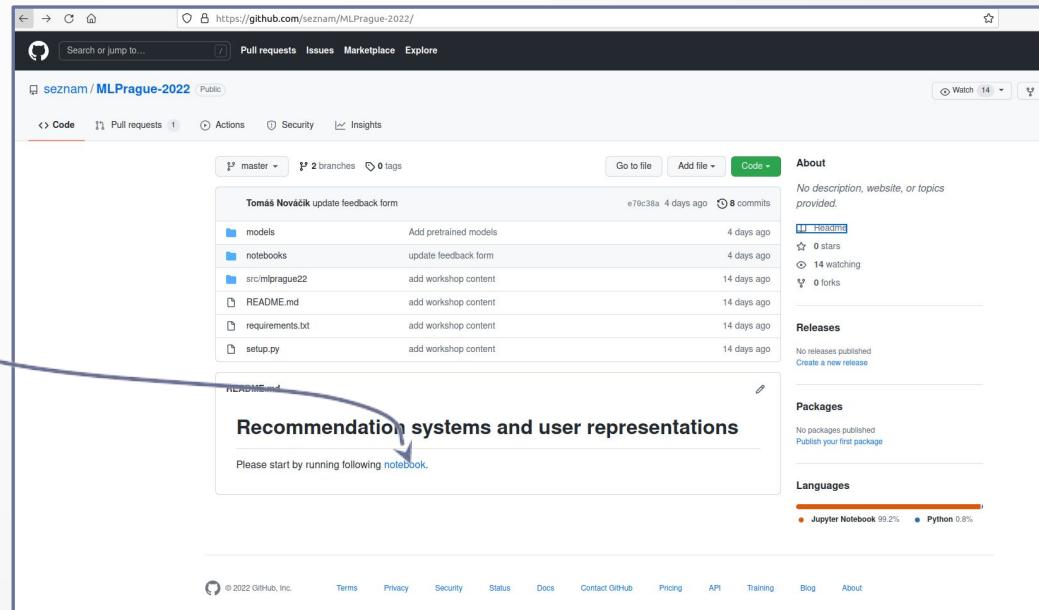
2) open the initial notebook (as in README.md)

3) open the “001-prepare-dataset”

4) copy

5) launch

2.



# practical part prep: notebook #001

1) go to: <https://tinyurl.com/48se6ze5>

2) open the initial notebook (as in README.md)

3) open the “001-prepare-dataset”

4) copy

5) launch

3.

Workshop on Recommendation systems and user representations

Motivation

Following tutorial tries to illustrate typical production setting in a company which provides multiple types of content and have various recommender systems in place.

Single recommender system is typically focused only on the part of the content portfolio in which it operates and will not have access to full user interaction sequence - which is typically due to RAM/CPU/storage/logistics restrictions.

As an example we can take look at the [seznam.cz](#) page on which we can find recommended articles for website novinky.cz.

At the same time user might have visited different website in Seznam's ecosystem e.g. [zbozi.cz](#) which has its own set of recommender systems which serve various purposes and implicitly do not model user behavior in Seznam's ecosystem.



In order to achieve efficient personalization across many models one needs to create efficient user representation which might also alleviate [cold-start problem](#).

Following tutorial tries to:

- illustrate various techniques which might yield such efficient user representation
- demonstrate how such representation might be used in ranking model which is an essential part of recommender system

Agenda

1. At first we will download and prepare [MIND dataset](#) by running notebook [001-prepare-dataset](#)
2. At the next phase we will investigate attributes of newly created dataset by running notebook [002-explorative-data-analysis](#)
3. User representation will be computed in notebook [003-user-representation-embedding](#)
4. At the end we will train [DCN model](#) on created user representation by running notebook [004-train-dcn-model-user-representation](#)

# practical part prep: notebook #001

1) go to: <https://tinyurl.com/48se6ze5>

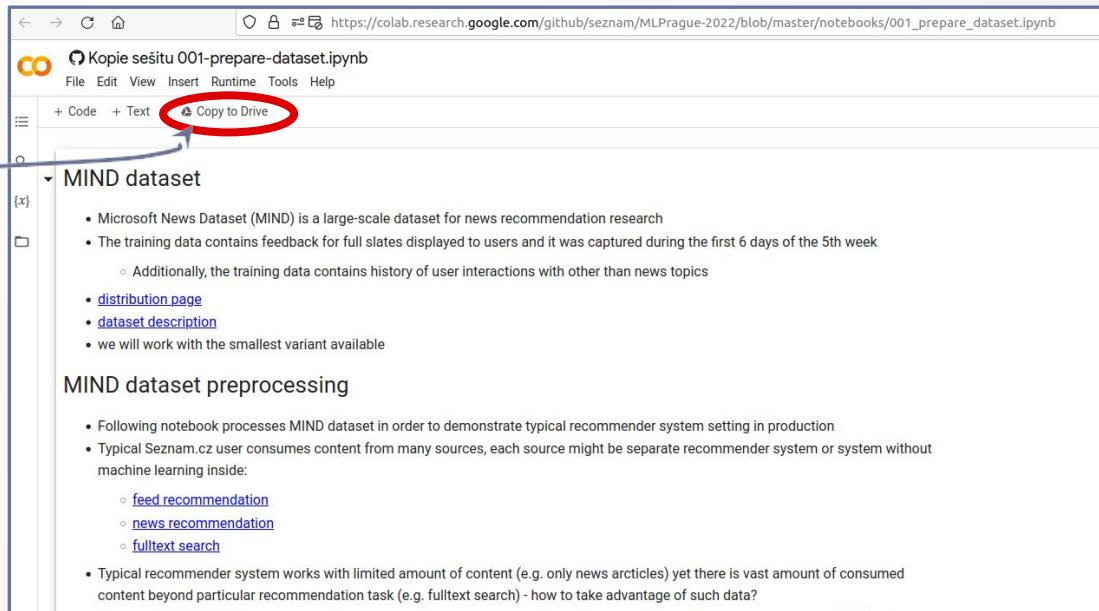
2) open the initial notebook (as in README.md)

3) open the “001-prepare-dataset”

4) copy (+sign in)

5) launch

4.



The screenshot shows a Google Colab notebook titled "Kopie sešitu 001-prepare-dataset.ipynb". The top navigation bar includes File, Edit, View, Insert, Runtime, Tools, and Help. Below the title, there are buttons for "+ Code" and "+ Text", followed by a "Copy to Drive" button, which is circled in red. The main content area is divided into sections: "MIND dataset" and "MIND dataset preprocessing". The "MIND dataset" section contains a bulleted list about the Microsoft News Dataset (MIND), mentioning it's a large-scale dataset for news recommendation research, training data with user feedback, and history of interactions with other topics. It also lists links for "distribution page", "dataset description", and notes about working with the smallest variant available. The "MIND dataset preprocessing" section lists steps for processing the dataset, including "feed recommendation", "news recommendation", and "fulltext search", along with a note about typical recommender system challenges like limited content and vast consumed content.



# practical part prep: notebook #001

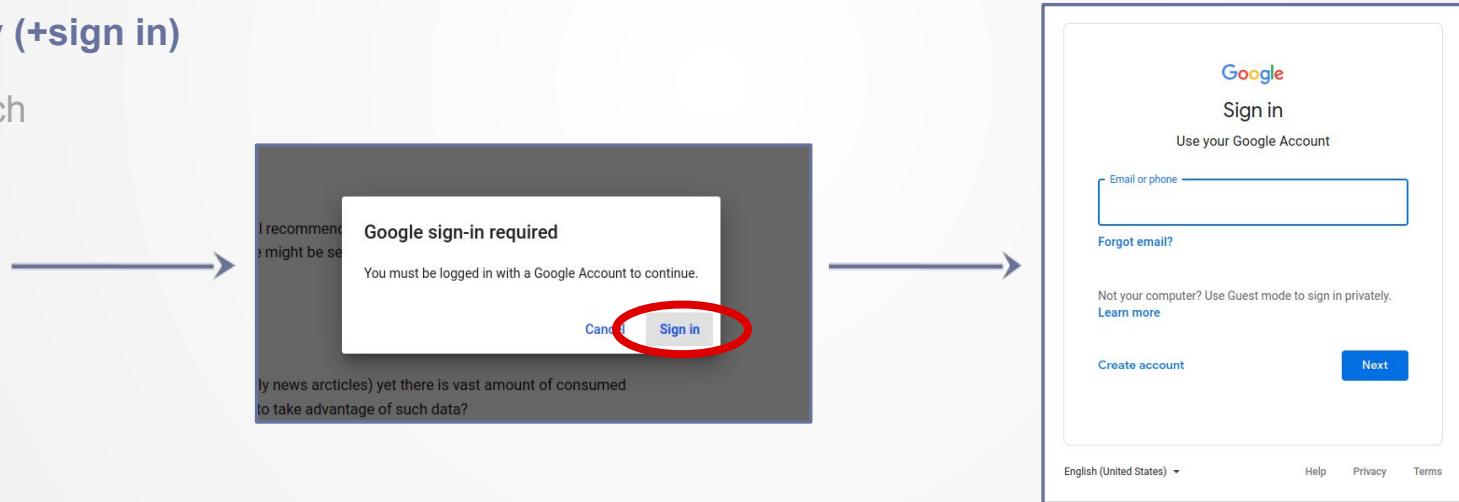
1) go to: <https://tinyurl.com/48se6ze5>

2) open the initial notebook (as in README.md)

3) open the “001-prepare-dataset”

4) copy (+sign in)

5) launch



# practical part prep: notebook #001

1) go to: <https://tinyurl.com/48se6ze5>

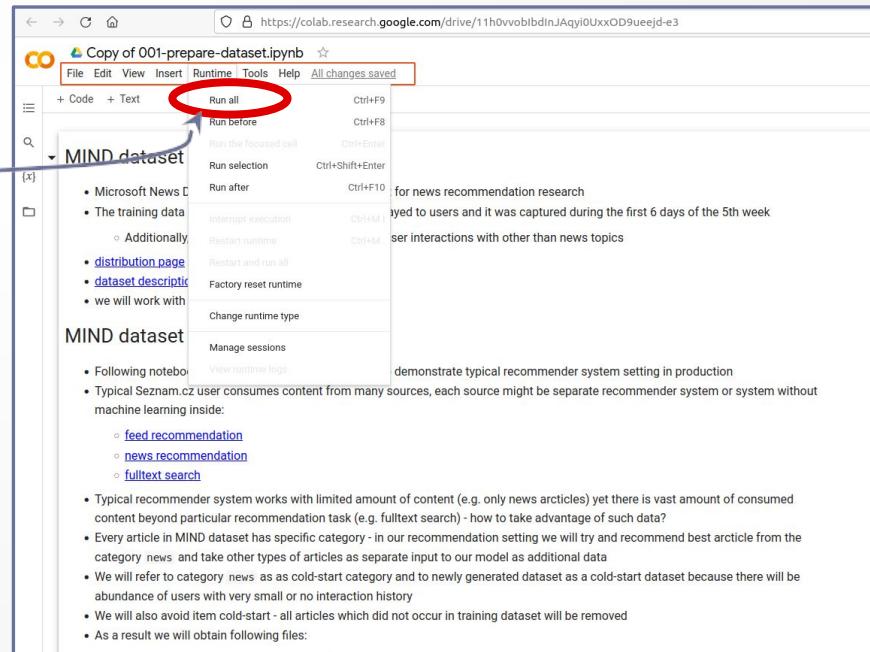
2) open the initial notebook (as in README.md)

3) open the “001-prepare-dataset”

4) copy (+ sign in)

5) Launch (+permissions)

5.



# practical part prep: notebook #001

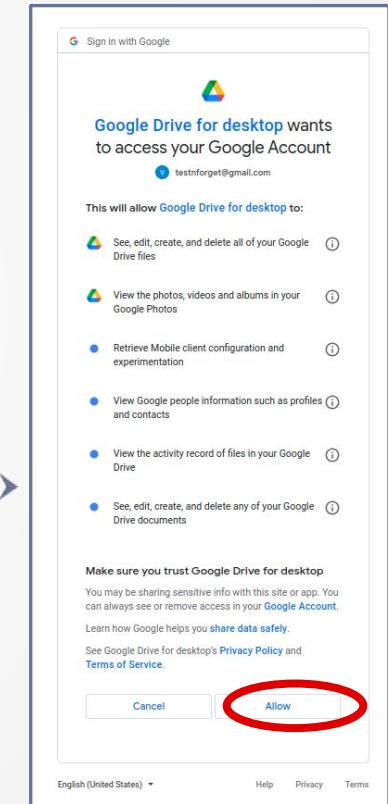
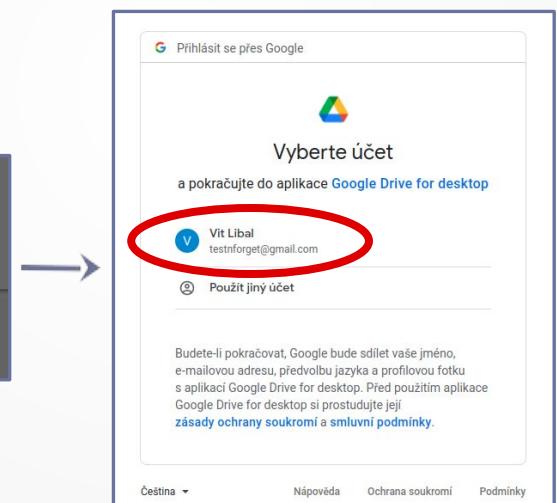
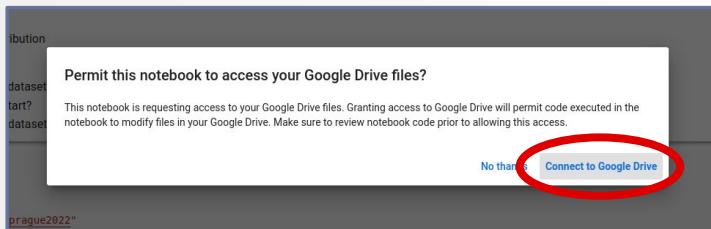
1) go to: <https://tinyurl.com/48se6ze5>

2) open the initial notebook (as in README.md)

3) open the “001-prepare-dataset”

4) copy (+ sign in)

5) Launch (+permissions)



# Agenda

## Theoretical part

- Advance the notebooks
- **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

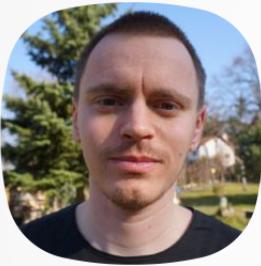


# Authors



**Václav Blahut**

Recommender  
Systems



**Radek Tomšů**

Recommender  
Systems



**Tomáš Nováčik**

Targeting  
& personalization



**Adam Jurčík**

Targeting  
& personalization



**Vít Líbal**

Relevance  
in display advertising

# Agenda

## Theoretical part

- Advance the notebooks
- 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

S | Televize Seznam

S

# About Seznam.cz

- Most visited content websites on Czech Internet
- About 7.3 million unique users visit Seznam.cz services every month

 PROŽENY.CZ

Novinky.cz

Seznam Zprávy |

 GARÁŽ.CZ

**SUPER.CZ**

**SPORT.CZ**



**95 %**

Monthly reach of the  
Czech online population



# Agenda

## Theoretical part

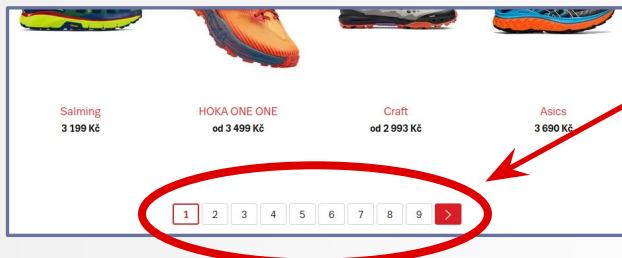
- Advance the notebooks
- About us
- About Seznam
- **Introduction to recommender systems**
- Recommender system at Seznam
- Recommender system infrastructure
- Ranking models
- Cold start problem

## 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?

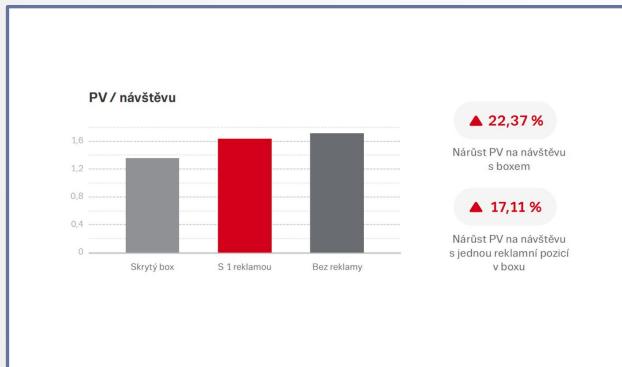


- “Tyranny of choices”:

- 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

- Advance the notebooks
- 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

Zkuste vyhledat "Křížové cesty"

Vyhledat

Internet Firmy Mapy Zboží Obrázky Slovník Jízdní rády Video

Právě se hledá: Počasi Velikonoce 2022 Brzobohatý a Pisařovicová Pašijový týden Nehoda na D5

Sauto Kupi Obrázky Volná místa Zboží Slovník Recepty Deníky Sdovolená Hry Mobilní aplikace Podcasty Pohádky Prohlížeč Sitky

**SEZNAM.CZ**

Nákupy v Česku podrážily o 12,7 %. Bude ještě hůř, varuj 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 svindl

**Novinky**

Bude až 20 stupňů, na Velikonoce se ale citelně ochladí

Nevyzpytatelnost dubnového počasa se naplno projeví o velikonočním týdnu, který právě startuje...

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

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

Údaje o spotřebě tepla dostanou lidé každý měsíc

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

Částečně známky nizkého sebevědomí a sebevěty

Zabavené jachty oligarchů milardy. Kdo to bude platiť?

Uzavřenou Šanghaj ovládla hlad. Z oken zničených nářek obyvatel

**Sport**

Šílenství v cili. Emoce bouchly po závodě, piloti šili do sebe pěstmi

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

Sedmdesátiny brankářského gentlemana. NHL legendu stále mrzí

Ronaldov zkratí Hvězda zaútočila na malého fanouška, pak se omlouvala

Fantastic Krejčí! Český basketbalista v NBA vylepšil rekord

**Stream**

Takhle jste jí ve Tváři ještě neviděli. Z vystoupení Nesváčilové vám bude běhat mráz po zádech

Tvoje tvář má známý hlas Denisa Nesváčilová měla v show Tvoje tvář má známý hlas šesté 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

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

**Garáž**

Za volant o 17, vyšší testy za hraní s mobilem. Ministerstvo dopravy chystá změny

Rozdávání nižších testů za malichernosti, ale přísnější test větší hříšníky. Tak by měl podle ministra...

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

**Válka na Ukrajině** - Ukrainské noviny

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

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

**Super**

Právě jí vypadávala z dekoltu: S výstříhem 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 pochybil rozkošnou dcerou: Malá Alex oslavila první měsíc na světě a je celý táta

**Prozřet**

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 pocítiv...

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

Týdeník horoskop: Střelce popadne „Jenora“, Panny, pozor na sňatkoveného podvodníka

**Koronavirus**

Positivní případy	V nemocnicích	Umrtí	Aktuální opatření
+2 648	-181	+8	
Reinfekce: 397	1 258	39 880	Cestování

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



Seznam

SEZNAME.CZ

Zkuste vyhledat "Velikonoční dekorace"

Vše

Formule

Newgarden se poprvé dočkal vítězství v Long Beach

Před 4 hodinami

Joséf Newgarden si k nadcházejícímu rodičovství nadělil perfektní dárek. V neděli ovládla závod IndyCar na městsk...

Libi se 0 Komentáře

AutoForum

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

Před 3 dny

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

Libi se 16 Komentáře 8

Aktuálně

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

Před 1 dnem

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

Libi se 1 Komentáře

Mall.cz

Slevové kupony na Mall.cz

Překlada

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

Prima Cool

Rychle a zběsile 10 naverbovalo velkou hvězdu

Před 1 hodinou

Avengers. Hlavním záporákem pak bude Aquaman

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

Libi se 0 Komentáře

fZone



Zprávy > Svět > Finsko má propracovaný plán pro případ ruské invaze

## 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, evakuacní 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?

### DOPORUČOVANÉ



Concorde: Velký nadzvukový švindl

VČERA 19:21

Naštívání stážistů: Nechceme dotovat předsednictví EU z našich brigád

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

#213 STASÍNEK

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

Seznam Native

Proč z české krajiny mizí ovocné stromy?



# Recommender system at Seznam

## First experiments

Combining articles with targeting categories from Targeting & user personalization team

## Started works on a ranking model

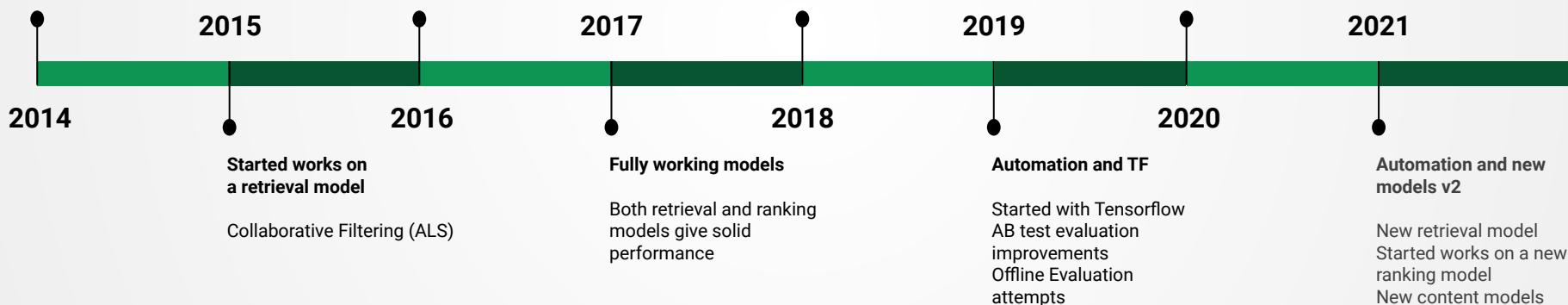
Logistic Regression (Vowpal Wabbit)  
Neural networks (Too slow at the time)

## Improving current models and data pipelines

Elastic search integration  
Diversification  
New interactions - remove, like

## Automation and new models

Spent time models  
First tensorflow models  
FastText for article embeddings  
Automation of AB testing

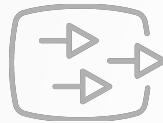


# Recommender system at Seznam



**~10M**

Clicks  
per day



**~10K**

Requests  
per second



**~1K**

New items  
per day



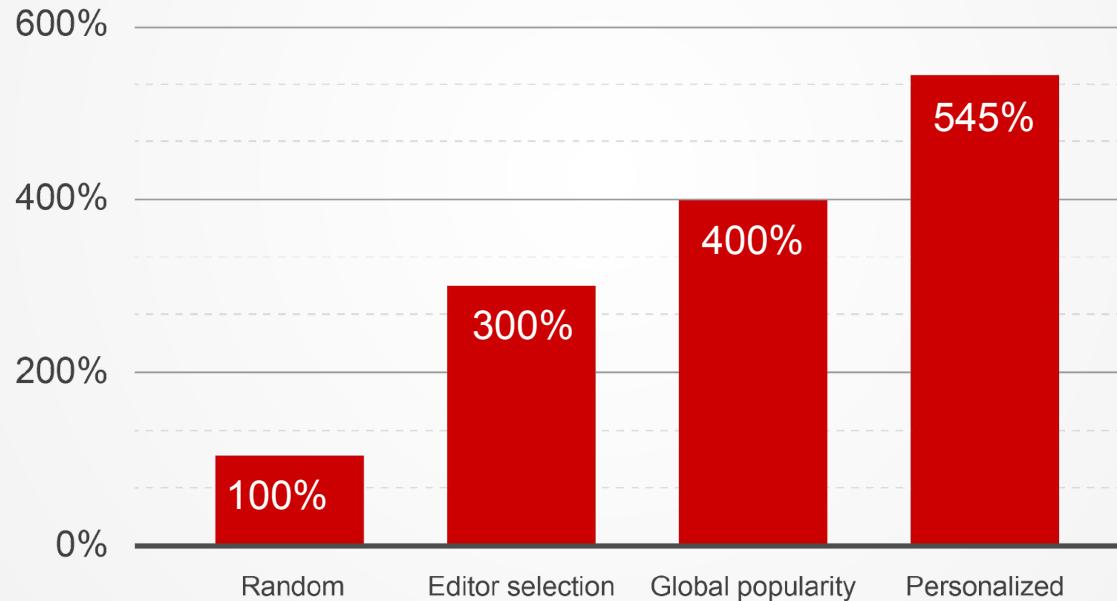
**~1K**

Experiments  
per year



# Algorithm performance

## Click-through-rate



# Agenda

## Theoretical part

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

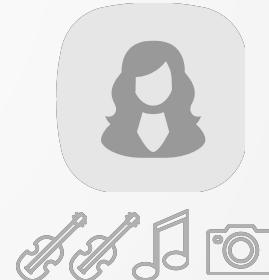
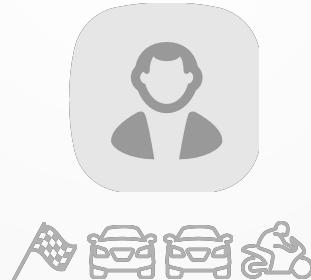
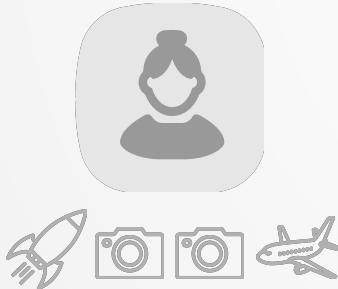
## Practical part

- 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

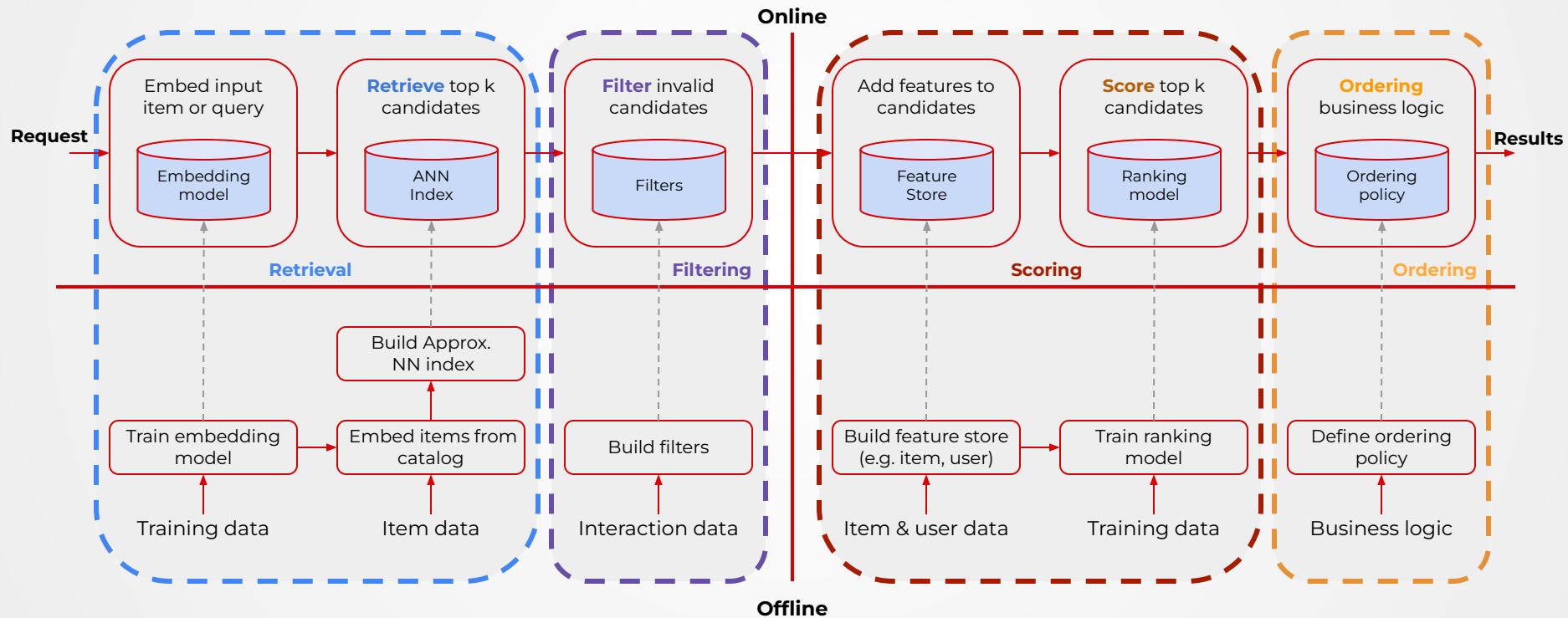


# Recommender systems

- 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>



# practical part prep: notebook #003

6) go open the “003\_user\_representation\_embedding”

7) copy

8) launch

6.

Workshop on Recommendation systems and user representations

Motivation

Following tutorial tries to illustrate typical production setting in a company which provides multiple types of content and have various recommender systems in place.

Single recommender system is typically focused only on the part of the content portfolio in which it operates and will not have access to full user interaction sequence - which is typically due to RAM/CPU/storage/logistics restrictions.

As an example we can take look at the [seznam.cz](#) page on which we can find recommended articles for website novinky.cz.

At the same time user might have visited different website in Seznam's ecosystem e.g. [zbozi.cz](#) which has its own set of recommender systems which serve various purposes and implicitly do not model user behavior in Seznam's ecosystem.



In order to achieve efficient personalization across many models one needs to create efficient user representation which might also alleviate [cold-start problem](#).

Following tutorial tries to:

- illustrate various techniques which might yield such efficient user representation
- demonstrate how such representation might be used in ranking model which is an essential part of recommender system

Agenda

1. At first we will download and prepare [MIND dataset](#) by running notebook [001-prepare-dataset](#)
2. At the next phase we will investigate attributes of newly created dataset by running notebook [002-explorative-data-analysis](#)
3. User representation will be computed in notebook [003-user-representation-embedding](#)
4. At the end we will train [DCN model](#) on created user representation by running notebook [004-train-dcn-model-user-representation](#)



# practical part prep: notebook #003

6) go open the “003\_user\_representation\_embedding”

7) copy

8) launch

7.

The screenshot shows a Google Colab notebook titled "Kopie sešitu 003-user-representation-embedding.ipynb". At the top, there are two buttons: "+ Code" and "+ Text". To the right of "+ Text" is a "Copy to Drive" button, which is circled in red. An arrow points from the number 7 on the left towards this button. The main content of the notebook is organized under a section titled "User representation". It contains the following text and bullet points:

As was already mentioned earlier it is impossible to supply every recommender system with all raw data coming from other systems due to computational limitations and certain amount of compression is therefore required

- Also supplying recommender system with additional data will allow us to personalize model and help us deal with user [cold-start problem](#)
- We will try to create user representation and generally describe methods for obtaining user dense representation also known as embedding from user page visits

Supervised approach:

- Assume that you own large portfolio of various websites and you could categorize websites into the following categories: sport, news and tabloid
- Assume user  $U$  visited following webpages [PV1](#), [PV2](#), [PV3](#)
- Then one could represent  $U$  as sequence of the following categories: sport, sport, tabloid
- This sequence can be further preprocessed into histogram or viewed as text document and be processed by some NLP technique such as [tf-idf](#)
- Having such information associated with user or a page will be important for user and item cold-start mitigation
- The main question is how can one receive such classification? One needs to build online classification service that will detect newly created web pages and classify them appropriately

Unsupervised approach:

- In an unsupervised approach one does not need any additional information regarding user content labels - only sequence of visited webpage views for every user
- Assume that user  $U$  has visited pages PV1, PV2, PV3 then we can treat user as a document and visited web pages as words and create word embeddings by using NLP techniques
- One can then represent  $U$  as a sequence of vectors which can be further aggregated into single vector
- There are many NLP libraries available:
  - [fasttext](#)
  - [starspace](#)

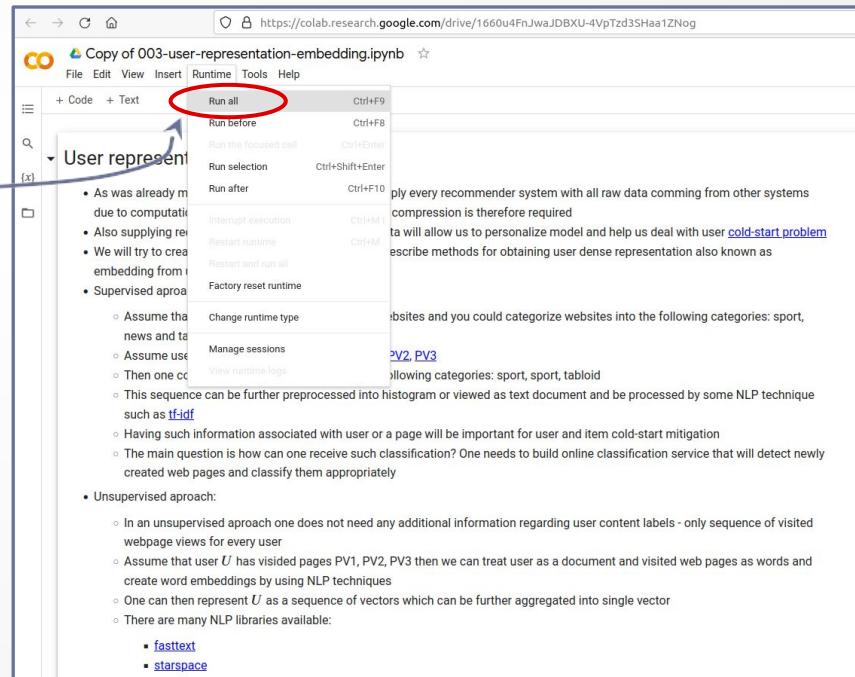
# practical part prep: notebook #003

6) go open the “003\_user\_representation\_embedding”

7) copy

8) launch

8.



# Agenda

## Theoretical part

- Advance the notebooks
- 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

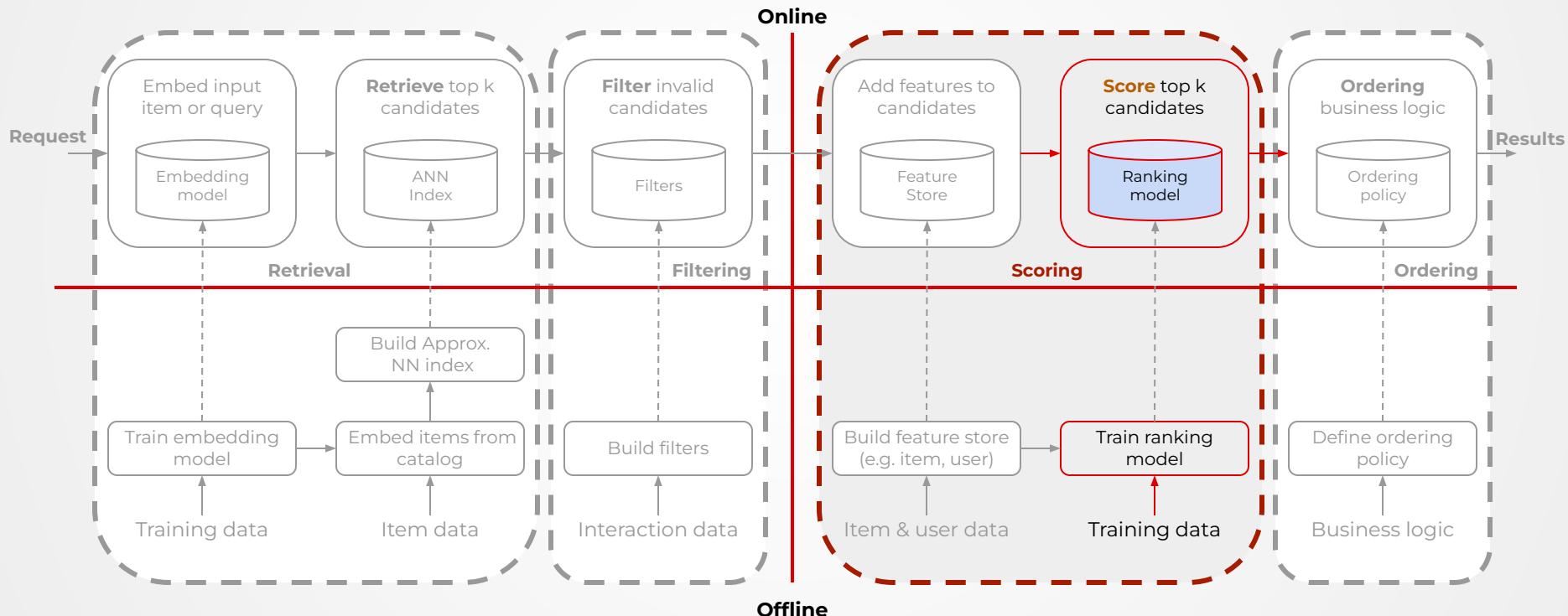


# Practical part

- ✓ 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

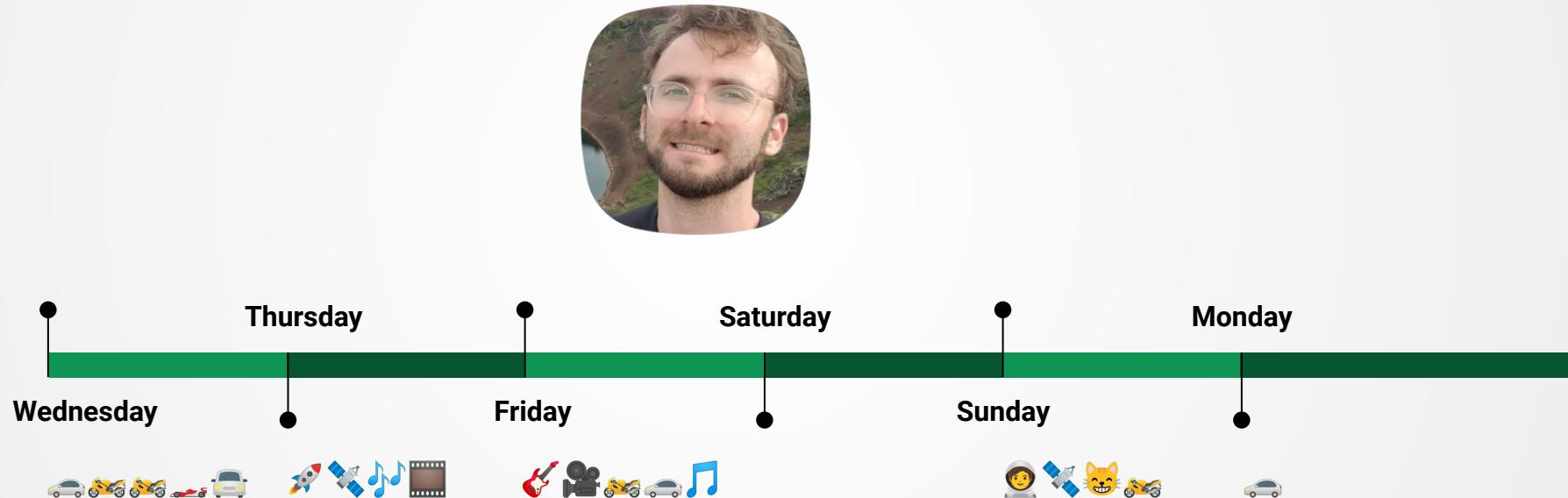
**Age:** 30s

**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 input - contextual features



- currently displayed item features
- current time of day/day of week
- time since last interaction
- current section
- ...

# Ranking models output - label



  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

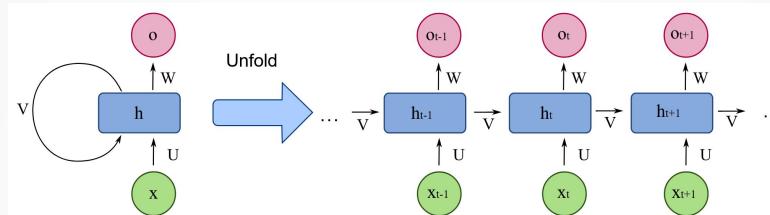
- Based on any features available
- Incl. content-based
- Examples
  - Logistic regression (Vowpal Wabbit)
  - Simple NN
  - Factorization Machines
  - ...

*Advanced neural network models*

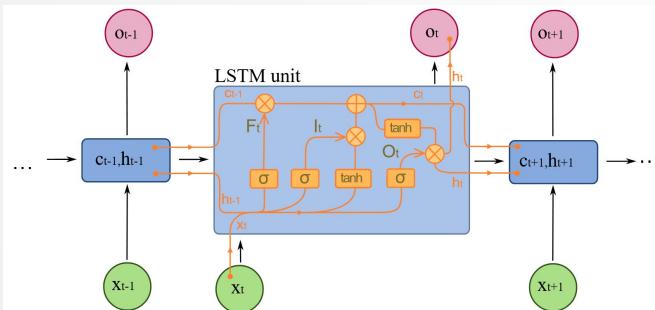


# Ranking models - preliminaries

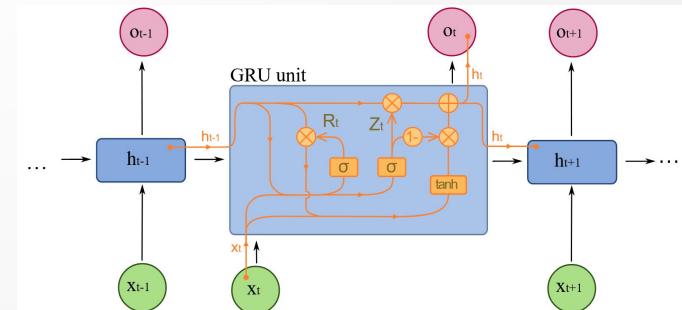
## RNN - Recurrent Neural Network



## 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

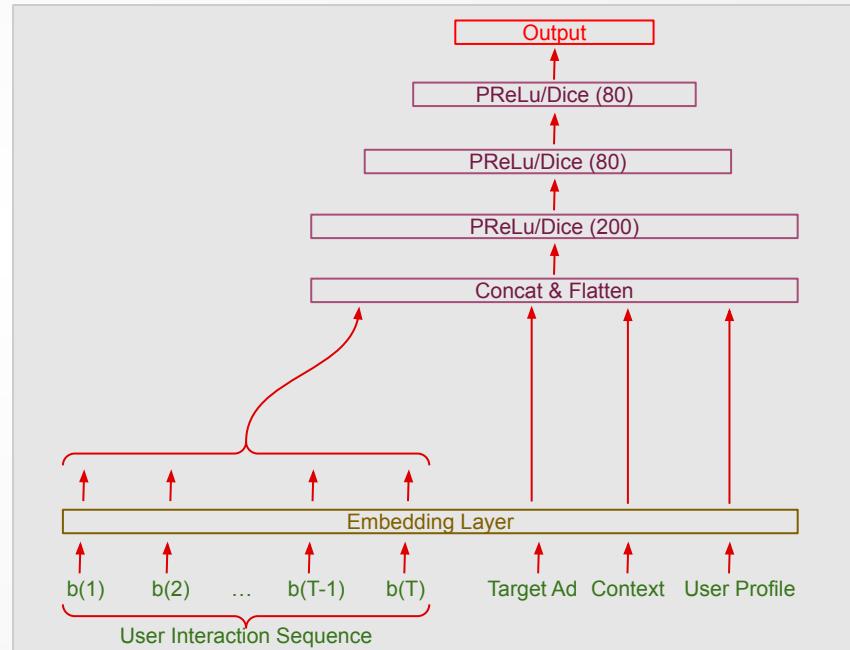
- Mainly focused on the core, usually accompanied by standard DNN
- Ordered from least to most successful attempts:
  - DIEN
  - SLi-Rec
  - SUM
  - **DCN**



# Deep Interest Evolution Network (DIEN)

## Embeddings

- + Interest Extractor
- + Interest Evolution Model
- + MLP



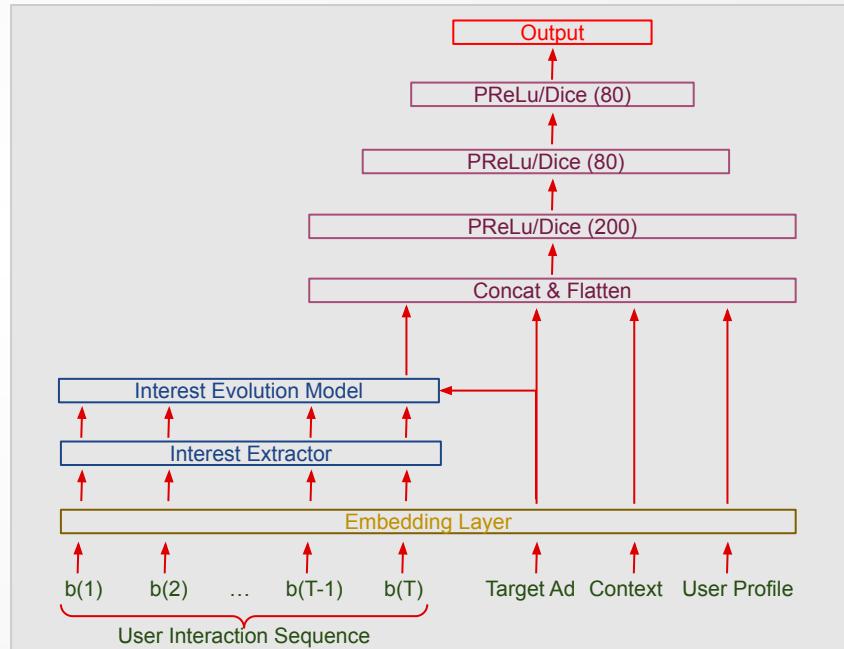
# Deep Interest Evolution Network (DIEN)

## 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 - Short-term and Long-term preference Integrated RECommender 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



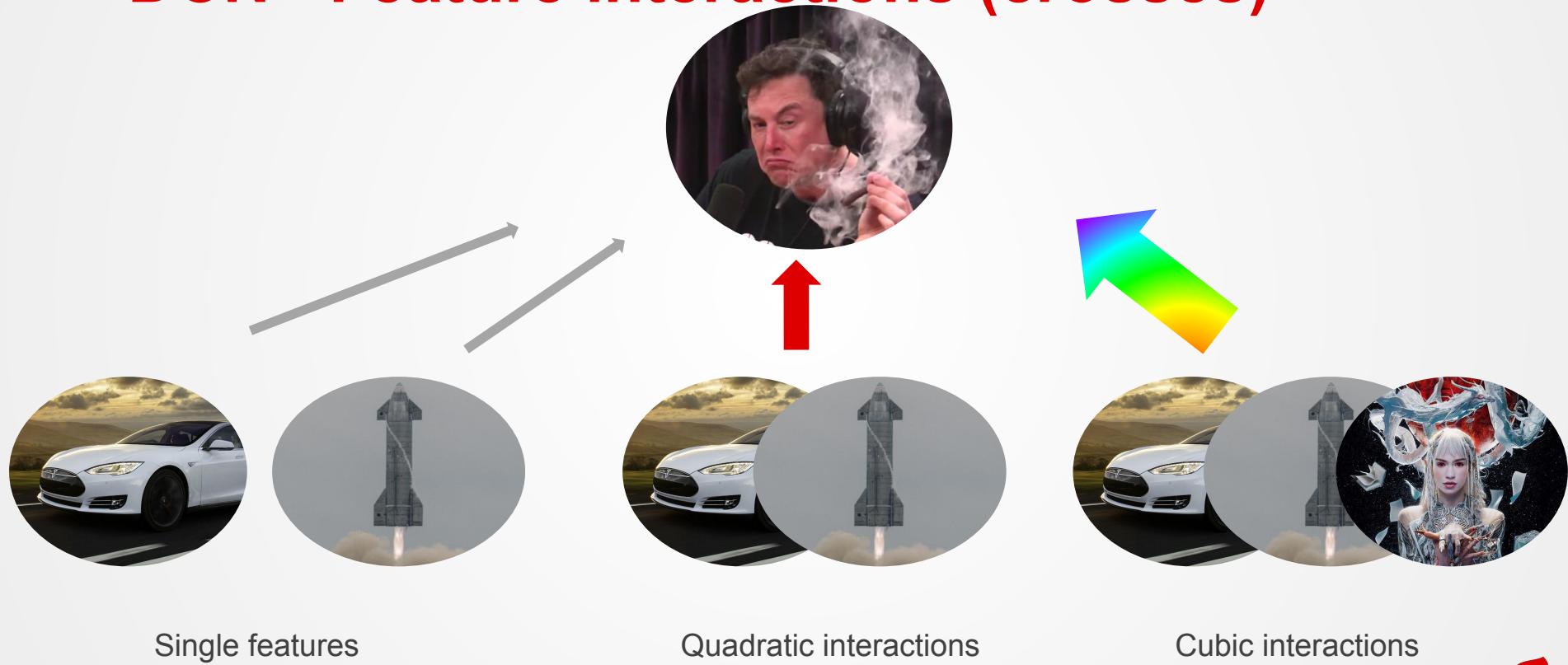
## Pros

- 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)



# DCN - Deep & Cross Network v2

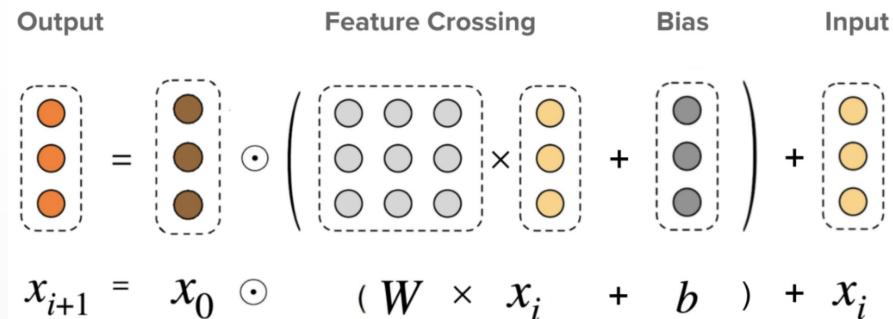
- Implicit f.i. modelled by DNN are not efficient
- Explicit f.i. were formerly hand-designed, now created by cross-layers
- Each cross-layer adds an order of f.i.

## Pros

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

## Cons

- Not sequential



# Agenda

## Theoretical part

- Advance the notebooks
- 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



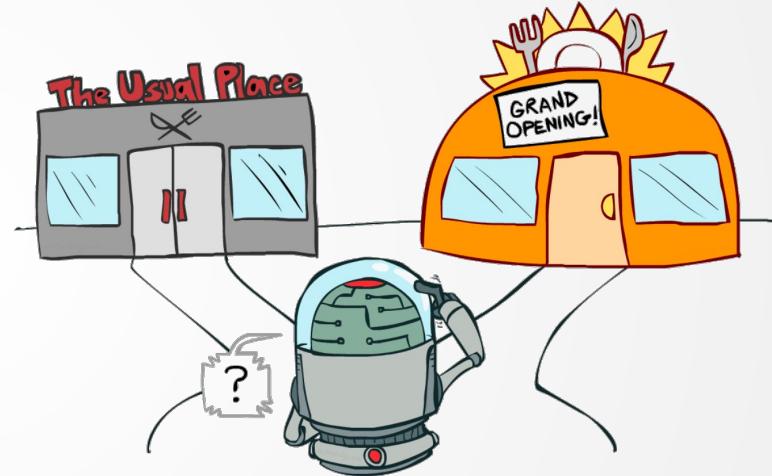
# Cold start mitigation

How to cope with cold start?



# Cold start mitigation

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

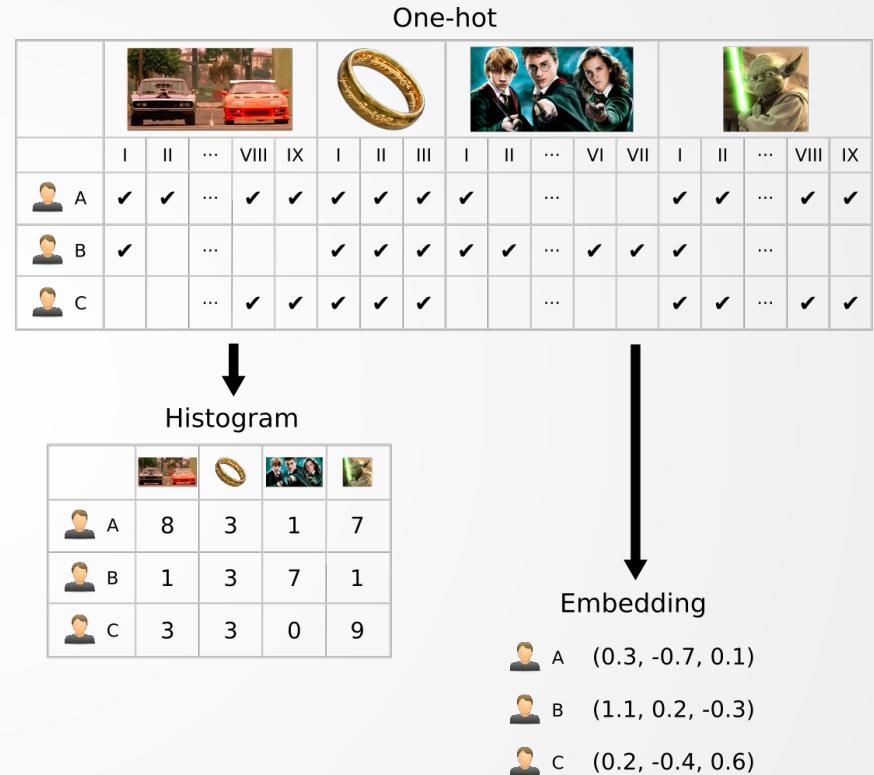


Zdroj:

<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



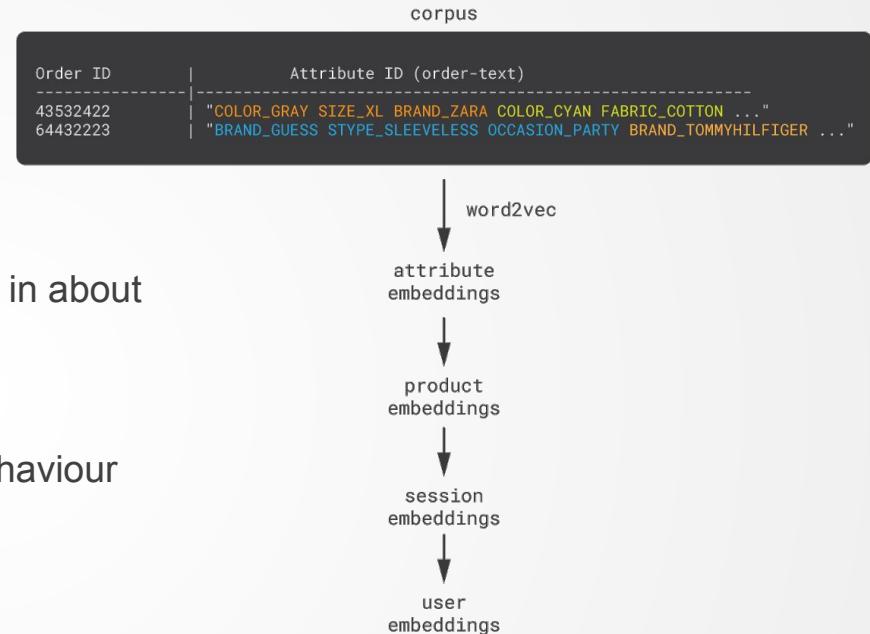
# User embedding

- Embedding (NLP)

- Classification, translation
- Word similarity  $\Leftrightarrow$  vector similarity
- Very efficient - can process english wiki in about a day

- User history  $\Leftrightarrow$  Text document

- Model user similarity using common behaviour
  - User actions  $\Leftrightarrow$  words
  - Users / sessions  $\Leftrightarrow$  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/>



# 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](mailto: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-doporucovacni-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-e-commerce-growth-decade/>  
[https://faculty.washington.edu/jdb/345/345%20Articles/Iyengar%20%26%20Lepper%20\(2000\).pdf](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](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:

<https://alammar.github.io/illustrated-word2vec/>  
<https://fasttext.cc/>  
<https://medium.com/wisio/a-gentle-introduction-to-doc2vec-db3e8c0cce5e>  
<https://ai.facebook.com/tools/starspace/>





Please leave us your feedback:

<https://forms.gle/s3cEEDV1hoQ16fLEA>