

IT akademie 2022:

Big Data a Machine Learning v praxi



Vít Líbal, Tomáš Nováčik
Výzkum reklamních systémů

SEZNAM.CZ

Obsah

O Seznamu

Když navštívíte Seznam.cz

- Časová linie:
 - čas nula: proklik
 - prvních sto milisekund
 - hodina poté
 - po roce



Na čem se v Seznamu pracuje

- Řešené úlohy a nástroje
- Výzkumné týmy
- Role a otevřené pozice

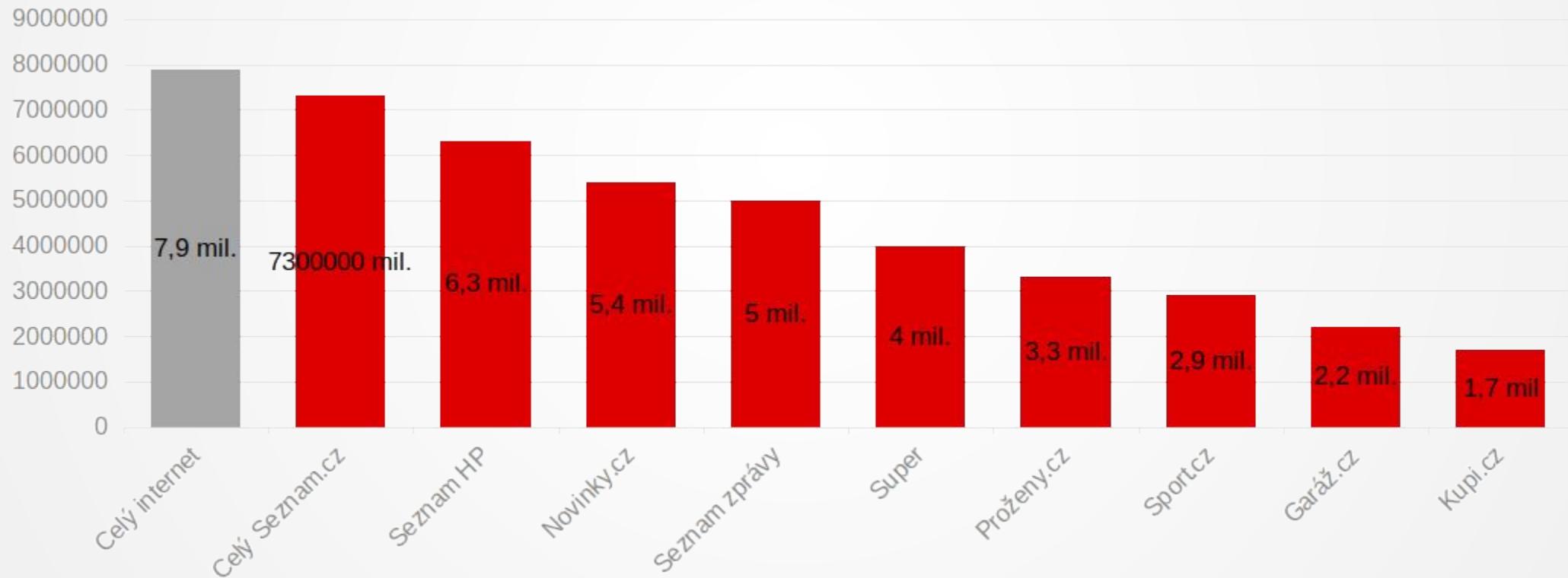
Hrajeme si s daty:

- EDA - Exploratory Data Analysis, aneb průzkum dat
- Klikne, nebo neklikne? – Modelování pravděpodobnosti prokliku
- Modelování s knihovnou Vowpal Wabbit



Seznam.cz - návštěvnost webů

Celkový počet uživatelů vs. uživatelé Seznamu



Zdroj: Netmonitor, Duben 2021



Seznam.cz - čísla



95 %
zásah české
internetové populace

8,4 mil.
zásah unikátních
uživatelů měsíčně*

78 %
přihlášených
uživatelů denně

1500+
zaměstnanců v 9
pobočkách

*Zdroj: Mediaguru, NetMonitor-SPIR-Gemius, 1. 1. – 30. 6. 2021, průměrně za měsíc, celý internet, všechny platformy



@Seznam.cz



S



➤ chce získat informace

- hledá, třídí, čte obsah
- přistál na hledané stránce



SEZNAM.CZ

...najdu tam, co neznám

Vyhledat

Koronavirus

Pozitivní případy V nemocnicí Úmrtí • Aktuální opatření
+278 -11 +1 Reinfekce: 59 202 40 325 • Cestování

Sz Seznam Zprávy • Čtvrtek 7. července Svátek má Bohuslava.

Fialův plán, když Putin odříznou Evropu: Kdo bude mít plyn, at' se rozdělí

Udržet Evropu jednotnou. Úkol, který si vytvořil Česko pro příští půlrok v čele Rady EU. Inflace, možný nedostatek...

Video: Ruský voják ukázal, jak nelikvidovat protiletadlové střely S-300

Krutý osud Haiti: Nejhudší země světa financovala francouzské banky

Nejnebezpečnější jsou parabolické jezy, říká lektor vadotví

Novinky

Poslanci zrychleně změnili zákon o evidenci skutečných majitelů

Babiš se postavil do poslance STAN Bernarda. Bureši, jste zloděj, práškač a lhář, opáčil Bělorusko hrozí údery na Polsko

U Černého jezera se utáborili němečtí turisté, dojeli si tam i přes zákaz autem

Časté důvody, proč lidé přerušují kontakt se členy rodiny

Volali policii k rozkládající se mrvotle. Nález hildku pobavil

Turci nechali odpout ruskou lodě s ukrajinským oblibím, Kyjev zuří

Skandál v Nizozemsku: Policisté stříleli na protestující farmáře ostrými

Super

Herečka z populárního seriálu oplakává manžela

(†33): Při výjízdce na lodi ho zabil blesk

Herečka Bevin Prince (39), nejvíce proslulá seriálem One Tree Hill, který v Americe běžel v letech 2003-2012, prožila tragickou smrt

Jiří Dvořák vyděl ve Varech novou přítelkyni: Krásnou, o 27 let mladší brunetku

Můj Zamilovaná Agáta Hanychová se pochlubila další fotkou s Jaromírem Soukupem. Místo Varů spolu vaří

Sport

Djokovičova podezřelá inhalace ve Wimbledonu. Čím si během zápasu pomáhá?

Bizarní návyk tenisové hvězdy Novaka Djokoviče zadělává během Wimbledonu u fanoušků na podezřívavé spekulace...

Stream

Videorozbor: Manžela přepadli, tak vytáhla pistoli a začala střílet Chokotopus rozbroy

Zobrazit všechna rádia

RÁDIO KISS

To nejlepší od 90. let

Republiku

17 °C Večer 13 °C V noci 18 °C Zítra

Rádio

Firma

Přidat firmu

Reklama • Koupit reklamu

Máte plynový kotel?
Až 80 000 Kč dotace od Maliny

Výměnu na tepelné čerpadlo místo státu zadotujeme my!

Chci čerpadlo

Služby

Akce / Letáky Email Profi Podcasty Recepty

Auto / Moto Hry Pohádky Prohlížeč Slovník Stream

Bazar Jízdní řády Rádio / Klasika Volná místa

Deníky Mapy Mobilní aplikace Reality Zboží / Magazín

Dovolená

Počasí • Česká republika • Predpověď větru

17 °C Večer 13 °C V noci 18 °C Zítra

Rádio

RÁDIO KISS To nejlepší od 90. let

Republiku

1 Frekvence Country 100.0 BEAT RADIO 100.0 Bonton 100.0 Signál 100.0

Zobrazit všechna rádia

Firma

Přidat firmu

@Seznam.cz

- chce získat informace
 - hledá, třídí, čte obsah
 - přistál na hledané stránce

The screenshot shows a web browser window with the URL <https://search.seznam.cz/?q=návratnost+fotovoltaiky&oq=návratnost+fotovoltaiky&aq=1&ms=1>. The search term 'návratnost fotovoltaiky' is entered in the search bar. The results page displays several search results related to solar panel returns and installations:

- Návratnost Fotovoltaiky - Dotace až 50 %**
bce.cz/ Reklama
Postavíme vám fotovoltaickou elektrárnu na klíč. Zajistíme vše od dotací až po montáž.
- Fotovoltaika na rodinný dům - Pro domácnosti i firmy**
malinagroup.cz/fotovoltaika Reklama
Nabízíme nejvýnosnější panely na trhu díky naši nejvyšší řadě celočerných panelů Gen2. Solární systém na mru. Volejte 800 880 878. Nejdělsí záruky na trhu. Instalace do 5 měsíců
- Fotovoltaika na klíč - Nyní s dotací až 225 000 Kč**
ilios.cz/fotovoltaika Reklama
Na Vaši elektrárnu Vám zajistíme dotaci až 225 tis. Kč. Vyřídíme veškerou dokumentaci. Rychlá návratnost · Garance ziskání dotace · Česká rodinná firma · 7000+ zapojených panelů · Absolonova 1071/25a, Brno
- Fotovoltaická elektrárna - Fotovoltaika pro váš dům**
fotovoltaika.tipa.eu/ Reklama
Vše zařídíme na klíč včetně dotace. Úspora za energie až 80 %. Nezávazná konzultace zdarma
- Rozumná návratnost domácí fotovoltaiky? Pozor na správný...**
estav.cz/9418.rozumna-navratnost-domaci-fotovoltaiky-pozor-na...
-  **10. 3. 2021** · Zvažujete fotovoltaiku na svém rodinném domě? Jakou bude mít návratnost, záleží i na střídači. Střídač schopný asymetrické dodávky elektřiny je v českých podmínkách klíčový pro zajištění návratnosti domácí...
- Jaká vás čeká návratnost u fotovoltaické elektrárny? | Voltair**
voltair.cz/jaka-vas-ceka-navratnost-u-fotovoltaicke-elektrarny
Doba navrácení financí se odvíjí od spotřeby, ceny za elektřinu, od velikosti fotovoltaiky i od vašeho zvoleného způsobu akumulace.
- Návratnost fotovoltaiky: kdy vám ušetří statisice a kdy se...**
oze.tbz-info.cz/21585-navratnost-fotovoltaiky-kdy-vam-usetri...
 **10. 12. 2020** · Pokud uvažujete o střešní solární elektrárně, je právě tento nejlepší čas na její pořízení. Přes zimu vyříďte instalaci a od jara už můžete využívat vlastní energii. Logicky ale přichází otázka, jestli se investice do...
- Návratnost fotovoltaiky při zdražování energií | BCE.cz**
bce.cz/zdrazeni-energi-a-navratnost-fotovoltaiky
Jak rychle se vrátí náklady na solární panely? Jak se do toho promítně každoroční zdražování energií?

Obrázky > Návratnost fotovoltaiky

@Seznam.cz

- chce získat informace
 - hledá, třídí, čte obsah
 - přistál na hledané stránce



https://woltair.cz/blog/fotovoltaika/jak-a-vas-cek-a-navratnost-u-fotovoltaick-elektrarny

WOLTAIR

Tepelná čerpadla Fotovoltaika Nabíjecí stanice Servis Dotace Spolupráce & školení O nás Kontakt Přihlásit se

Zpět na články

Jaká vás čeká návratnost u fotovoltaické elektrárny?

Olga Knoflíčková
Expert
09. září 2021



Investice do solární elektrárny je v dnešní době častějším a častějším řešením, se kterým lidé šetří za energie, stávají se soběstačnými a ekologickými zároveň. Jak zjistit, kdy se vám vrátí vložené peníze?

Běžná doba návratnosti fotovoltaické elektrárny bývá 7-10 let, což ale nemusí být pravidlo. Někdo se může s elektrárnou sžít natolik, že se mu vložené peníze vrátí už za 5 let. Stává se tak u návrhů, které jsou síté na míru zákazníků.

Vyšší kvalita bydlení

S fotovoltaikou se stanete nezávislými. Hlavní rada zní – co nejvíce vyrobené elektřiny spotřebovat ve vašem objektu. S tím souvisí i to, jakým způsobem sladíte život s vás.



0 ms



time



Putinova realita se může zhroutit jako domeček z karet. Ruská veřejnost je rozpolcena, mladší ročníky invazi příliš nepodporují

Před 19 hodinami

Putin postavil svou legitimitu na myšlence obnovení stability, prosperity a globálního statusu Ruska. Válka na Ukrajině ...

 Libí se 90  Komentáře 21

 EnergetikaOnline



Nově schválené dotace podpoří malé fotovoltaické elektrárny

Před 23 hodinami

Ministerstvo životního prostředí schválilo dotace šesti desítkám menších fotovoltaických elektráren, které vznikn...
 Libí se 5  Komentáře

 Forum 24



Zmařili jsme atentát na ruského moderátora, řekl Putin. Obvinil Západ a Kyjev

Před 22 hodinami

Ruská tajná služba FSB podle ruského prezidenta Vladimira Putina zmařila pokus Kyjeva a Západu zavraždit známého...
 Libí se 29  Komentáře 3

 Betonbau



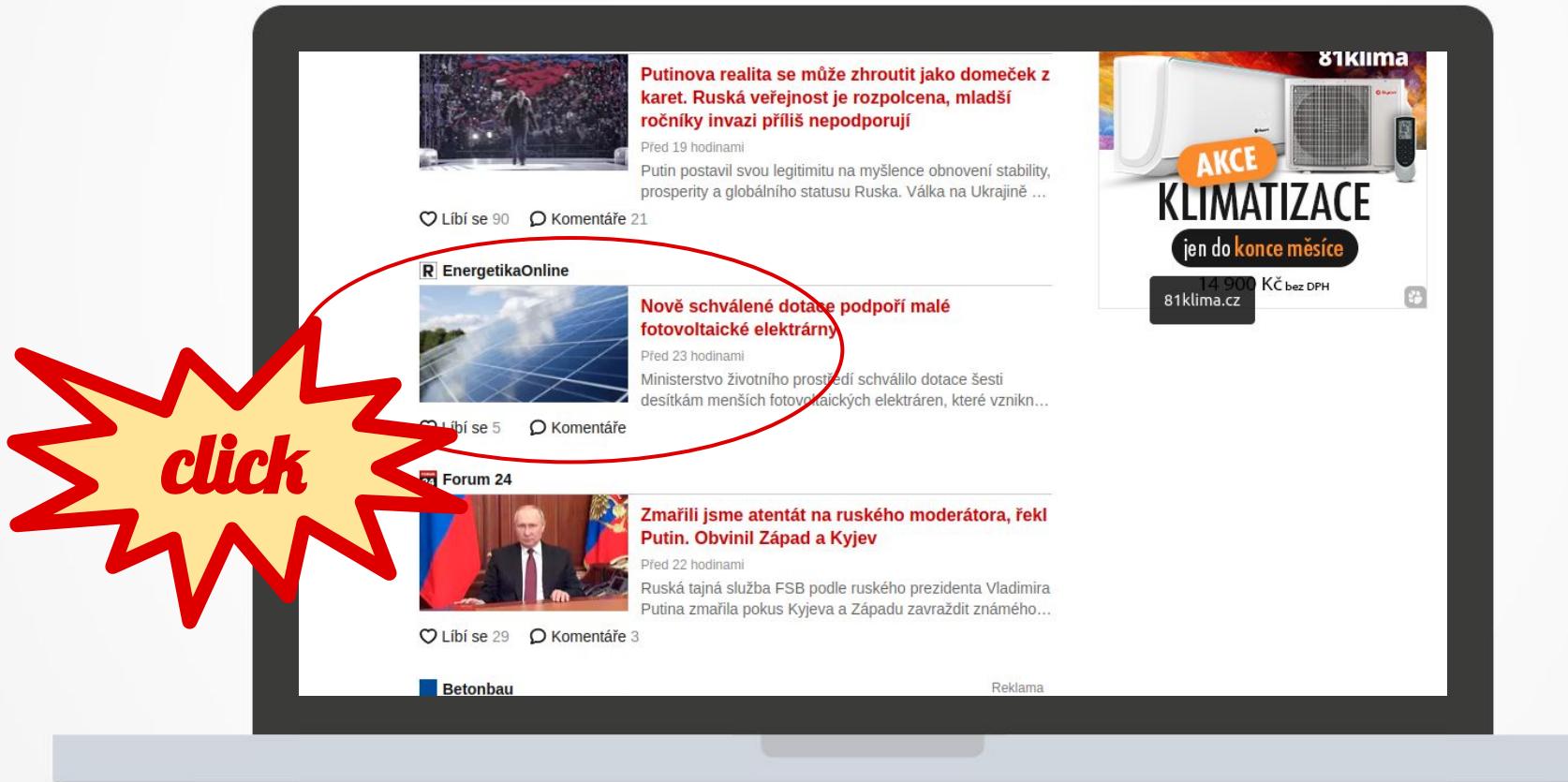
AKCE KLIMATIZACE
jen do konce měsíce
14 900 Kč bez DPH
81klima.cz

Reklama



0 ms

time



0 ms

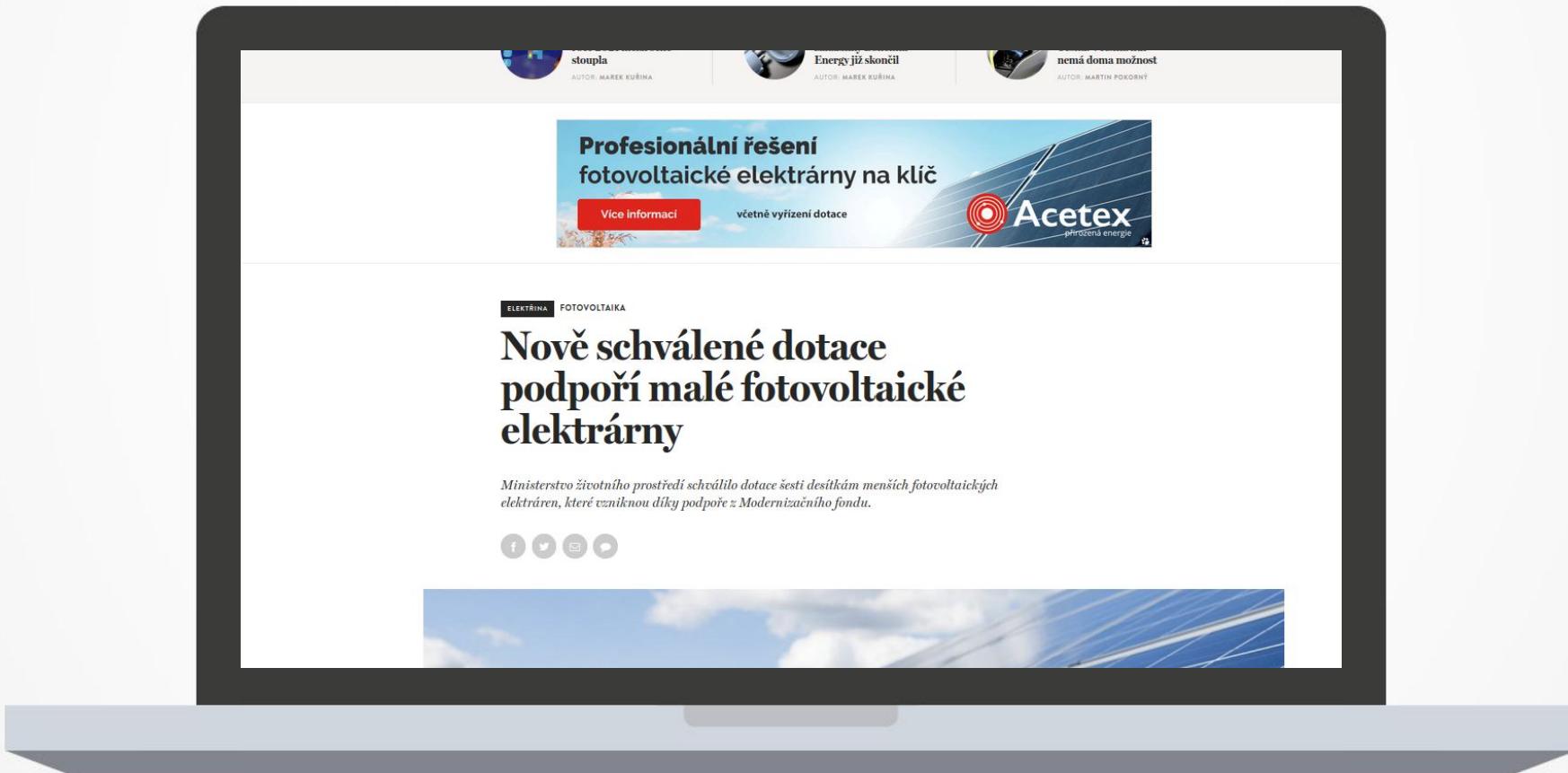


- klik → požadavek zobrazení obsahové stránky

time



0 - 100 ms



0 - 100 ms



- klik → požadavek zobrazení stránky
- odpověď serveru → stránka zobrazena

time



0 - 100 ms



- klik → požadavek zobrazení stránky
- odpověď serveru → stránka zobrazena
 - vlastní obsah stránky
 - odkazy na doporučené položky
 - reklamy

time

https://energozroutei.cz/z/co-strecha-to-solarni-elektrarna-ekologove-vyzyvaji-k-odchodu-od-plynu-z-ruska?utm_source=www.seznam.cz&utm_medium=sekce-z-

The screenshot shows a news article titled "„Co strecha to solární elektrárna“. Ekologové vyzývají k odchodu od plynu z Ruska". The article discusses the transition away from Russian gas. The page includes a header with various products, a sidebar with advertisements for Toyota RAV4 and a gas boiler, and a sidebar with the most viewed articles.



Služba doporučování

Doporučování – “pípa na sudu dobrého obsahu”



Služba doporučování



Dočte článek a...

... odejde jinam

- ❖ ztrácíme uživatelu pozornost
- ❖ posílí konkurenci
- ❖ ztrácíme uživatele

... vrátí se na HP

- ❖ ztrácí čas
- ❖ ztrácíme uživatele

... bude pokračovat

- ❖ ušetří čas
- ❖ získáváme uživatelu pozornost
- ❖ získáváme uživatele

0 - 100 ms



Služba doporučování

Uživatel

Gender: Muž
Age: 30s
Location: Vizovice
Device: Android
Interests:      

...

Historie interakcí uživatele



@ D -8

@



D -7

D-5

-4

@

@

@ D-3

Kontext

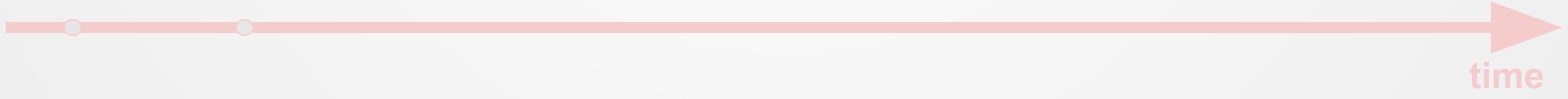
Den v týdnu: pátek
Čas: 22:30
Počasí: jasno
...

Co může chtít číst dále?

Položka

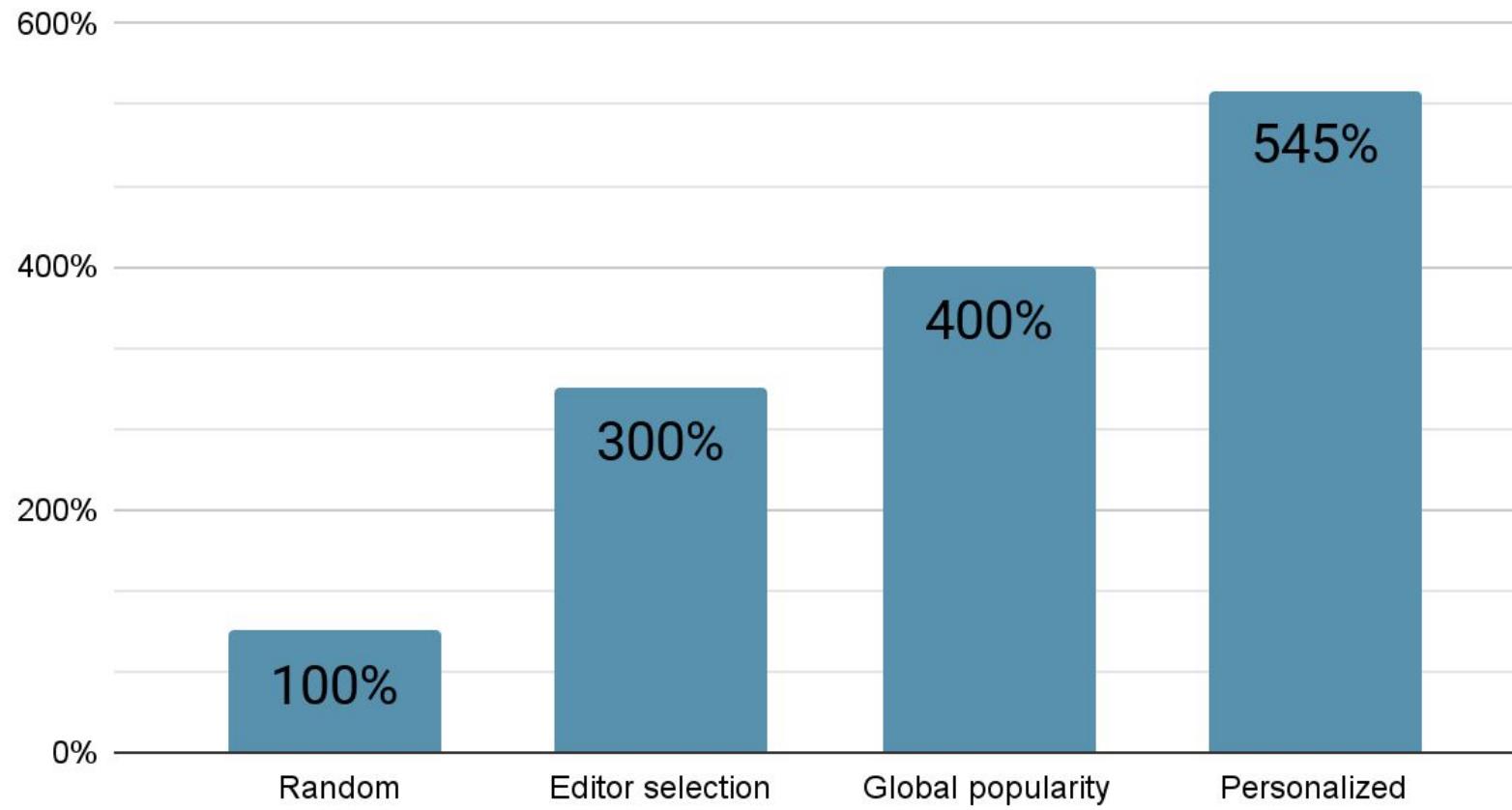
Titul: "Jeden z posledních Wartburgů: Svezli jsme..."
Zdroj: Garáž.cz
Vydáno: 15 hours ago
Tagy: Auta, Historická auta, Německo

0 - 100 ms



Služba doporučování

Click-through-rate



Služba doporučování



~10M

prokliků za den



~10K

požadavků za sekundu



~1K

nových položek za
den

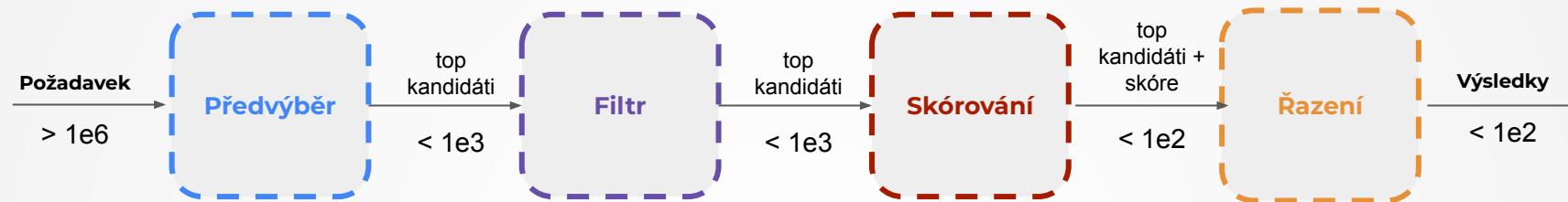


~1K

experimentů za rok



Doporučovací systém



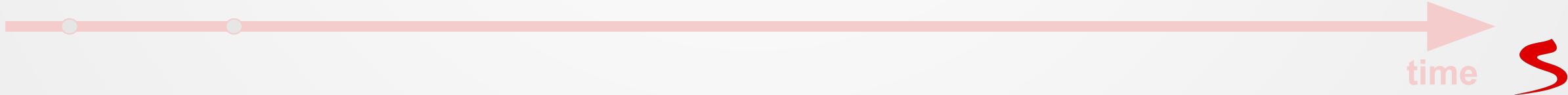
Rychlé zúžení výběru

Business logika zobrazení

Odstranění nechtěných kandidátů

Klasifikace zbylých kandidátů

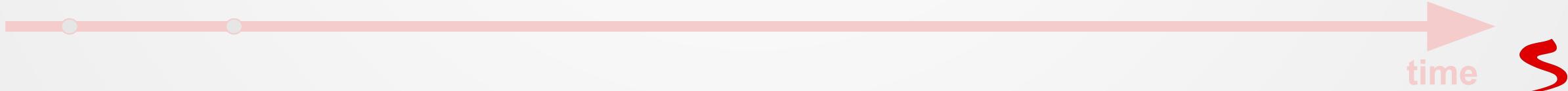
0 - 100 ms



Doporučovací systém

- Data:
- ML úlohy:
 - Predikce prokliku: Jaká velká je šance, že uživatel klikne na link?
 - Predikce dokoukanosti videa: Jaká je šance, že uživatel shlédne video?
 - Detekce clickbaitu: Jaká pravděpodobnost je, že uživatel po prokliku nenajde, co hledá?
- Role:
 - Data Scientist → ML Engineer → SW Developper

0 - 100 ms



0 - 100 ms

- klik → požadavek zobrazení stránky
- odpověď serveru → stránka zobrazena
 - vlastní obsah stránky
 - odkazy na doporučené položky
 - reklamy



time

https://energozroutei.cz/z/co-strecha-to-solarni-elektrarna-ekologove-vyzyvaji-k-odchodu-od-plynu-z-ruska?utm_source=www.seznam.cz&utm_medium=sekce-z-

The screenshot shows a news article titled "„Co střecha to solární elektrárna“. Ekologové vyzývají k odchodu od plynu z Ruska". The article discusses the transition away from Russian gas. The page includes a header with various products, a sidebar with advertisements for Toyota RAV4 and a heat pump, and a sidebar with the most viewed articles.



Služba reklamního systému

Reklamní systém – “platí Váš Internet zdarma”

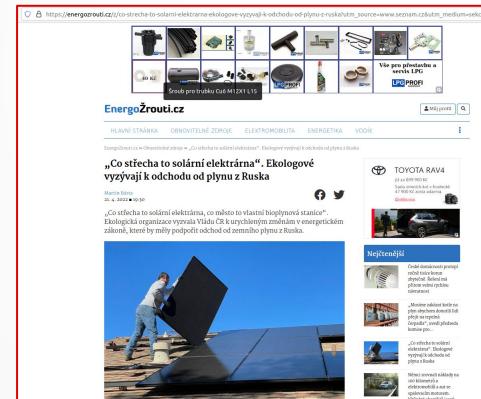


Služba reklamního systému

SKLIK.cz

Uživatel Internetu

Poskytovatel
obsahu



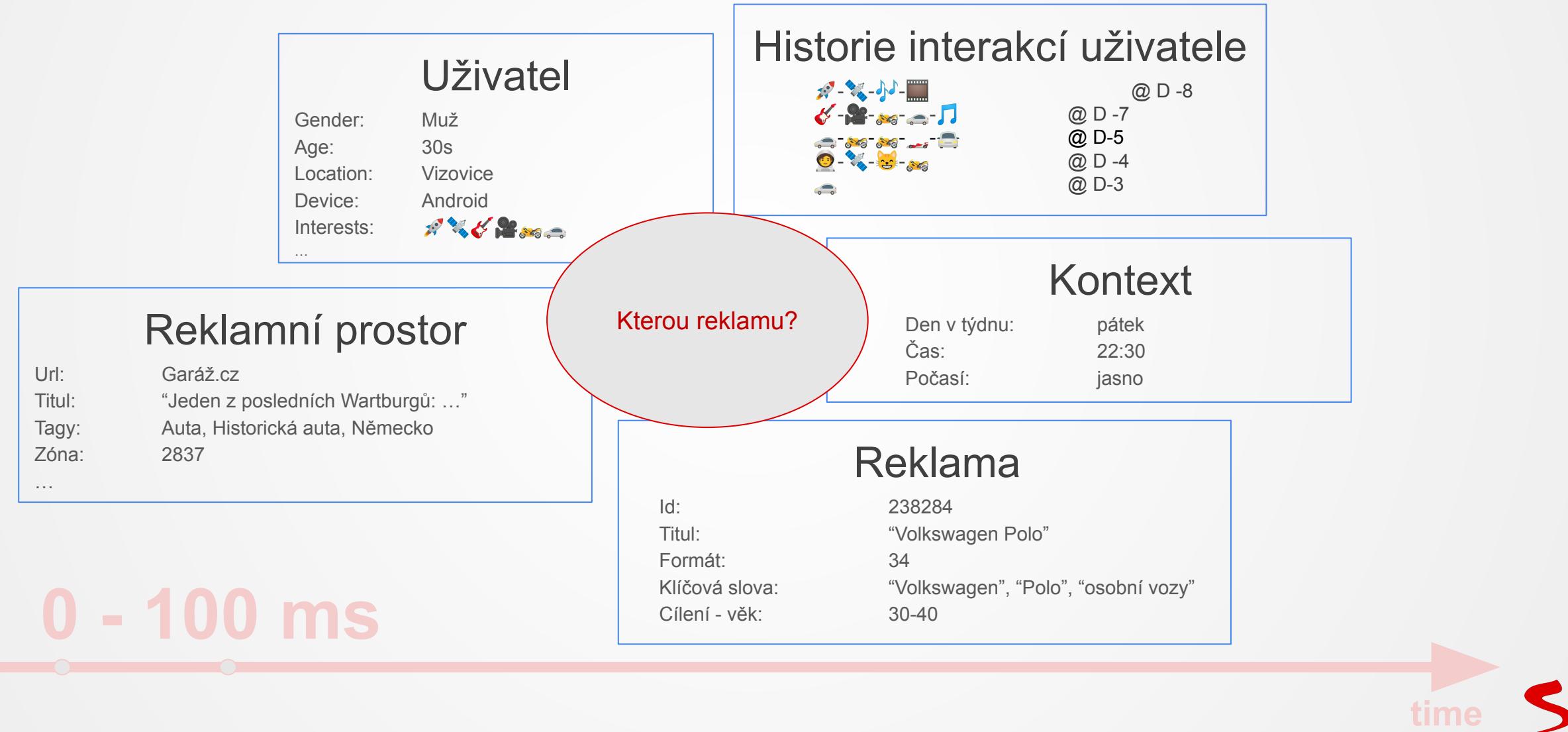
Inzerent

Provozovatel
reklamního systému

S

Služba reklamního systému

SKLIK.cz



Služba reklamního systému

SKLIK.cz



~15K
inzerentů



~500K
reklam



~10K
požadavků za
sekundu

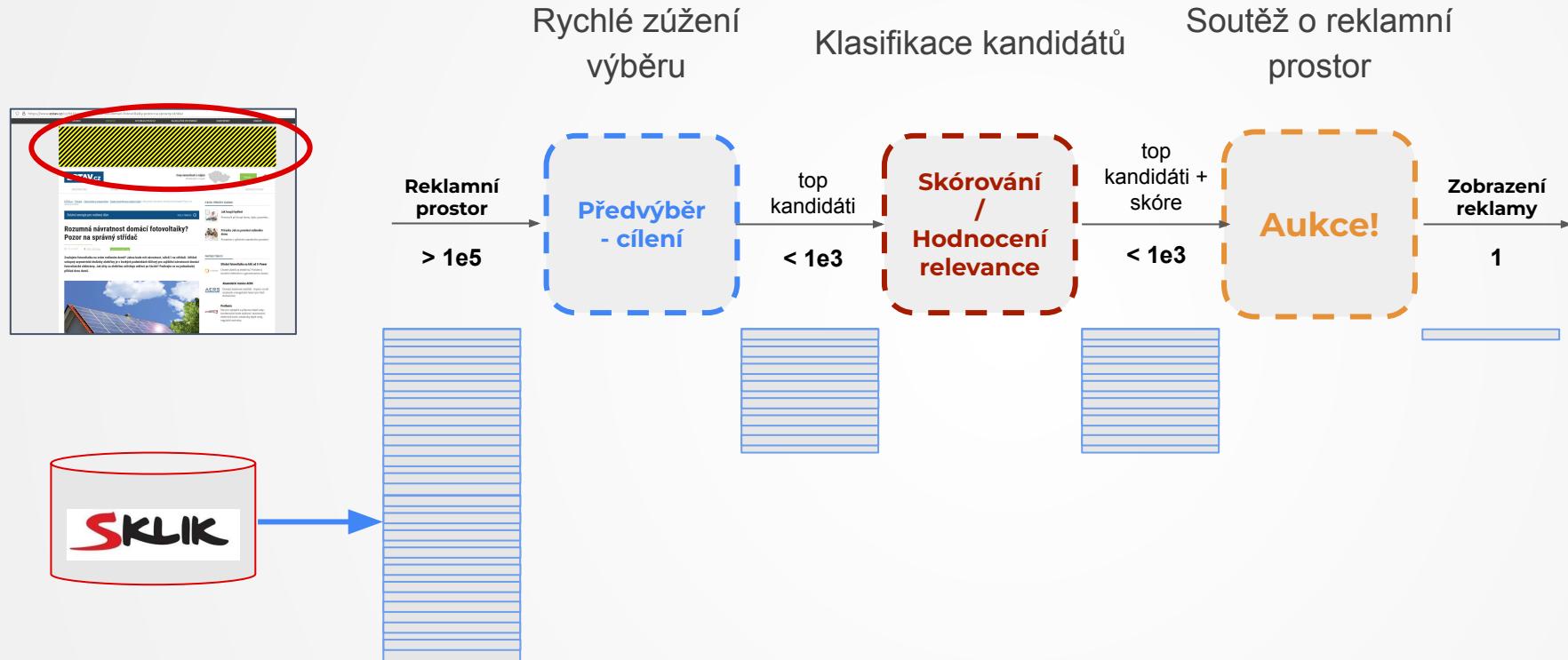


~10M
predikcí za sekundu

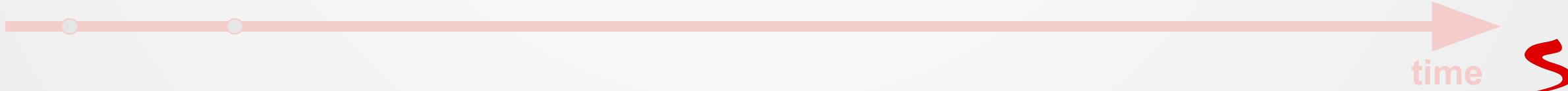


Reklamní systém

SKLIK.cz



0 - 100 ms



Reklamní systém

SKLIK.cz

REAL-TIME BIDDING



0 - 100 ms

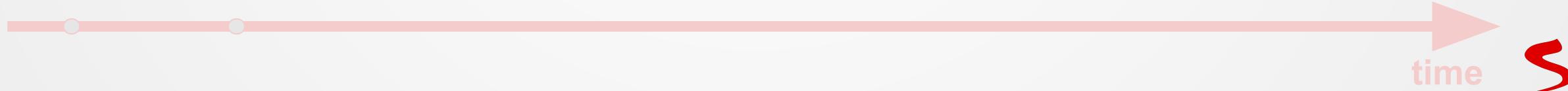


Reklamní systém

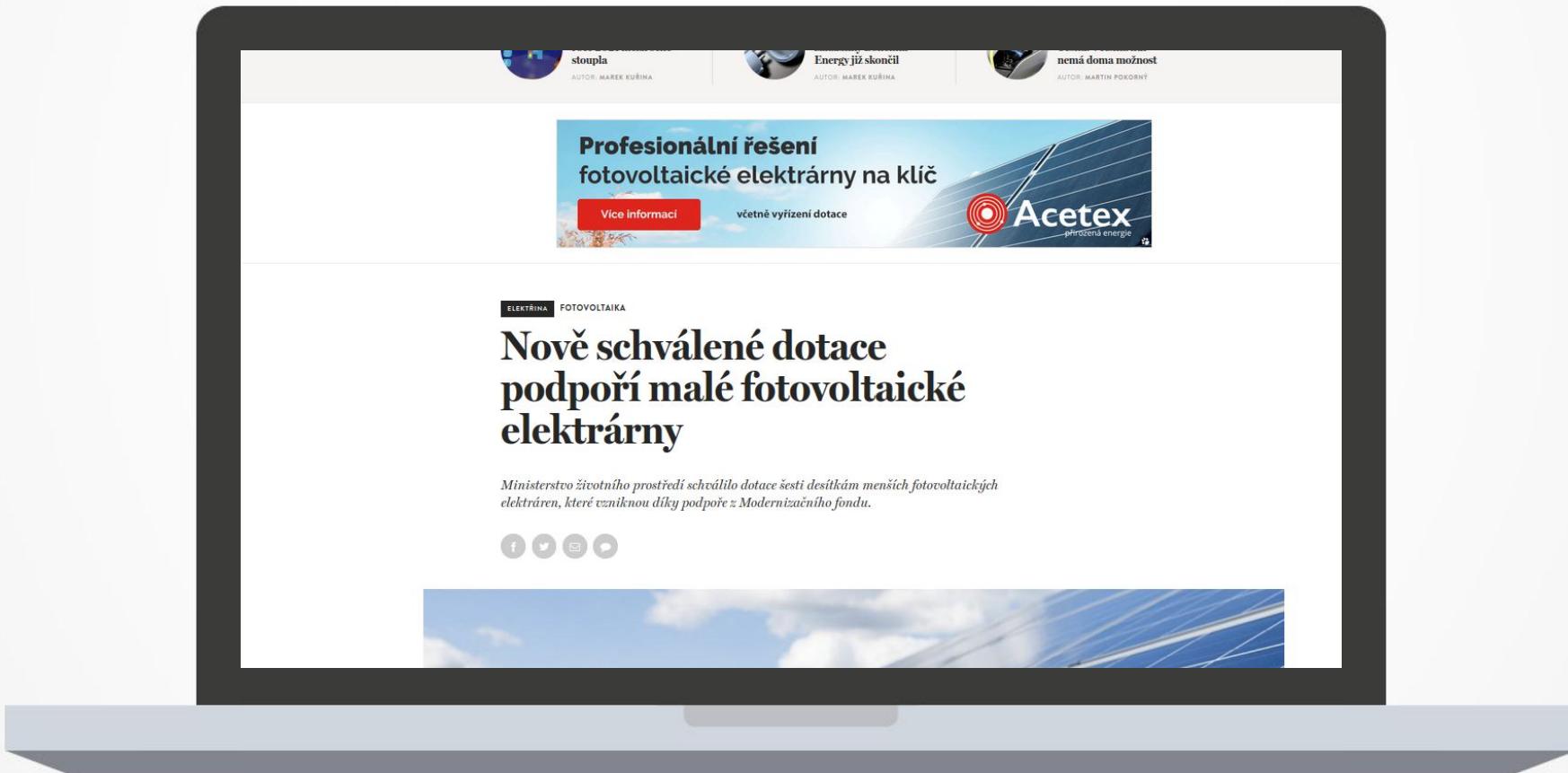
SKLIK.cz

- Data:
- ML úlohy:
 - Predikce prokliku/konverze/dokoukanosti: Jak velká je šance, že uživatel klikne na reklamu, shlédne videoreklamu, zakoupí zboží?
 - Optimalizace výše bidu: Jakou volit výši bidu aby bylo dosažen maximální ROI?
 - Predikce vítězné ceny: Jak nastavit výši bidu aby vyhrál s minimálním nákladem?
- Role:
 - Data Scientist → ML Engineer → SW Developper

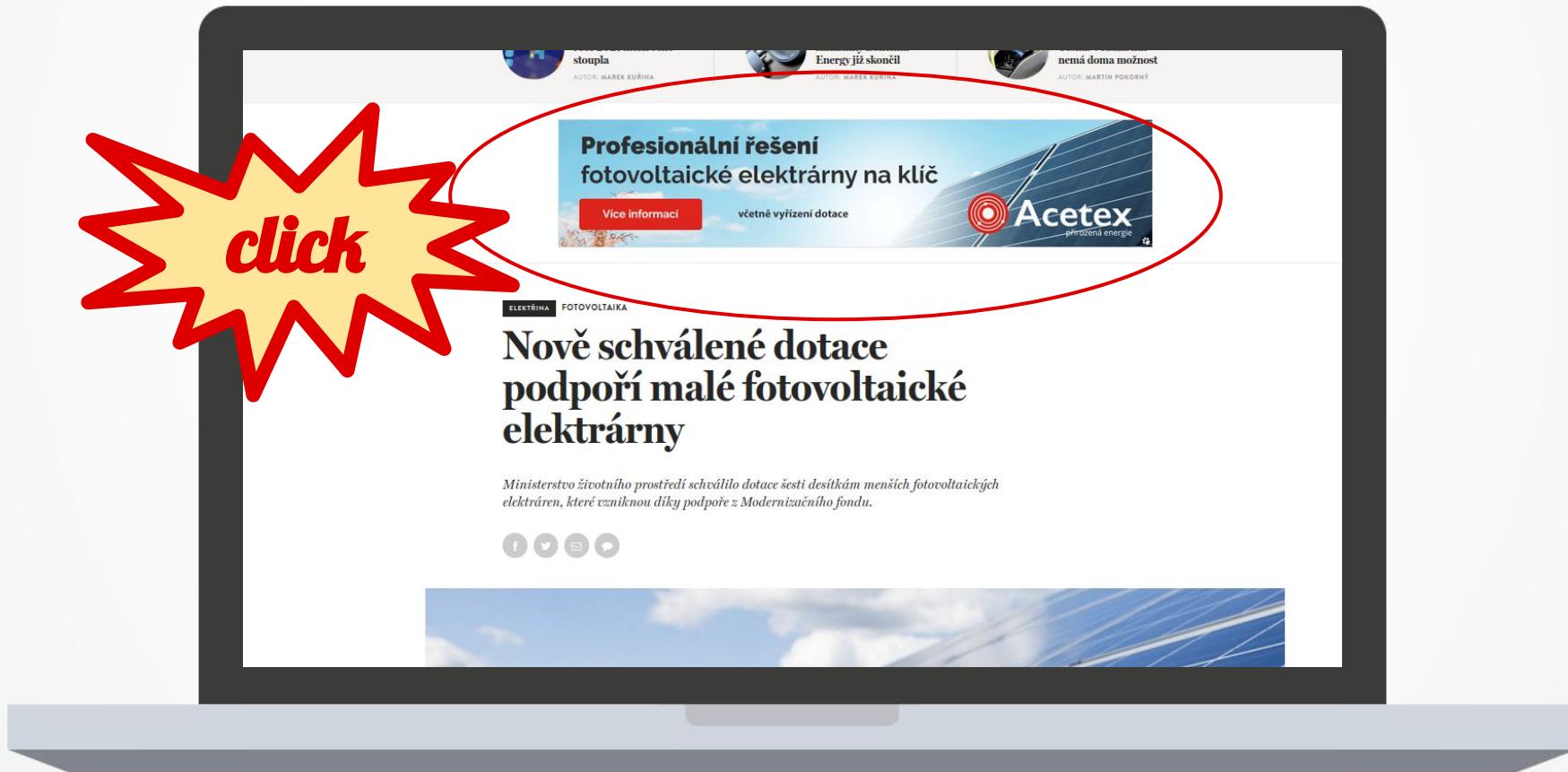
0 - 100 ms



0 - 1 h



0 - 1 h



0 - 1 h



time

- klik → požadavek zobrazení stránky
- odpověď serveru → stránka zobrazena
- klik → požadavek zobrazení stránky inzerenta



0 - 1 h



time

ELEKTRÍNA FOTOVOLTAIKA

Nově schválené dotace podpoří malé fotovoltaické elektrárny

Ministerstvo životního prostředí schválilo dotace šesti desítkám menších fotovoltaických elektráren, které vzniknou díky podpoře z Modernizačního fondu.



0 - 1 h



time

- klik → požadavek zobrazení stránky
- odpověď serveru → stránka zobrazena
- klik → požadavek zobrazení stránky inzerenta
- odpověď serveru → etc.



0 - 1 h



time

- klik → požadavek zobrazení stránky
- odpověď serveru → stránka zobrazena
- klik → požadavek zobrazení stránky inzerenta
 - agregace kliků na reklamy, na články, vyhledávací dotazy
 - kontroly a čištění dat
 - doučování modelů

Doporučovací systém

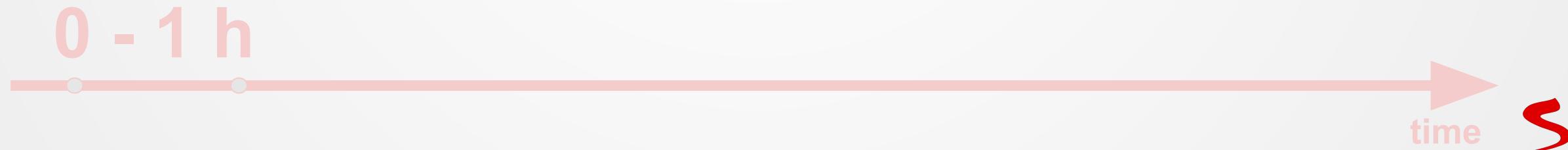
- Sběr zpětné vazby
 - linkování prokliků
 - anotace clickbaitů
- Modelování
 - predikce prokliku:
 - logistická regrese
 - DNN (DCN, DIEN)
 - clickbait detection:
 - DNN
- Metriky
 - precision/recall
 - perplexita
 - novelty
 - coverage



Reklamní systém

SKLIK.cz

- Sběr zpětné vazby
 - linkování prokliků
 - sběr výherních cen
- Modelování
 - predikce prokliku:
 - logistická regrese
 - CatBoost
- Metriky
 - log-loss
 - AUC
 - kalibrace
 - proklikovost
 - cena za proklik
 - výnos



Služba cílení reklamy

Cílení – “poznávání srdce a duše českého Internetu”



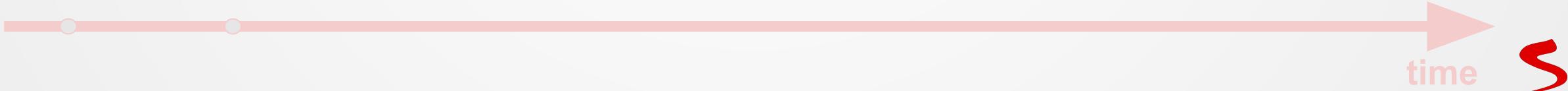
0 - 1 h



Systém cílení

- Co:
 - Vymezit cílovou skupinu pro reklamy
- Proč:
 - Adopce strategie inzerenta
- Jak:
 - Klasifikace uživatelů z dostupných dat
- Vstupní data:
 - Profil uživatele, historie interakcí uživatele
- Výstup:
 - Charakteristiky uživatelů, příslušnost do cílových skupin
- Dopad:
 - Vyšší spokojenost inzerentů
 - Zvýšení návratnosti reklam

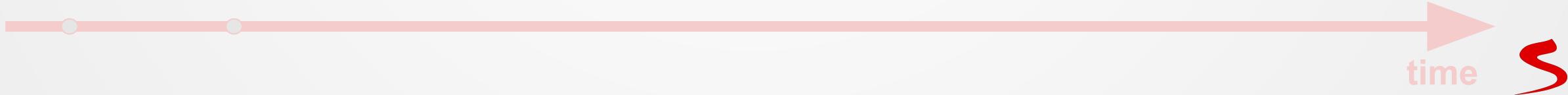
0 - 1 h



Systém cílení

- Data:
- ML úlohy:
 - Klasifikace kategorie uživatele (věk, pohlaví, zájmy, bonita)
 - Odhad identity uživatele
 - Klasifikace dokumentů, dotazů uživatele, shluková analýza
- Modelování:
 - XGBoost, Logistická regrese, KNN
- Role:
 - Data Scientist → ML Engineer → SW Developer

0 - 1 h



0 - 1 year



The image shows a tablet screen with two distinct content areas. On the left, a CSOB advertisement features a man holding a smartphone displaying various digital interfaces like a clock and calendar. The text reads: "KDYŽ POTŘEBUJETE SVOJÍ BANKU KDYKOLI A KDEKOLI" and "Založte si účet pro firmy a podnikatele online". On the right, the Seznam Zprávy news website is displayed, showing a navigation bar with categories like ZPRÁVY, BYZNYS, TECH, PODCASTY, MAGAZÍN, VIDEO, and REDAKCE. Below the navigation, there are several news cards with titles such as "Jaká bude zima?", "Video: Jako ze Star Wars.", and "Trpas britské premiérky". A sidebar on the right also promotes CSOB services for companies.



0 - 1 year



time

- klik, klik, klik ...
- Analýzy
 - Nové signály, nové modely/metody
 - Problémy s daty
 - Hodnocení přínosů
- Výzkum nových metod:
 - feature engineering - nové signály, transformace, selekce
 - modelování - nové modely
 - explorace, reinforcement learning
 - MLOps - automatizace ML pipeline



Recap

Reklamní systémy:

- Komplexní systém, mnoho hráčů, mnoho zájmů
- Obsluha pomocí provázaného systému ML algoritmů
- Velká, rychlá, komplexní data



Role

- Data Engineer
- Machine learning engineer
- Data scientist



Role - Data engineer

- uložení, zpracování a analýzy většího množství dat



Role - Machine learning engineer

- vytváření a udržování produkční infrastruktury pro strojové učení



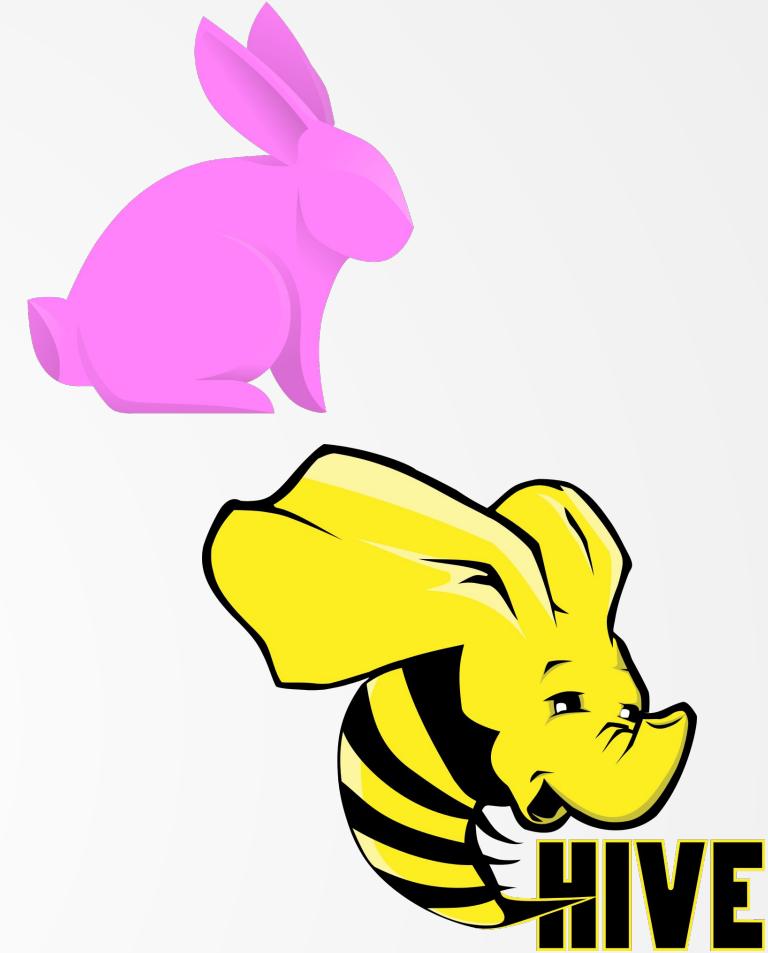
kubernetes



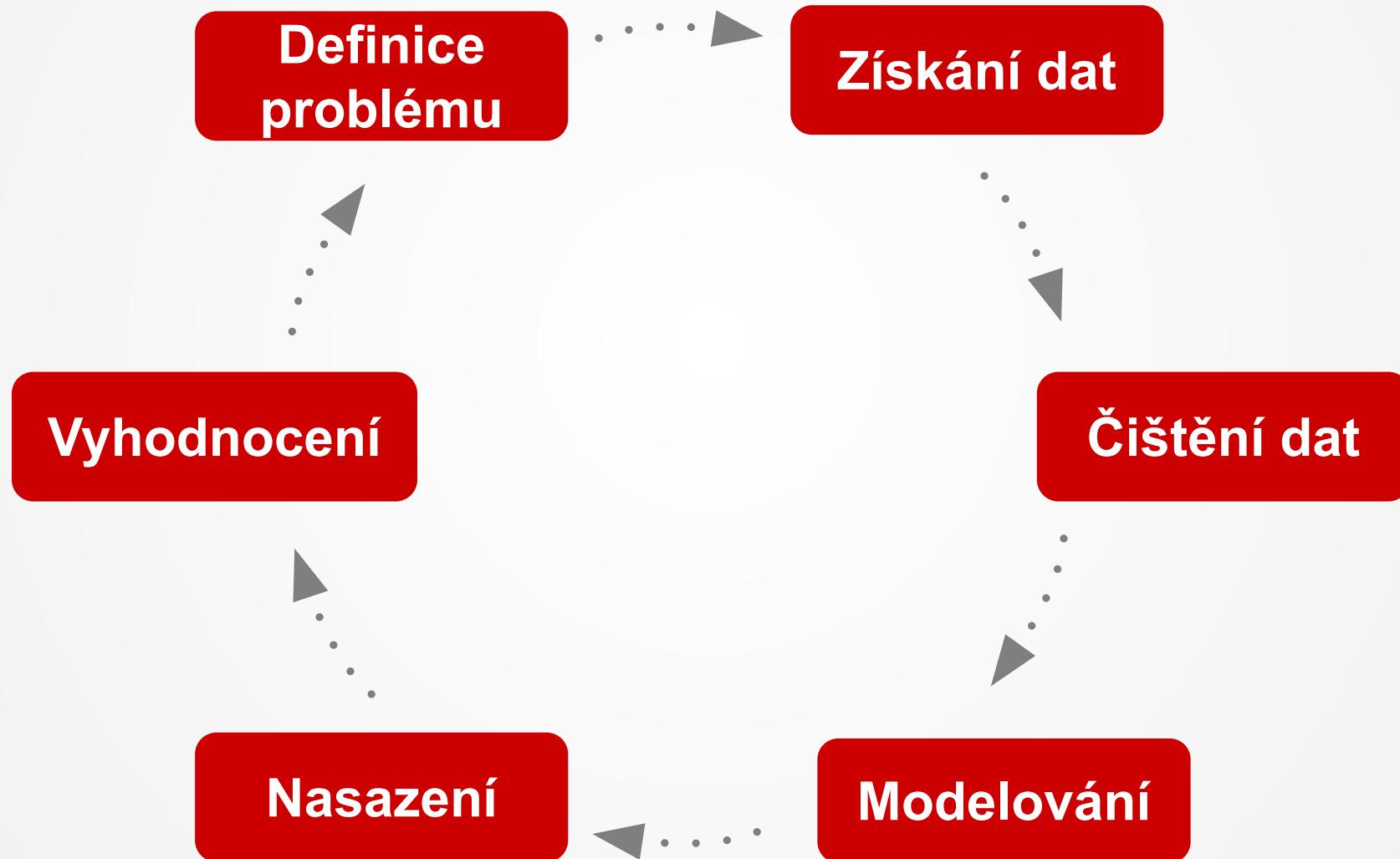
SCYLLA

Role - Data scientist

- čištění, analýza a modelování dat

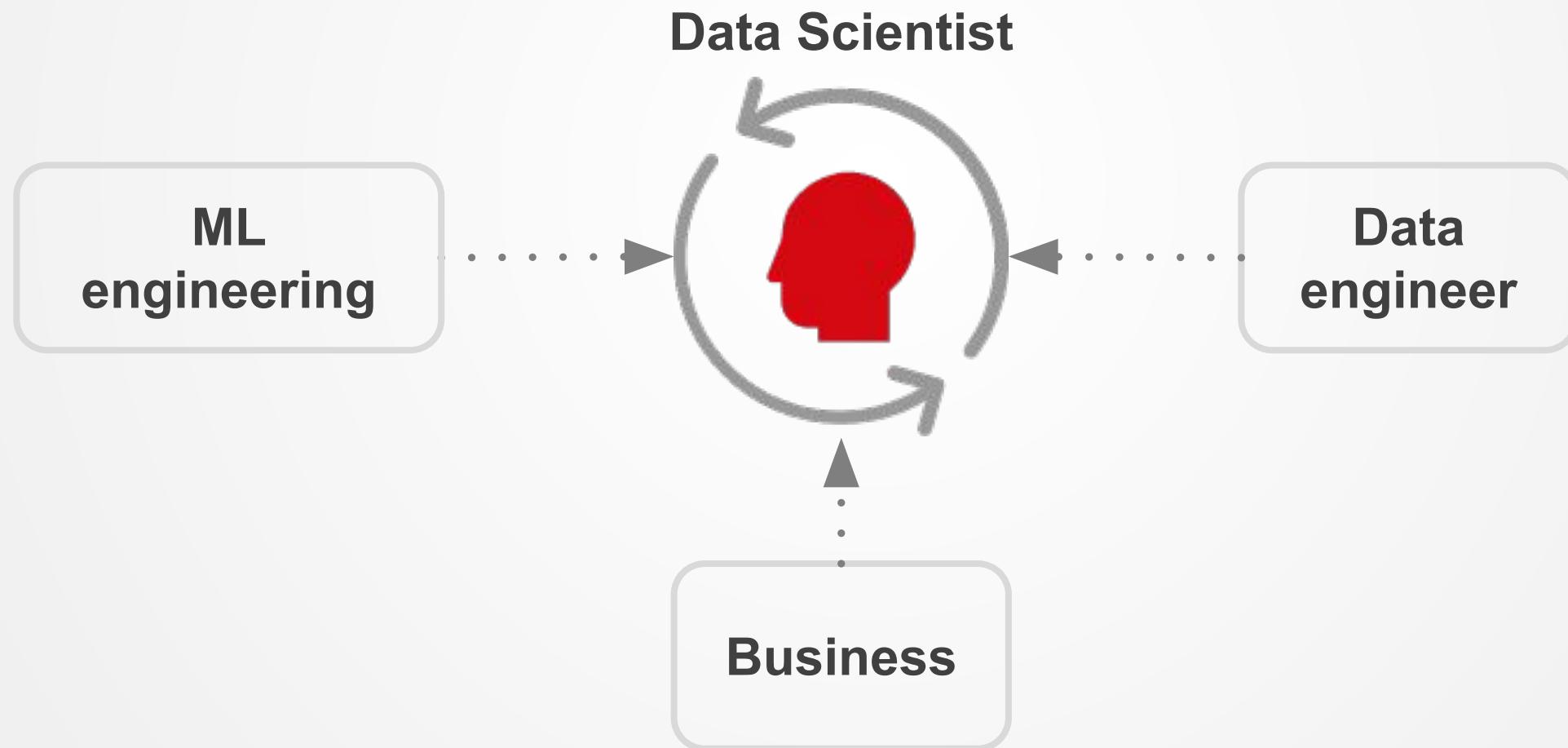


Role - Data scientist



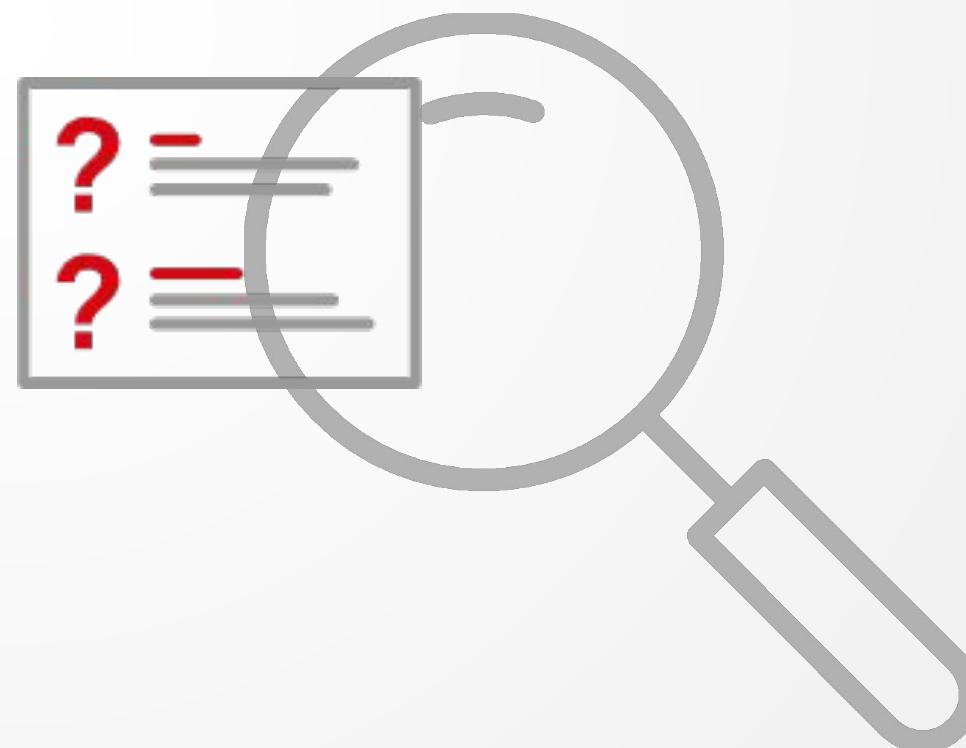
Role - Data scientist

- komunikace napříč spektrem rolí



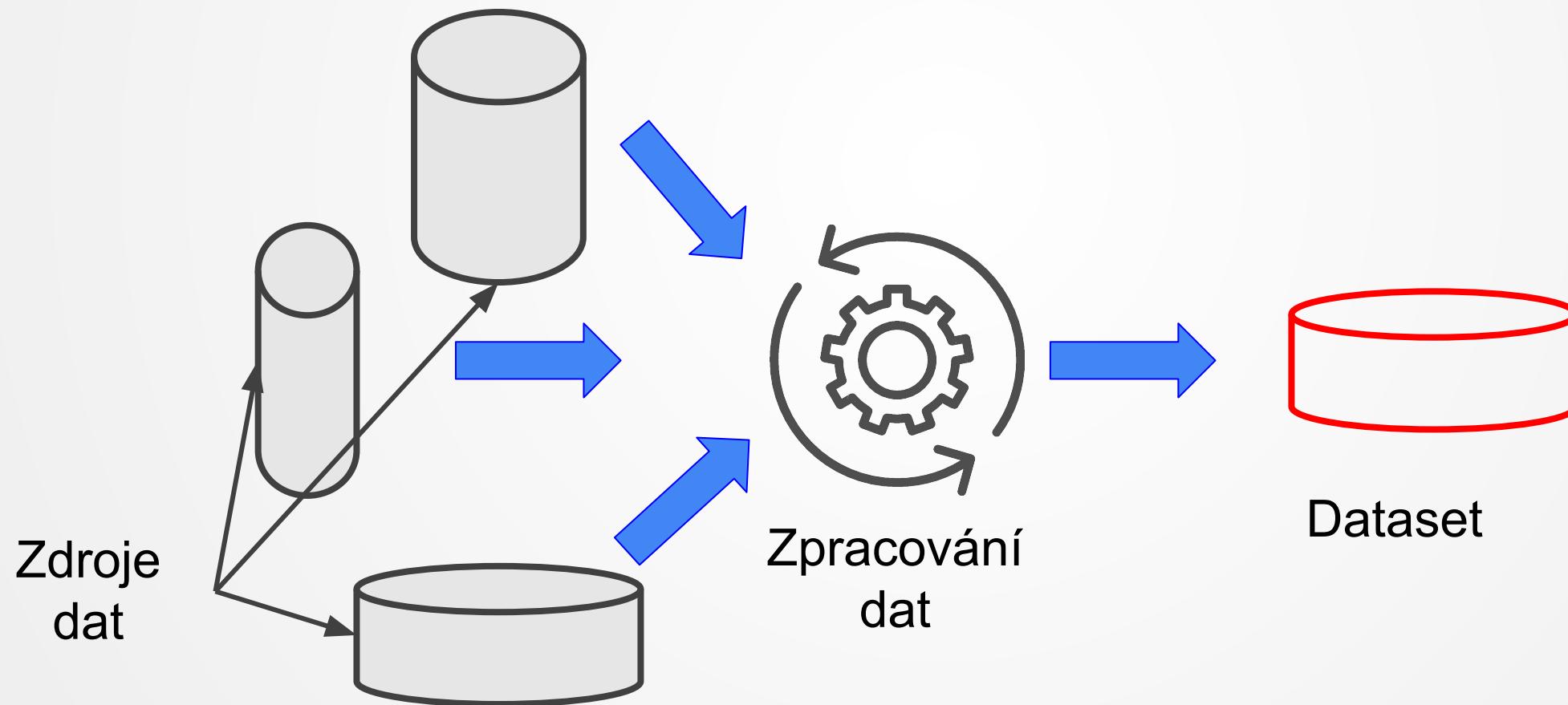
Definice problému

- klíčové je správně definovat řešený problém
- domluvit se na rozsahu řešení
- definice metrik a způsobu vyhodnocení projektu
- promyslet zdrojová data pro řešení



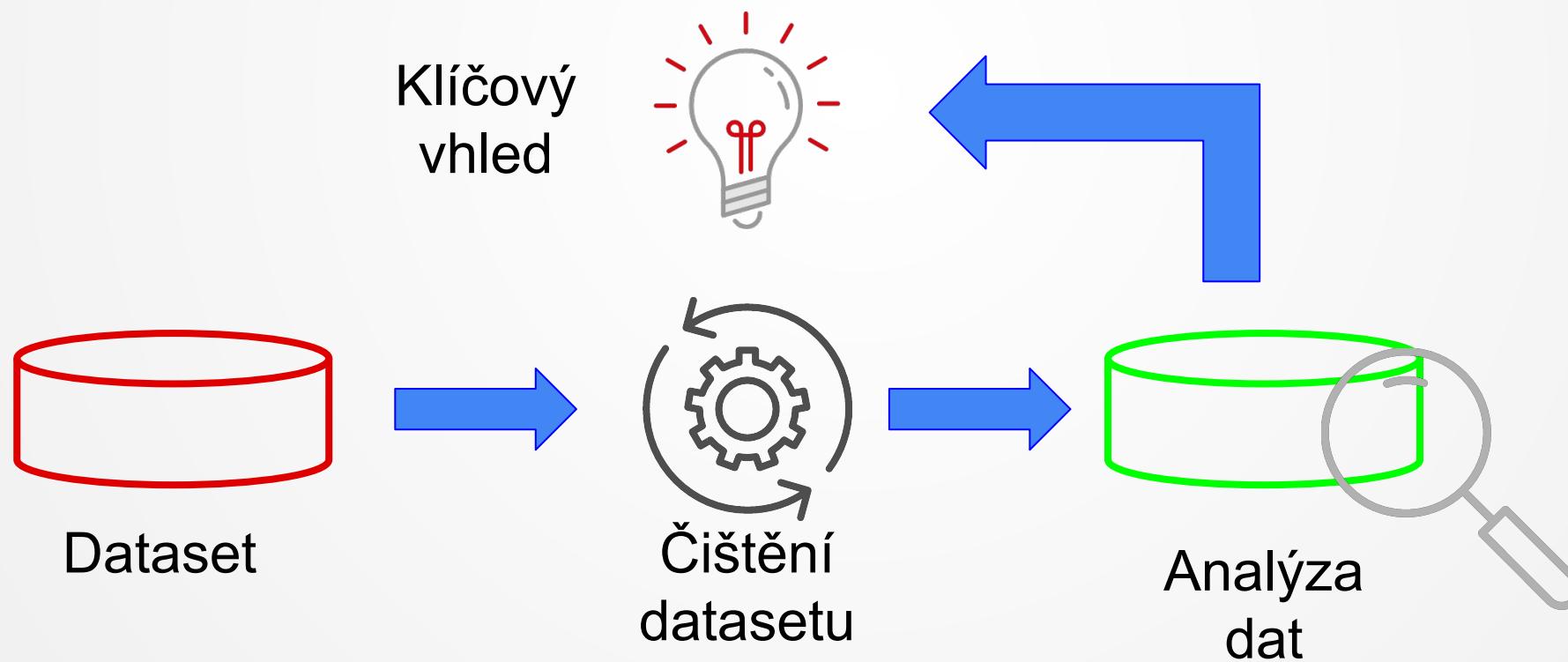
Získání dat

- klíčové pro jakýkoliv ML projekt
- typická spolupráce s data engineery



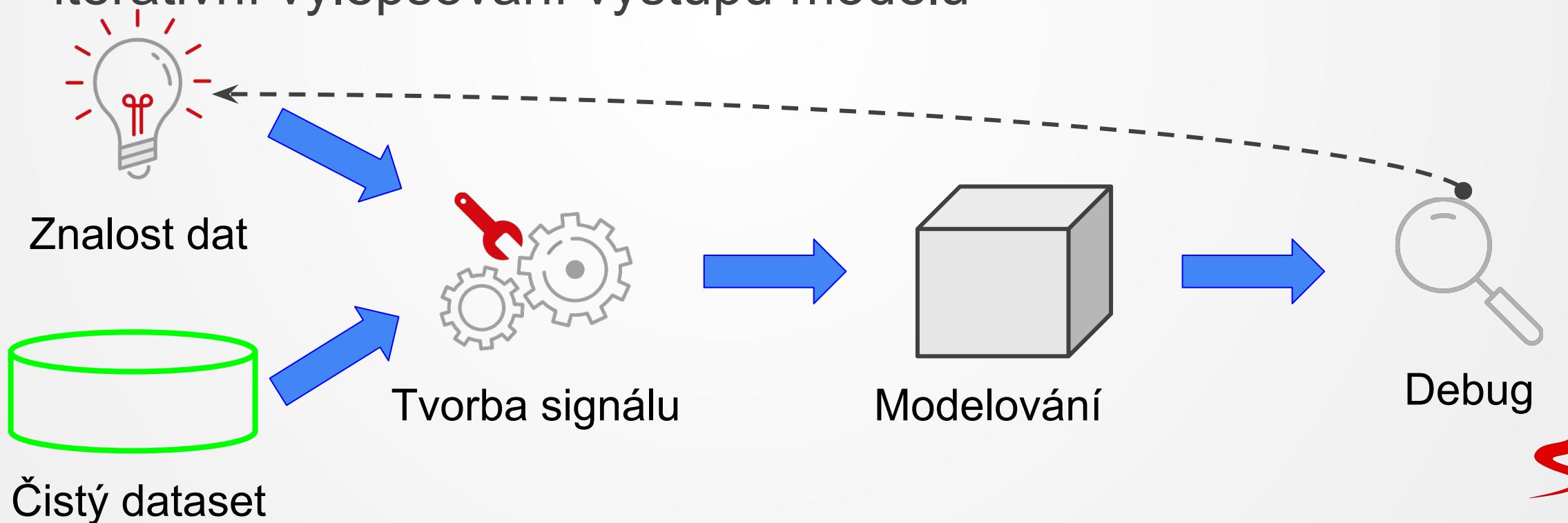
Analýza a čištění dat

- odstranění šumu a nevalidních dat
- získání klíčového výhledu do problematiky



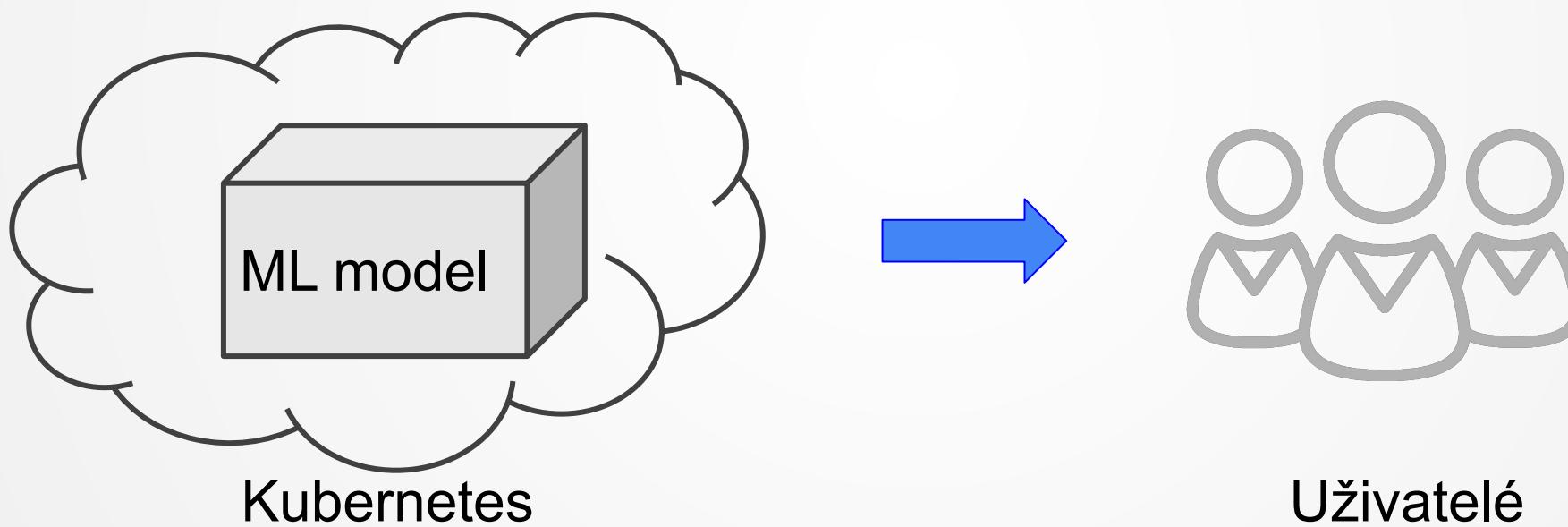
Modelování

- vytvoření signálu pro modely
- modelování
- analýza výstupu modelů
- iterativní vylepšování výstupu modelu



Nasazování

- spolupráce s machine learning engineery
- vysvětlení fungování modelů a převod do produkce



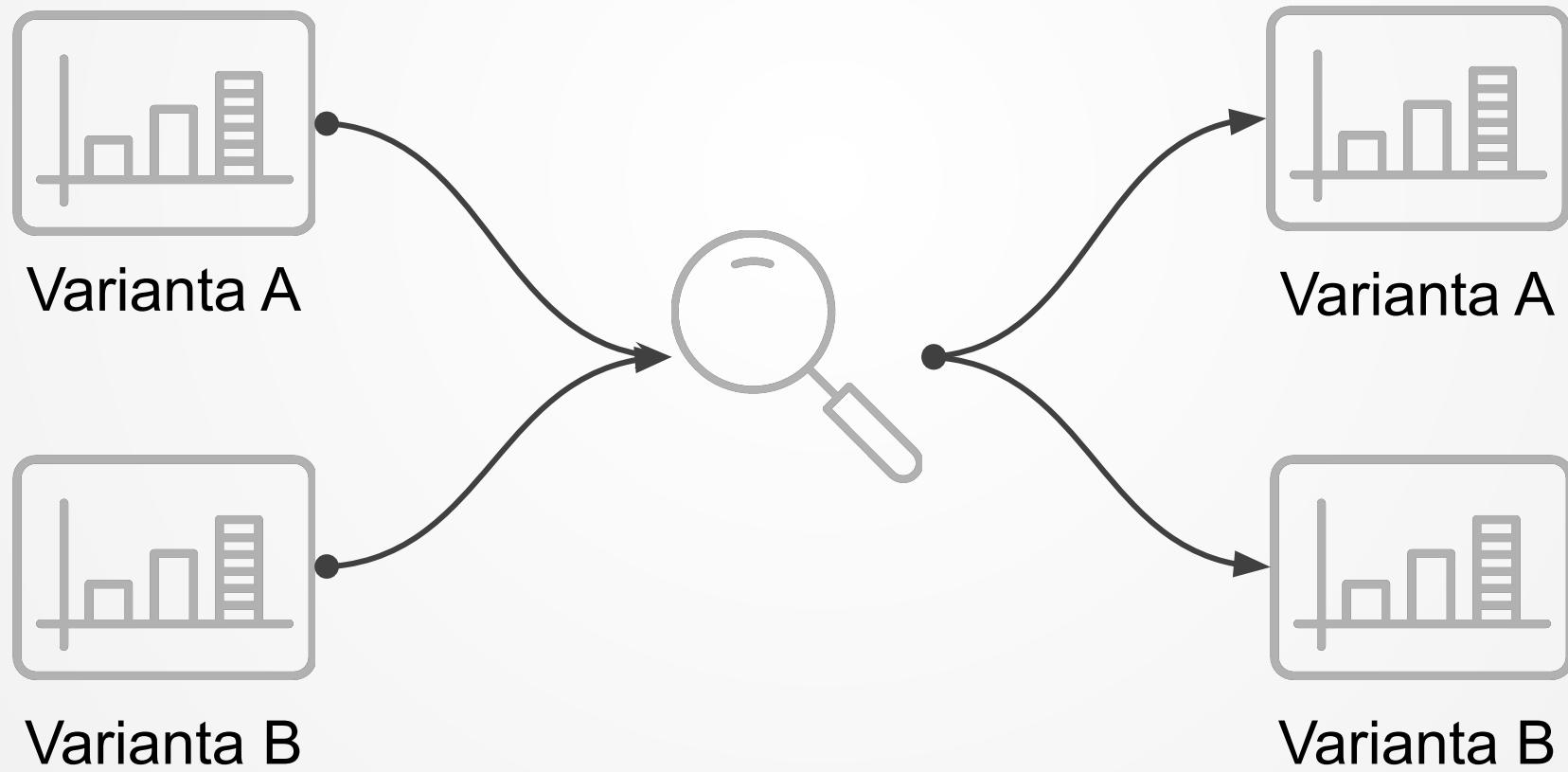
Vyhodnocování

- statistické vyhodnocení - typicky ab-testování



Vyhodnocování

- statistické vyhodnocení - typicky ab-testování



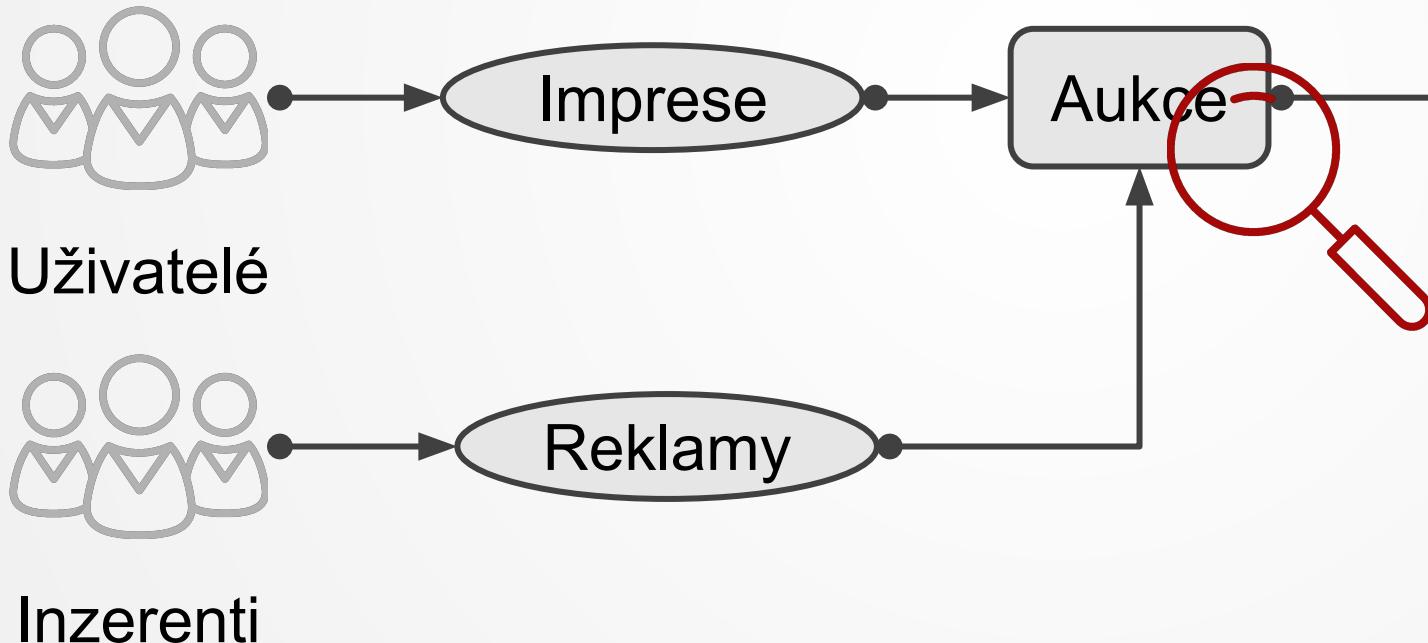
Týmy

- operační výzkum
- doporučování
- relevance v obsahové síti
- cílení & personalizace
- relevance ve vyhledávání



Týmy - operační výzkum

- optimalizace aukce s reklamou
- zaměření na maximalizace výdělku



Hledáte zubaře? Je jich dost, města je ale musí umět nalákat

DANIELA PŘÁDOVÁ



V Česku pracuje více než 8600 zubařských lekářů. V řadě míst je přesto péče nedostupná.

11:35 Stomatologů přibývá, přesto je péče v některých oblastech nedostupná. Na čem závisí, zda se k lékaři dostanete? Problém má hlubší kořeny a obec často marně hledají recept na efektivní řešení.



Článek si také můžete poslechnout v audioverzi.

Jedním z indikátorů neutěšeného stavu kolem stomatologické péče se stala všejína internetová fóra.

10:42 Ukrajina varuje: Kreml už se rozhodl napadnout Moldavsko

15:37 Putin zvažuje navázat rubl na zlato. Centrální banka má jiný názor

16:31 Sledujte, co se právě děje v redakci Seznam Zprávy. Budete s námi

DALŠÍ ČLÁNKY



Týmy - doporučování

- velké množství webového obsahu
- jak najít nejrelevantnější obsah pro uživatele?



Hledáte zubaře? Je jich dost, města je ale musí umět nalákat

DANIELA PŘÁDOVÁ

V Česku pracuje více než 8600 zubařských lékařů. V řadě míst je přesto péče nedostupná.

11:35

Stomatologů přibývá, přesto je péče v některých oblastech nedostupná. Na čem závisí, zda se k lékaři dostanete? Problém má hlubší kořeny a obec často marně hledají recept na efektivní řešení.

Byt zdarma nebo milion. Zubařů je dost, jen je umět nalákat

00:00 07:09

Článek si také můžete poslechnout v audioverzi.

Jedním z indikátorů neutěšeného stavu kolem stomatologické péče se stala velejívná internetová fóra.





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

**Nákupy v Česku podrážily 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 omlovala
Fantastický Krejčí! Český basketbalista v NBA vylepšil rekord

Garáž**Za volant od 17, vyšší tresty za hrani 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**ZIVĚ Bude až 20 stupňů, na Velikonoce se ale citelně ochladí**

Nevyzpytatelnost dubnového počasí se naplno projeví i o velikonočním týdnu, který právě startuje...
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
Nejštělejší 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 nelibí „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 pojíšťovacích 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	Aktuální opatření
+2 648	-181	+8	
Reinfekce: 397	1 258	39 880	Cestování

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

Týmy - relevance v contextové sítí

- obrovské množství reklam
- jak vybrat reklamu relevantní pro daného uživatele?



Týmy - relevance ve vyhledávací síti

- jak vybrat reklamu na základě vyhledávacího dotazu



Uživatelé



„tepelné čerpadlo“



The screenshot shows a search results page with the query "tepelné čerpadlo" in the search bar. Below the search bar are filters: Internet, Obrázky, Zboží, Mapy, Videa, Zprávy, Firmy, and Slovník. The results are listed in a grid:

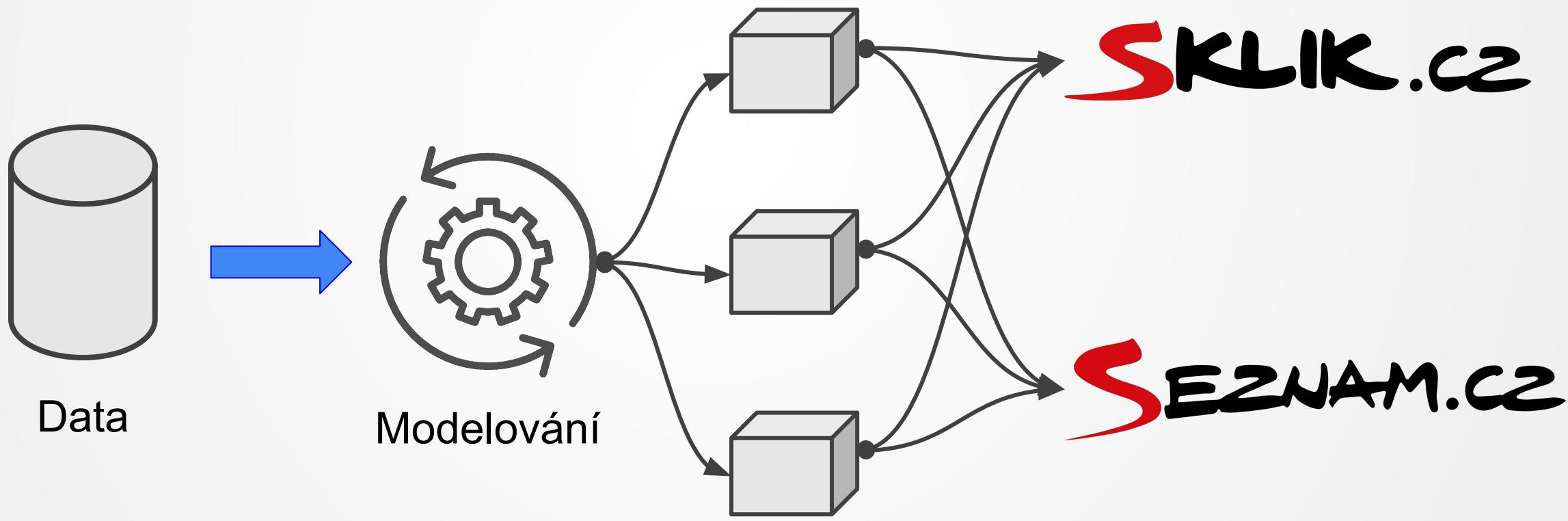
- Tepelné čerpadlo - Dotace Zelená úsporám**
81heat.cz/tepelnecerpadlo [Reklama]
Objednejte s tepelné čerpallo za předesezónní ceny. Možnost vrácení peněz z dotace.
Rychlé doručení · Akční nabídka · Nabídka profi montáže · Super ceny
• Nádražní 445/185, Ostrava
- Tepelná čerpadla voda vzduch - Od českého výrobce**
ceskatepelnacerpadla.cz [Reklama]
Pusťte starost s vytápěním Vašeho domu z hlavy. Vše s Vámi zkonzultujeme a zajistíme.
Návrh řešení zdarma · Konzultace zdarma · Český výrobce · Dotace až 127 500 Kč
Kontaktejte nás · Reference
Dotace · Fotovoltaika
- Tepelné čerpadlo | Vip Klíma - Ceny od 17.999 Kč s DPH**
vip-klima.cz/tepelnacerpadla [Reklama]
Montáž tepelného čerpadla máte u nás zcela zdarma. Ceny jsou již od 17.999 Kč s DPH.
Nejlevnější řešení · Doprava zdarma po celé ČR · 5 let záruka · Montáž zdarma
- Tepelná čerpadla na klíč - Tepelné čerpadlo s dotací**
arpeg.cz/tepelnacerpadla [Reklama]
Tepelná čerpadla voda/vzduch na klíč. Smluvně garantujeme navíc získání státní dotace NZÚ.

At the bottom of the search results is a map showing locations like Slaný, Kralupy nad Vltavou, Lysá nad Labem, Brandýs nad Labem, Nymburk, and Poděbrady. There are also buttons for "V mém okolí" and "Otevřené".



Týmy - personalizace a cílení

- zpracování uživatelských dat a personalizace modelů





kariera.seznam.cz

Machine learning výzkumník pro cílení
reklamy a personalizaci



Praktická část aneb Hrajeme si s daty

EDA - Exploratory Data Analysis

Účel analýzy dat:

- základní charakteristiky dat
- porozumění veličinám a vztahům mezi nimi
- ověření konzistence, odhalení chyb a anomálií
- (re)formulace úkolů/problémů
- návrh postupů zpracování dat včetně vhodných ML metod
- návrh dalších kroků

Důležité:

- První krok k řešení úlohy
- Velká data = průběžný/opakovaný proces (automatizace)
- Sleduje specifické zadání, účel zpracování dat

Naše zadání:

- Doporučování obsahu



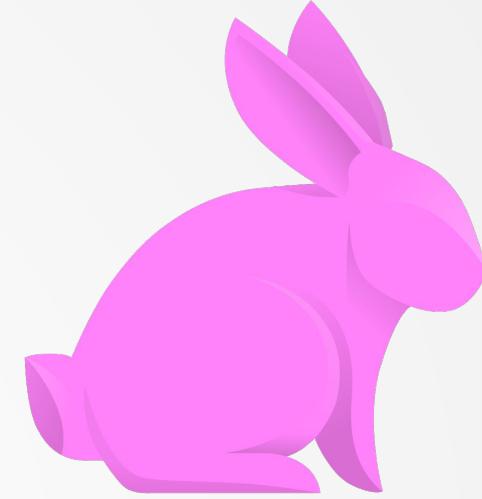
Predikce prokliku

- Co známe?
 - sekvenci interakcí uživatele s portálem msn.com
 - další kontextová data (ID, kategorie, subkategorie, URL, nadpis, abstrakt)
- Co chceme?
 - (udržet čtenáře na obsahové síti MSN)
 - nabídnout čtenáři to, co ho zajímá (tedy odkaz na článek na který si uživatel pravděpodobně klikne)
- Jak vyhodnotíme?
 - proklikovost nabízených odkazů / **maximalní pravděpodobnost prokliku**
 - vyhodnocujeme pouze offline! #ABtesting
- Jak provedeme?
 - logistická regrese vowpal wabbit



Vowpal wabbit

- open-source framework
- původně od Yahooo, nyní Microsoft
- původně logistická regrese v c++
- extrémně rychlá
- trénování pomocí online learning
- out-of-core -> škáluje pro velké množství dat
- podpora pro 1M+ signálu přes hashing trick
- interakce přes cmd (existují wrappery v pythonu)
- spousta dalších vlastností
- v rámci Seznamu používán pro doporučování obsahu a reklamy
- buggy

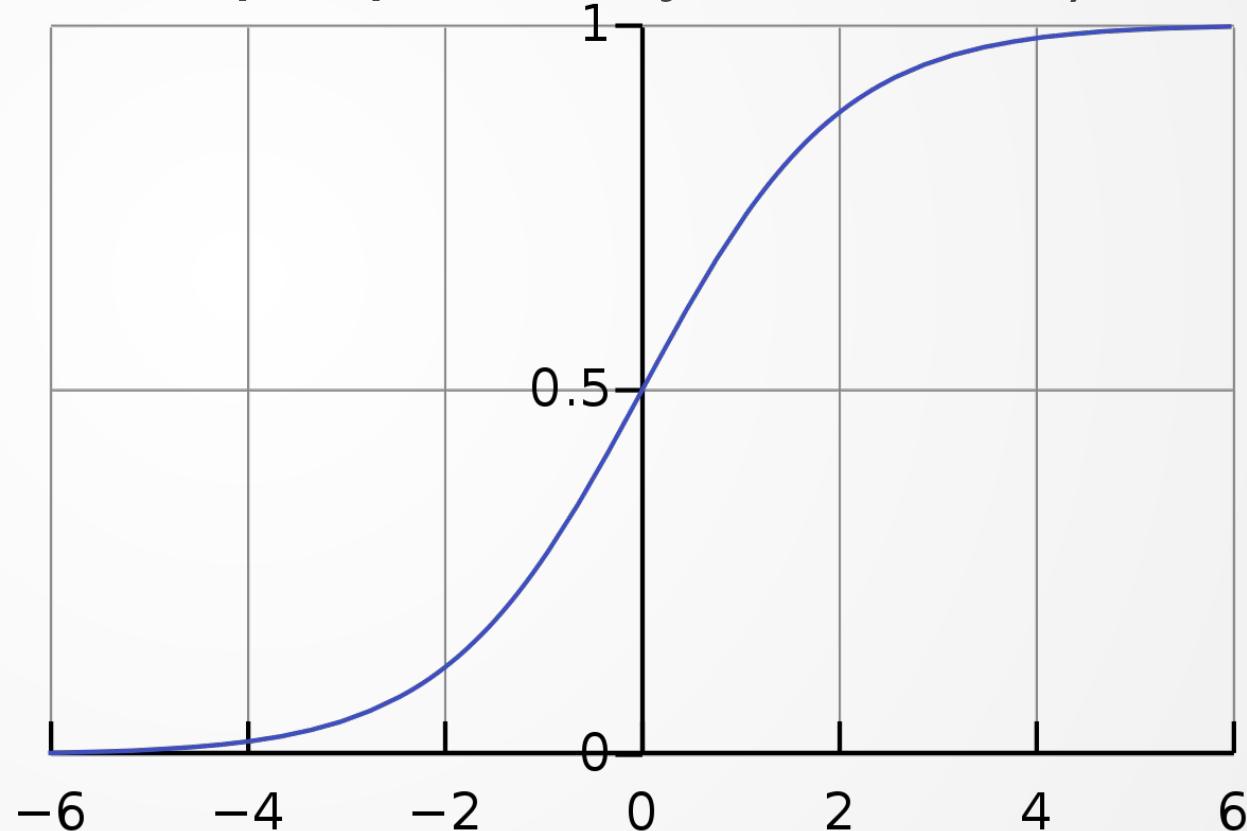


Vowpal wabbit - logistická regrese

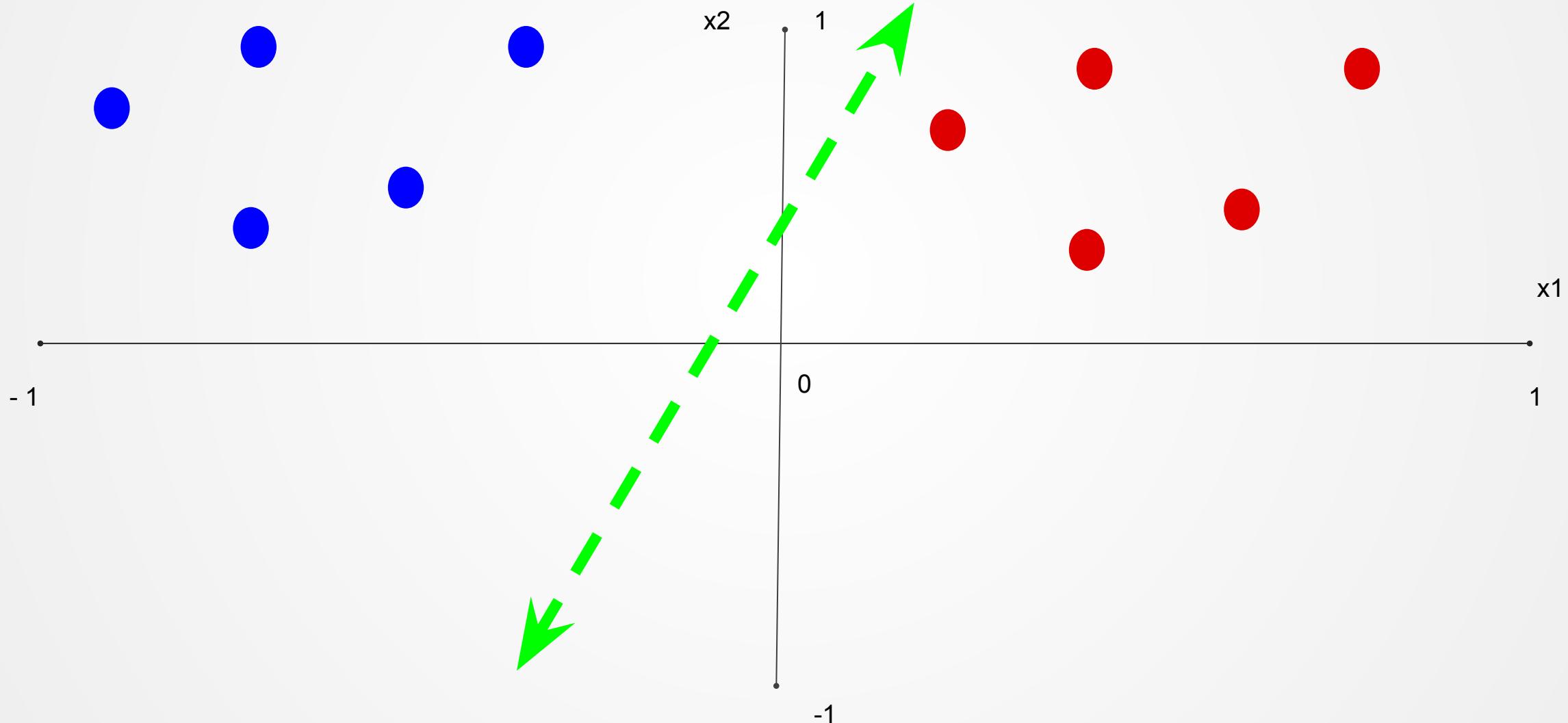
- binární klasifikátor (jeden z podporovaných modelů)
- lineární model

$$y = \sigma(w_0 + \sum_{i=1}^{|F|} w_i x_i)$$

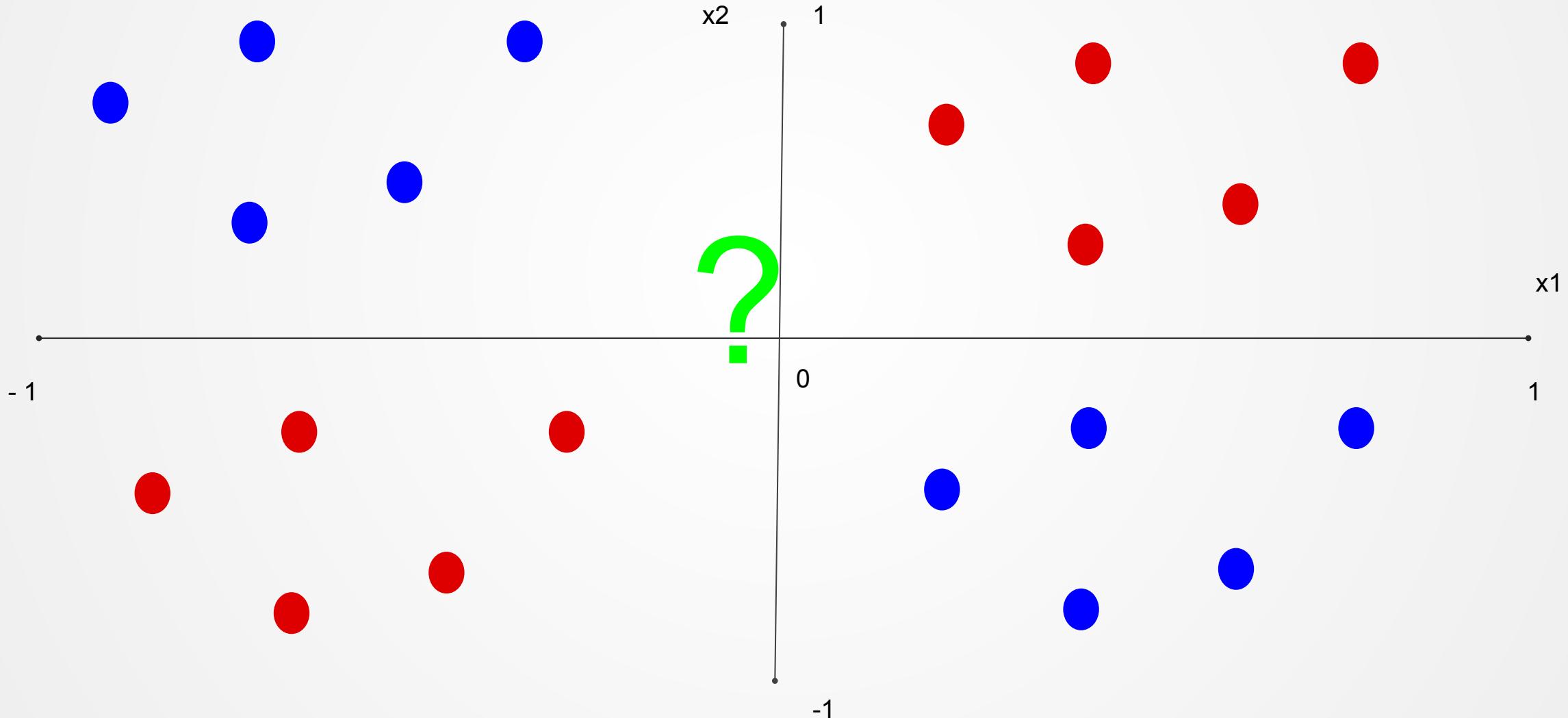
$$\sigma(x) = \frac{1}{1 + e^{-x}}$$



Vowpal wabbit - logistická regrese



Vowpal wabbit - logistická regrese

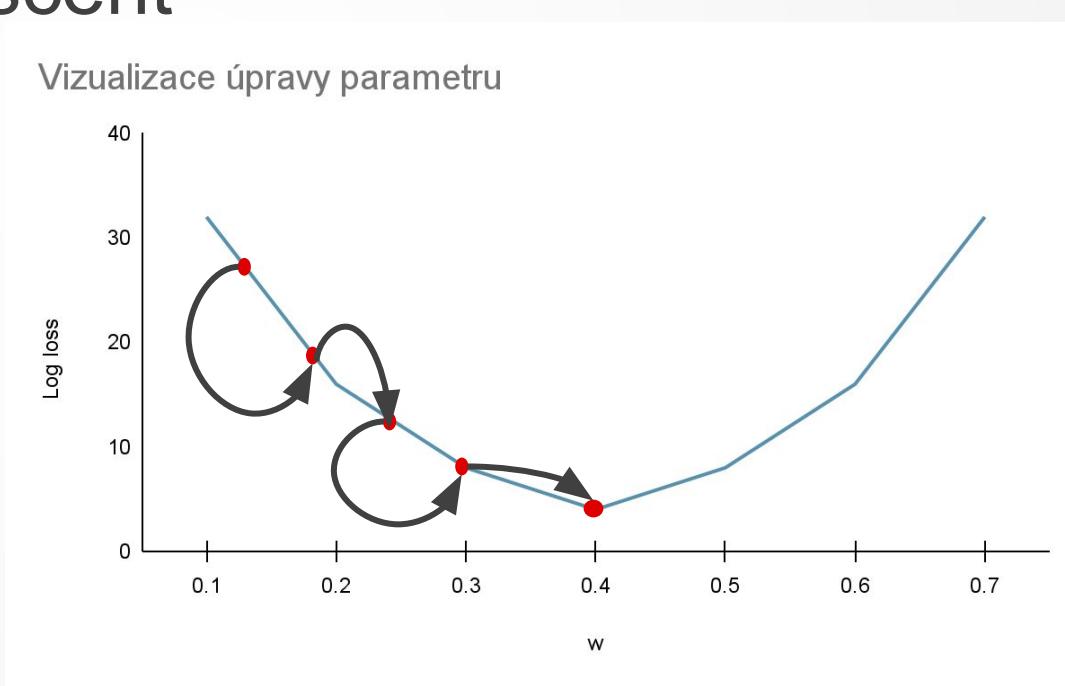


Vowpal wabbit - logistická regrese

- jak řešit nelineární závislosti u lineárního modelu?
- přidáním syntetického signálu
- $x3 = x1 * x2 <-$ díky této nové dimenzi bude již problém lineárně separovatelný

Vowpal wabbit - online learning

- minimalizace negativního log likelihoodu
- $\text{LogLoss} = -\frac{1}{N} \sum_i [y_n * \log(\sigma(x_n)) + (1 - y_n) * \log(1 - \sigma(x_n))]$
- stochastic gradient descent
- $w_j^n = w_j - \alpha * (\sigma(x_j^n) - y^n)x_j^n$
- doučování



Vowpal wabbit - formát vstupních data

- speciální znaky jsou ':' a '|'
- {Label} {Váha}|{Namespace} {Signál:hodnota*} | ...
- Label: -1 negativní, 1 pozitivní příklad
- Váha: váha trénovacího příkladu
- Namespace: oddělení typu signálu, např. uživatelská historie, imprese
- Signál: název signálu a binární nebo reálné hodnoty signálu

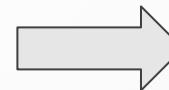
Vowpal wabbit - formát vstupních data

```
1 1.0|h N2 N1 N3 |i N5 |u age:25 location=praha gender=M  
-1 1.0|h N2 N1 N3 |i N8 |u age:25 location=praha gender=M  
1 1.0|h N2 N1 |i N10 |u age:35 location=ostrava gender=M  
....
```

Vowpal wabbit - hashing trick

- udržování mapování mezi indexem váhy a signálem je náročné
- jak během online učení inkorporovat nová data?
- bit precision

název signálu	index
věk	0
typ prohlížeče	1
imprese	2



$\text{index} = \text{hash}(\text{název signálu})$

Vowpal wabbit - feature engineering

- jednoduchý lineární model nezachytí komplexní interakci mezi signály
- řešením je použít kombinaci signálů
- generování kombinace signálu až v době trénování s využitím hashing tricku umožňuje zpracovat signály o obrovské kardinalitě
- součást modelovacího notebooku v rámci praktické části

Děkujeme za pozornost



SEZNAM.CZ

Kontakt



Vít Líbal, Tomáš Nováčik
Výzkum reklamních systémů

E-mail:

- vit.libal@firma.seznam.cz
- tomas.novacik@firma.seznam.cz

SEZNAM.CZ



kariera.seznam.cz

- Data science:
 - [Cílení a personalizace](#)
- Machine learning engineer:
 - [Machine learning výskumník pre relevanciu reklamy](#)
- Data engineer:
 - [Vývojár/ka DataOps platformy](#)



Budeme rádi za zpětnou vazbu!

- <https://tinyurl.com/ycy65m28>

