# Personalized Search

## Building a prototype to infer the user's interest

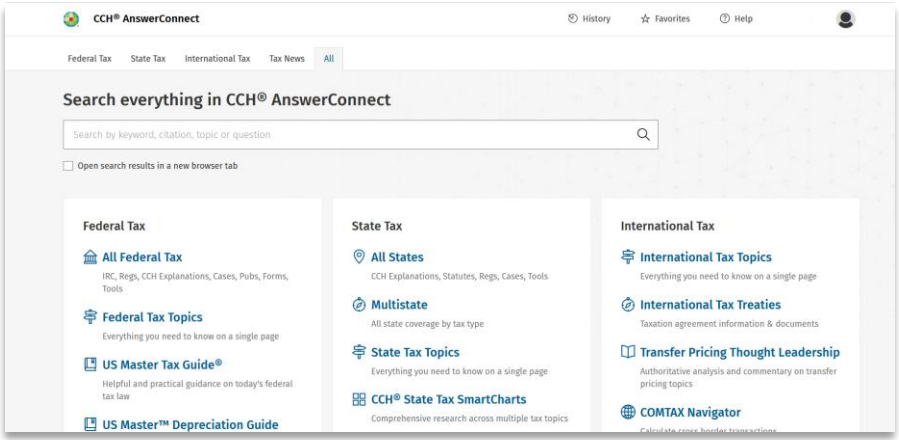
Tom Burgmans

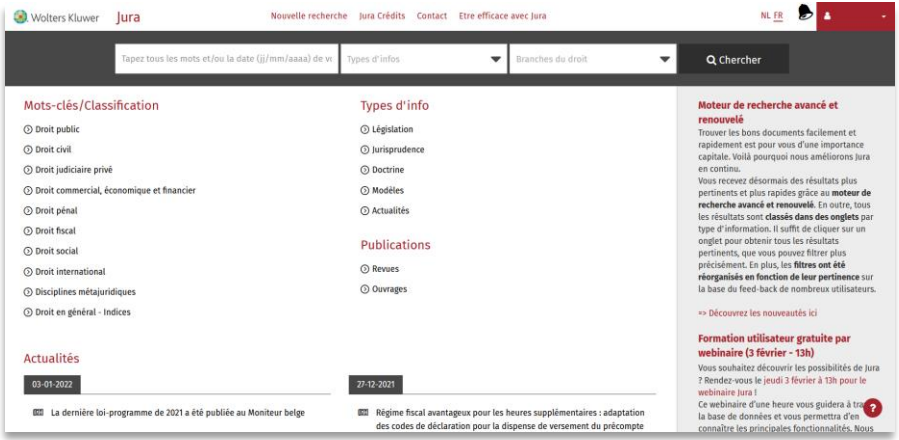
Technology Product Owner Search

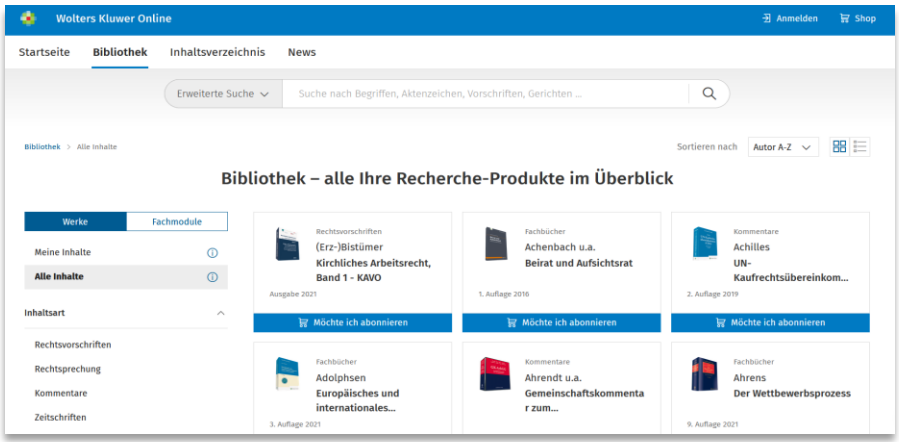
Wolters Kluwer

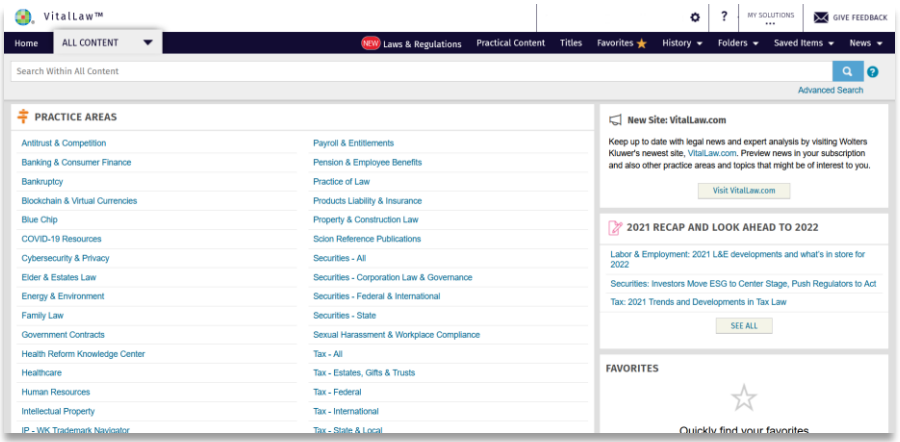
April 28th 2022

Wolters Kluwer


CCH® AnswerConnect
History
Favorites
Help
Federal Tax
State Tax
International Tax
Tax News
All
Search everything in CCH® AnswerConnect
Search by keyword, citation, topic or question
Open search results in a new browser tab
Federal Tax
All Federal Tax
IRC, Regs, CCH Explanations, Cases, Pubs, Forms, Tools
Federal Tax Topics
Everything you need to know on a single page
US Master Tax Guide®
Helpful and practical guidance on today's federal tax law
US Master™ Depreciation Guide
State Tax
All States
CCH Explanations, Statutes, Regs, Cases, Tools
Multistate
All state coverage by tax type
State Tax Topics
Everything you need to know on a single page
CCH® State Tax SmartCharts
Comprehensive research across multiple tax topics
International Tax
International Tax Topics
Everything you need to know on a single page
International Tax Treaties
Taxation agreement information & documents
Transfer Pricing Thought Leadership
Authoritative analysis and commentary on transfer pricing topics
COMTAX Navigator


Wolters Kluwer
Jura
Nouvelle recherche
Jura Crédits
Contact
Etre efficace avec Jura
NL FR
Tapez tous les mots et/ou la date (jj/mm/aaaa) de vi
Types d'infos
Branches du droit
Chercher
Mots-clés/Classification
Droit public
Droit civil
Droit judiciaire privé
Droit commercial, économique et financier
Droit pénal
Droit fiscal
Droit social
Droit international
Disciplines métajuridiques
Droit en général - Indices
Types d'info
Législation
Jurisprudence
Doctrine
Modèles
Actualités
Publications
Revues
Ouvrages
Actualités
03-01-2022
La dernière loi-programme de 2021 a été publiée au Moniteur belge
27-12-2021
Régime fiscal avantageux pour les heures supplémentaires : adaptation des codes de déclaration pour la dispense de versement du précompte
Moteur de recherche avancé et renouvelé
Trouver les bons documents facilement et rapidement est pour vous d'une importance capitale. Voilà pourquoi nous améliorons Jura en continu.
Vous recevez désormais des résultats plus pertinents et plus rapides grâce au moteur de recherche avancé et renouvelé. En outre, tous les résultats sont classés dans des onglets par type d'information. Il suffit de cliquer sur un onglet pour obtenir tous les résultats pertinents, que vous pouvez filtrer plus précisément. En plus, les filtres ont été réorganisés en fonction de leur pertinence sur la base du feed-back de nombreux utilisateurs.
=> Découvrez les nouveautés ici
Formation utilisateur gratuite par webinaire (3 février - 13h)
Vous souhaitez découvrir les possibilités de Jura ? Rendez-vous le jeudi 3 février à 13h pour le webinaire Jura !
Ce webinaire d'une heure vous guidera à tra
la base de données et vous permettra d'en
connaître les principales fonctionnalités. Nous


Wolters Kluwer Online
Anmelden
Shop
Startseite
Bibliothek
Inhaltsverzeichnis
News
Erweiterte Suche
Suche nach Begriffen, Aktenzeichen, Vorschriften, Gerichten ...
Bibliothek > Alle Inhalte
Sortieren nach
Autor A-Z
Bibliothek – alle Ihre Recherche-Produkte im Überblick
Werke
Fachmodule
Meine Inhalte
Alle Inhalte
Inhaltsart
Rechtsvorschriften
Rechtsprechung
Kommentare
Zeitschriften
Rechtsvorschriften
(Erz-)Bistümer Kirchliches Arbeitsrecht, Band 1 - KAVO
Ausgabe 2021
Möchte ich abonnieren
Fachbücher
Achenbach u.a. Beirat und Aufsichtsrat
1. Auflage 2016
Möchte ich abonnieren
Kommentare
Achilles UN-Kaufrechtsübereinkom...
2. Auflage 2019
Möchte ich abonnieren
Fachbücher
Adolphsen Europäisches und internationales...
3. Auflage 2021
Kommentare
Ahrendt u.a. Gemeinschaftskommentar zum...
Fachbücher
Ahrens Der Wettbewerbsprozess
9. Auflage 2021


VitalLaw™
MY SOLUTIONS
GIVE FEEDBACK
Home
ALL CONTENT
Laws & Regulations
Practical Content
Titles
Favorites
History
Folders
Saved Items
News
Search Within All Content
Advanced Search
PRACTICE AREAS
Antitrust & Competition
Banking & Consumer Finance
Bankruptcy
Blockchain & Virtual Currencies
Blue Chip
COVID-19 Resources
Cybersecurity & Privacy
Elder & Estates Law
Energy & Environment
Family Law
Government Contracts
Health Reform Knowledge Center
Healthcare
Human Resources
Intellectual Property
IP - WK Trademark Navigator
Payroll & Entitlements
Pension & Employee Benefits
Practice of Law
Products Liability & Insurance
Property & Construction Law
Scion Reference Publications
Securities - All
Securities - Corporation Law & Governance
Securities - Federal & International
Securities - State
Sexual Harassment & Workplace Compliance
Tax - All
Tax - Estates, Gifts & Trusts
Tax - Federal
Tax - International
Tax - State & Local
New Site: VitalLaw.com
Keep up to date with legal news and expert analysis by visiting Wolters Kluwer's newest site, VitalLaw.com. Preview news in your subscription and also other practice areas and topics that might be of interest to you.
Visit VitalLaw.com
2021 RECAP AND LOOK AHEAD TO 2022
Labor & Employment: 2021 L&E developments and what's in store for 2022
Securities: Investors Move ESG to Center Stage, Push Regulators to Act
Tax: 2021 Trends and Developments in Tax Law
SEE ALL
FAVORITES
Quickly find your favorites

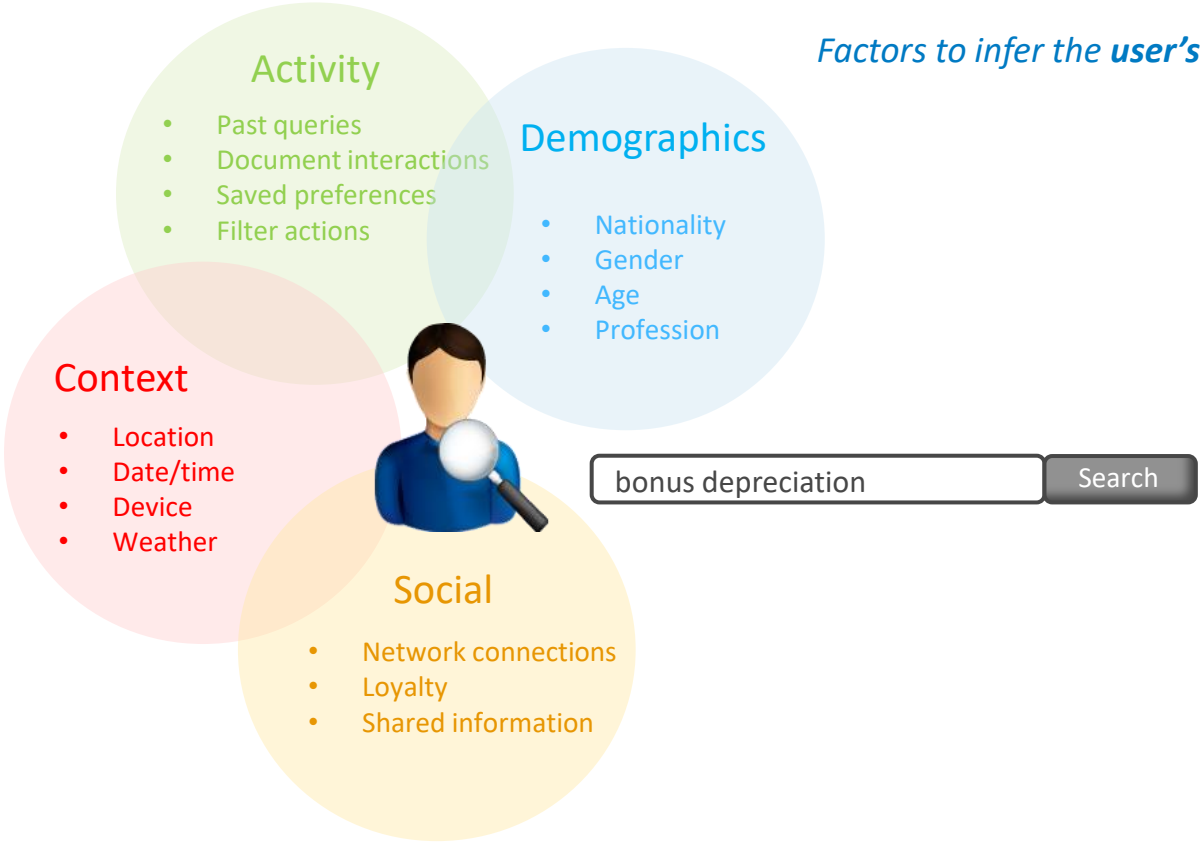
# Make search better through **Personalization**….a controversial topic

## Wikipedia says

**Personalized search** is web based search results that are tailored specifically to an individual's interests by incorporating information about the individual beyond the specific query provided.

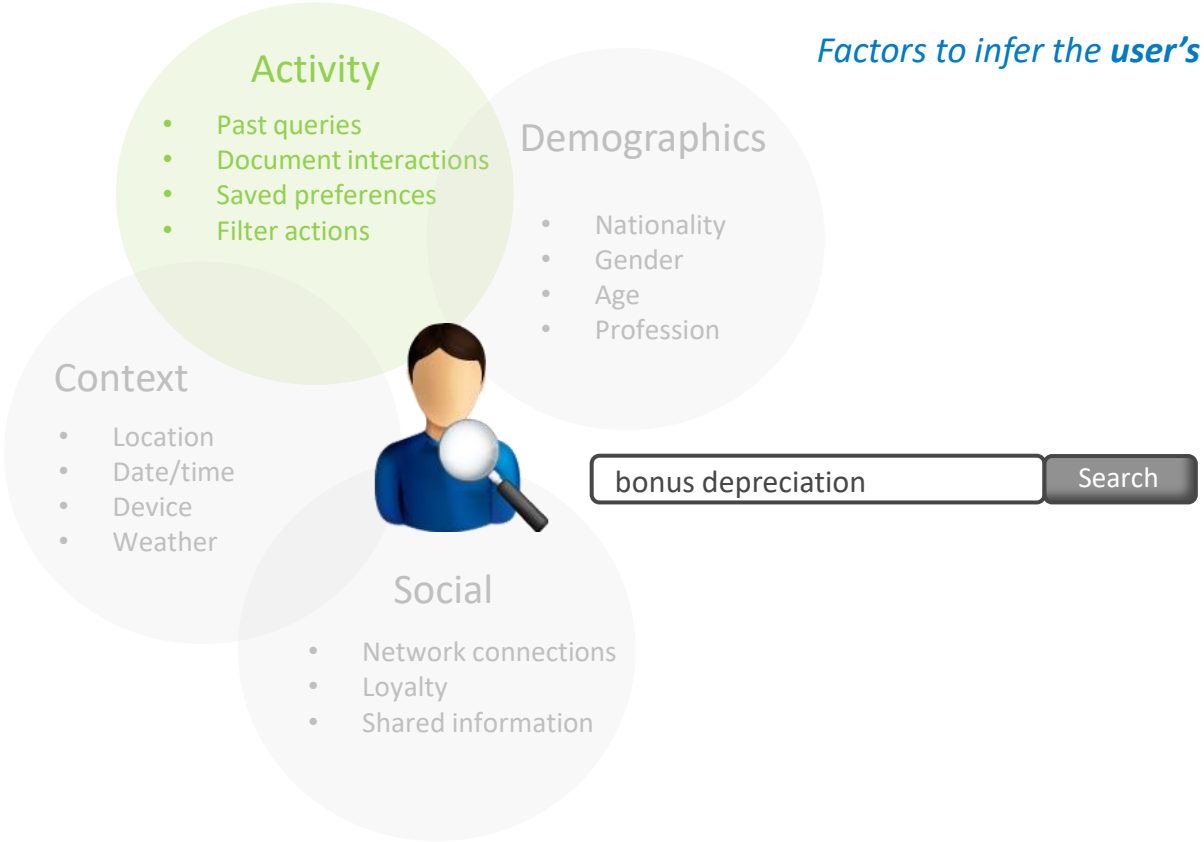
# What could make search “personal”

*Factors to infer the **user’s interest***



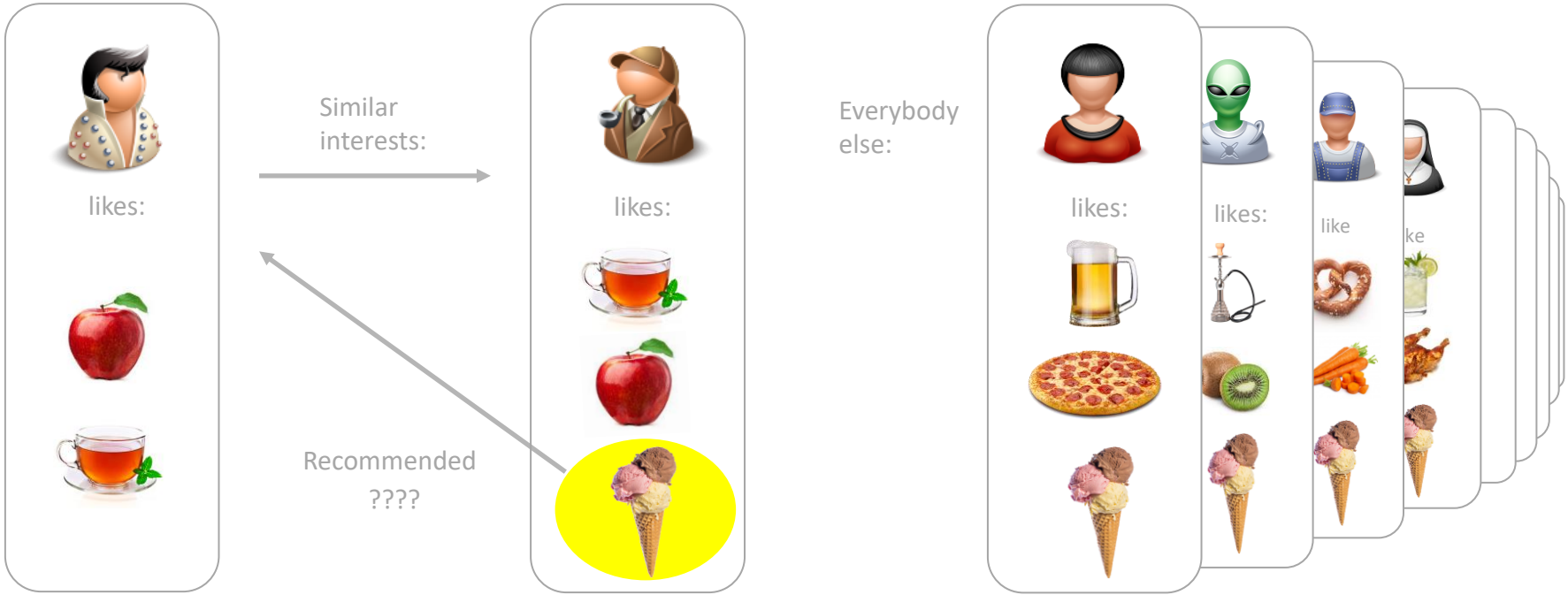
# What could make search “personal” (this PoC)

*Factors to infer the **user’s interest***



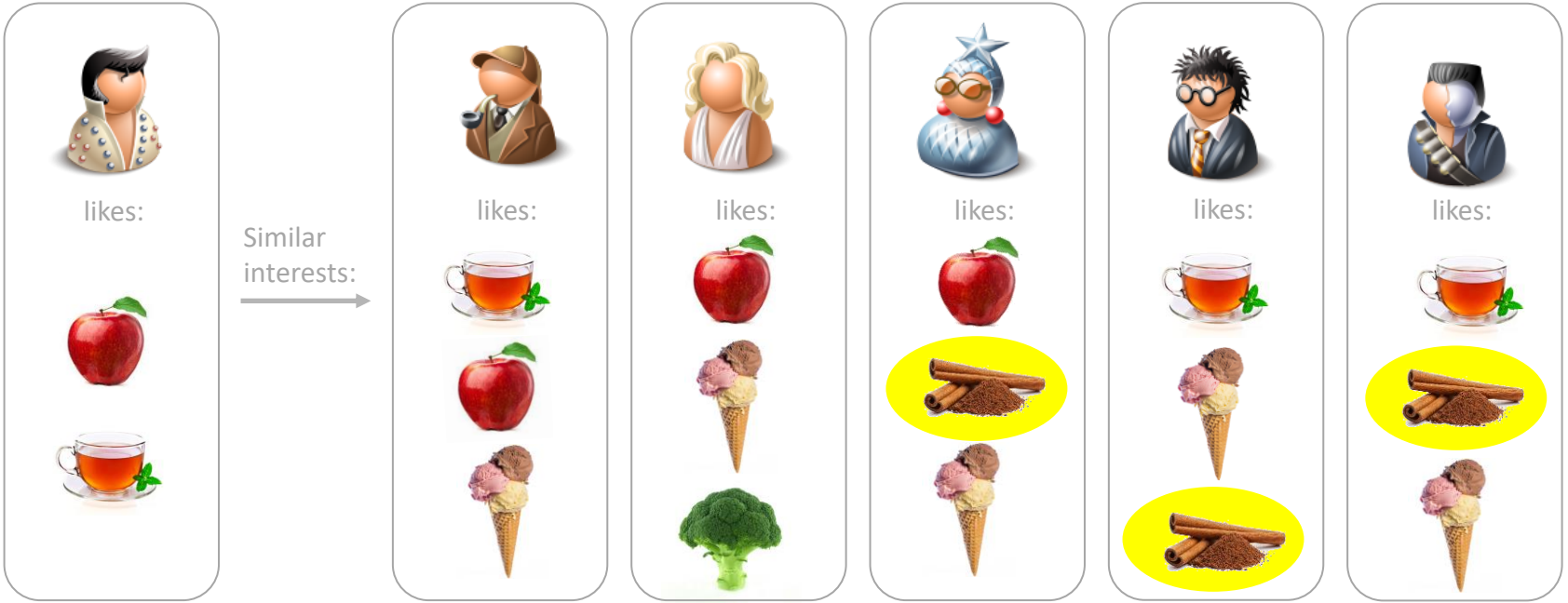
# Hypothesis

*The basis for **Personalized Search** is a **Recommendation Engine** which is based on an index of **user activity** that allows to infer the **user's interest** via **collaborative filtering**.*



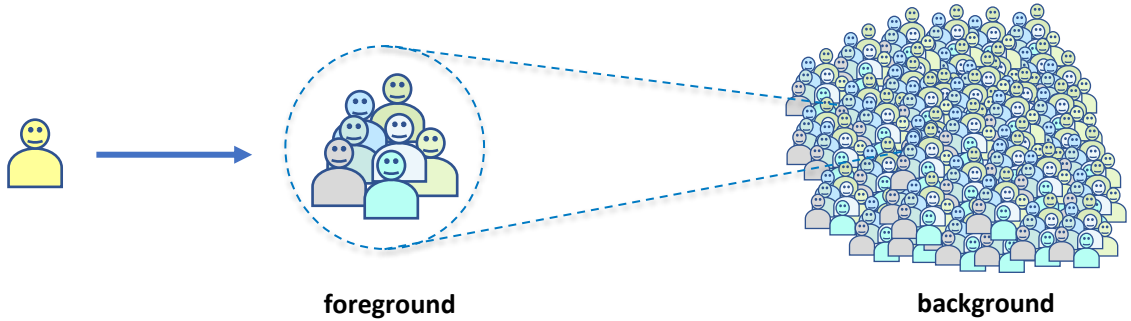
# Collaborative filtering & anomaly detection

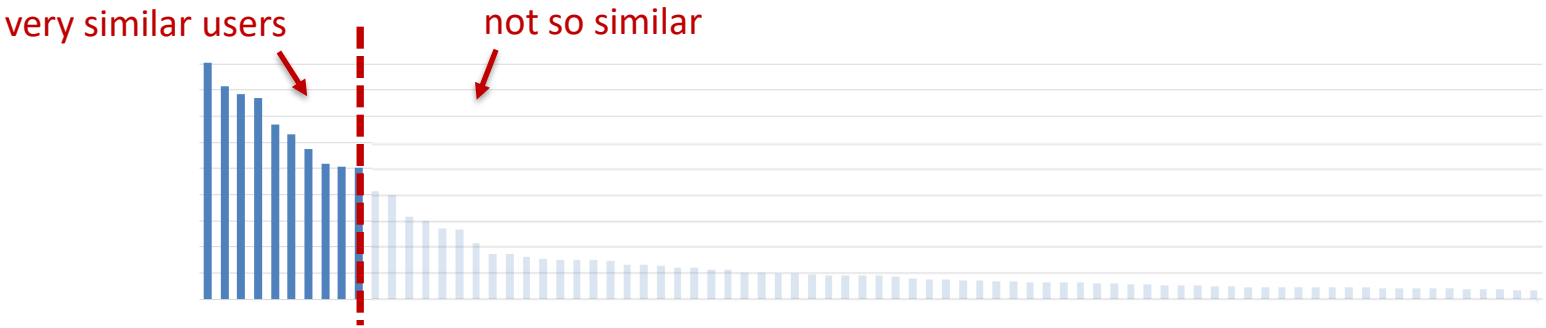
- *Find users with **similar interests***
- *Recommend items with an **unusually high presence** in this group compared to the rest*



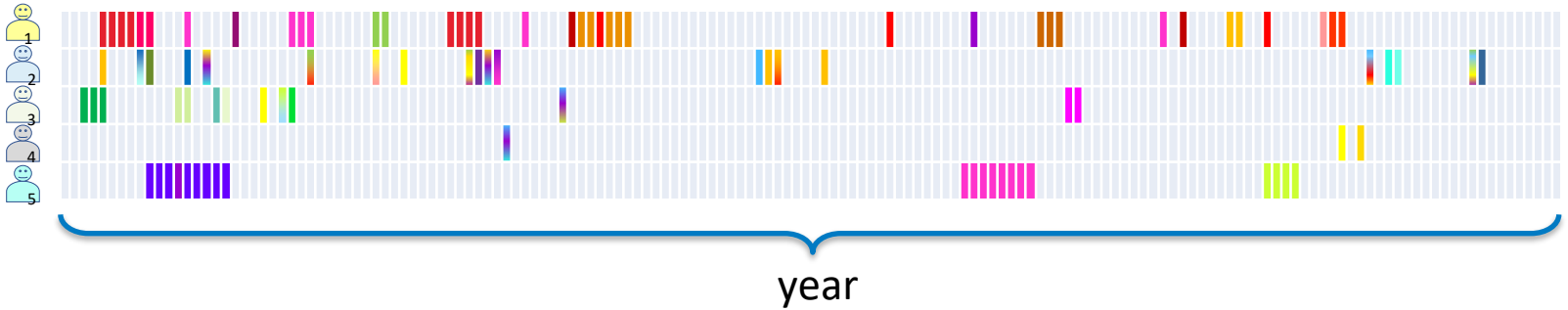
# Collaborative filtering & selecting the foreground

- *Find only users with **very similar interests***
- *Balance between similarity and volume*



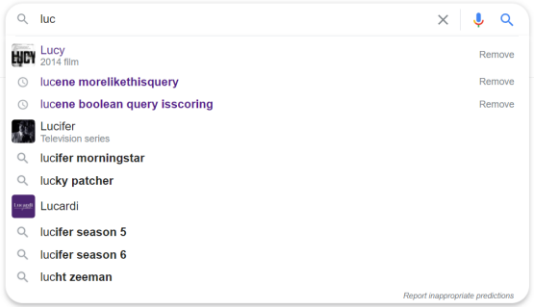
# What is the ideal time window to infer the user's interest?



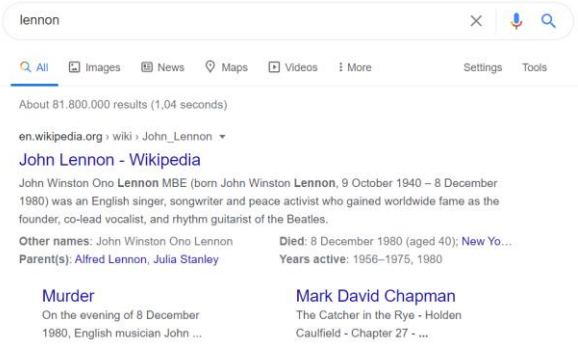


# Example extremely temporary personalization

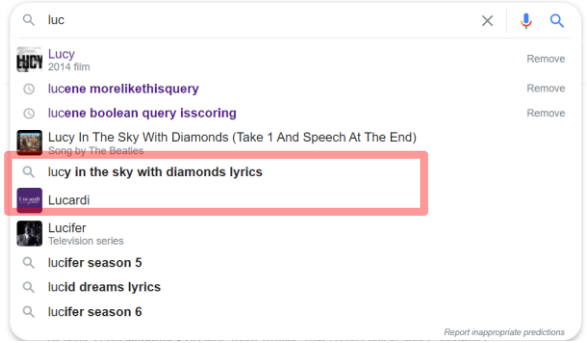
1



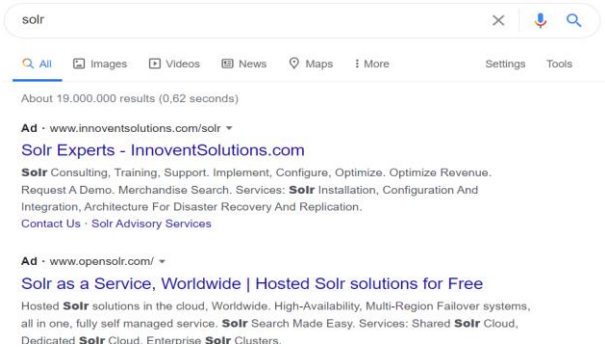

2



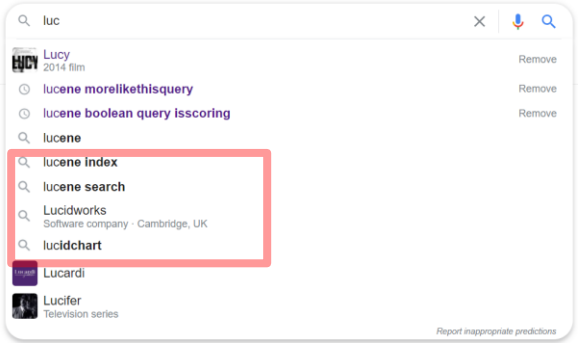

3




4

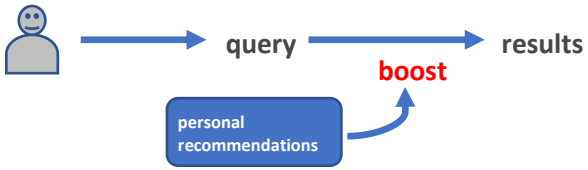



5

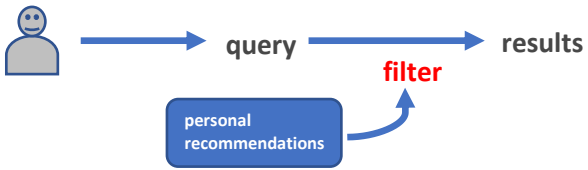


# How to apply the inferred user's interest?

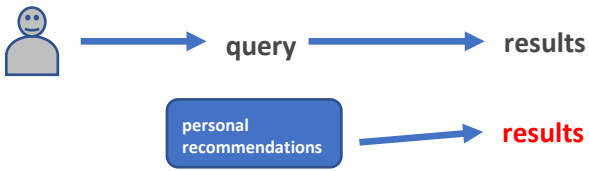
as a boost:




as a filter:

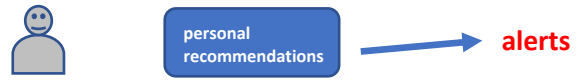



as a separate result set:




as interesting items:


as alerts:




as query suggestions:

# Goals of this PoC

## How to build it?

What can we do on our existing technology stack?

## In what cases to apply it?

Where is the need for personalization? What form is needed?

## What data do we need?

Define ‘additional’ data sources; cleaning of noise

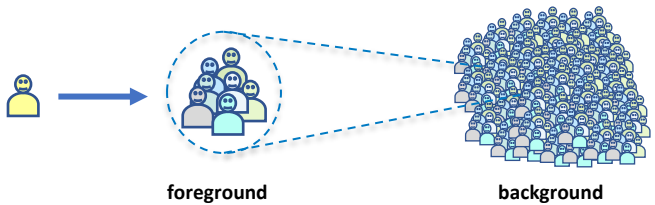
## How to measure success?

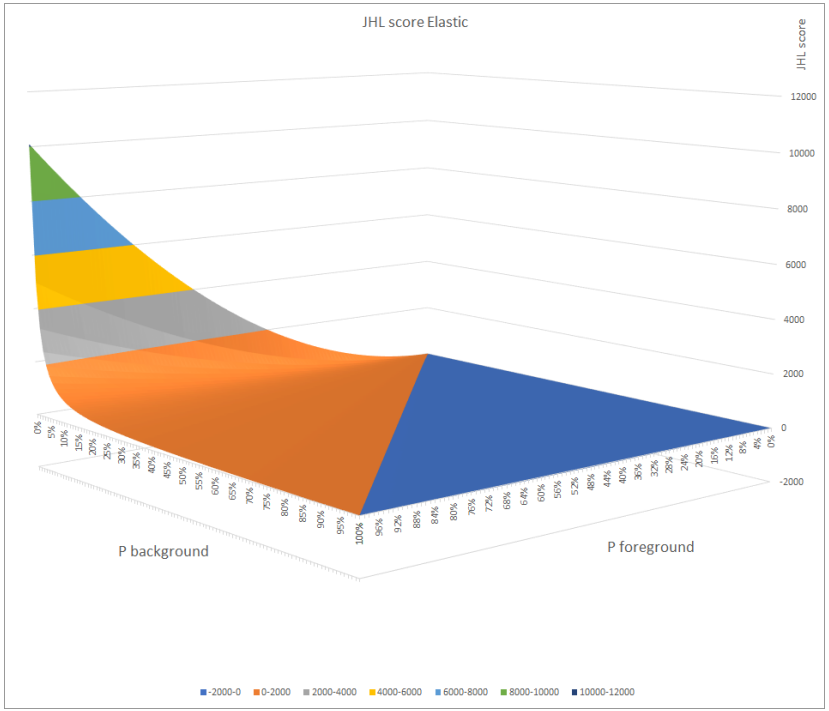
Can we test/tune it offline? How to A/B test online?

- Technology
- Need
- Data
- Verification

# Finding a technology for anomaly detection

elasticsearch Significant Terms



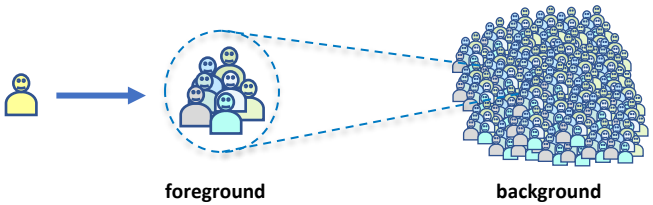

JLH score in Elastic = `(fg_percentage - bg_percentage) * (fg_percentage / bg_percentage)`

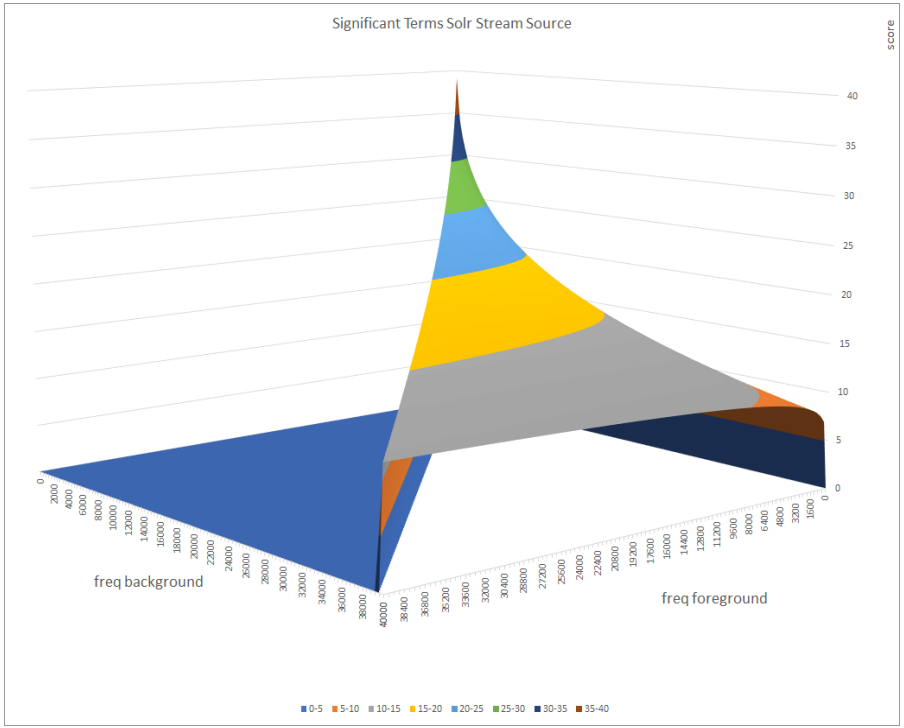
# Finding a technology for anomaly detection

Solr

## Significant Terms

Stream source




Solr's SignificantTerms score = `(log(freq_foreground)+1) * (log( (numDocs + 1)/(freq_background + 1) ) + 1)`

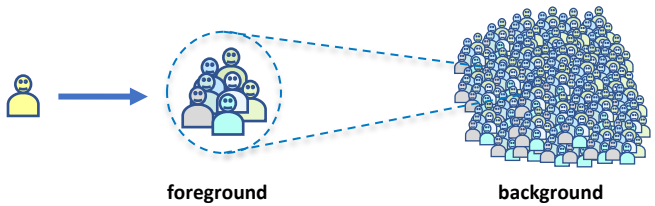
(SignificantTermsStream.java)

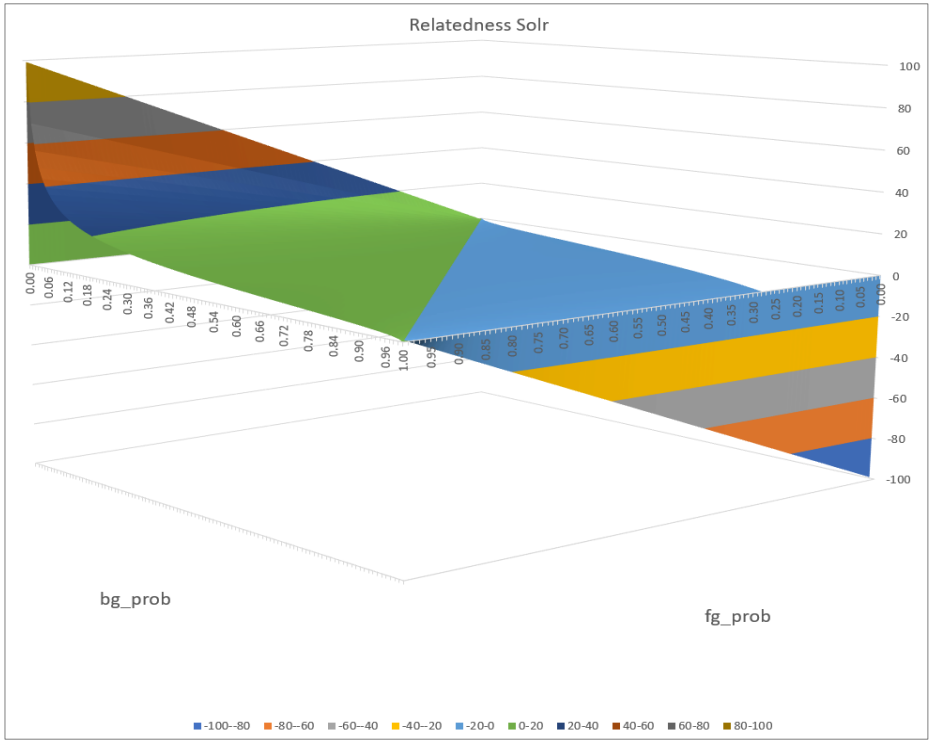
# Finding a technology for anomaly detection

Solr

## Reletedness() JSON Facet function

a.k.a. Semantic Knowledge Graph




$$\text{Z-score} = \frac{\texttt{fq\_count - fg\_size * bg\_prob}}{\texttt{sqrt(fg\_size * bg\_prob * (1 - bg\_prob))}}$$

(RelatednessAgg.java)

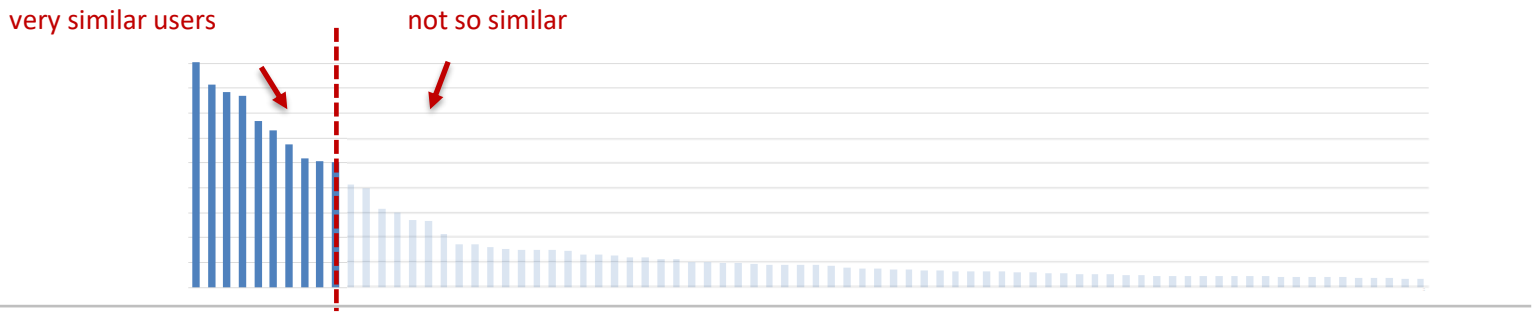
```
bg_prob = bg_count / bg_size
fg_prob = fg_count / fg_size
```

# Selecting similar users

- MLT standard Solr query parser to find all similar documents
- MLTPLUS home made query parser to only keep documents that are *very* similar.

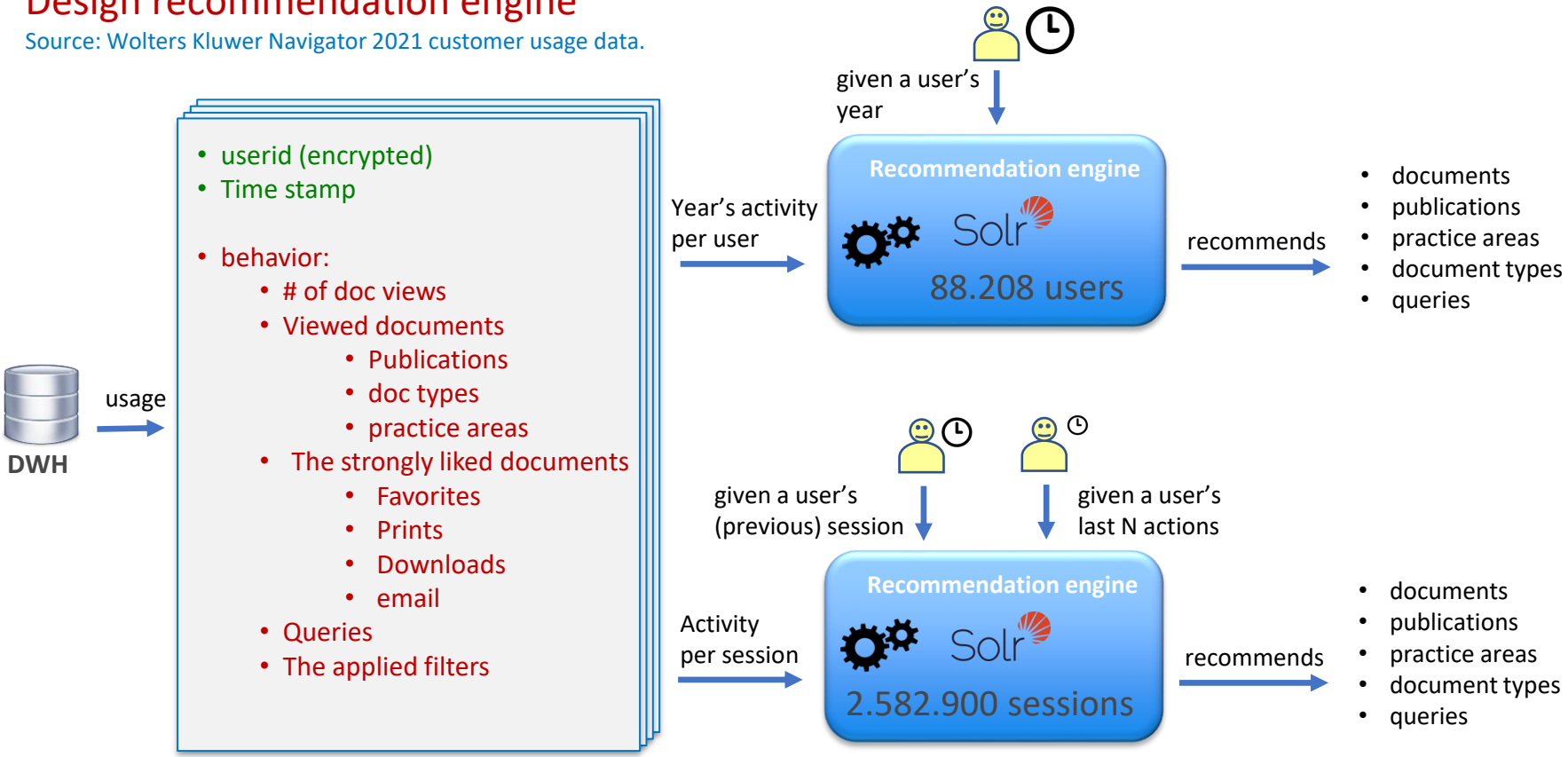
## Wrapping **mlt** into **mltplus**

```
(
  ({!wkmltplus minqt=3  mm=3  mmp=20 score=false}{!mlt v=$user maxqt=200 mintf=1 mindf=1 qf='suas'}) OR
  ({!wkmltplus minqt=10 mm=5  mmp=10 score=false}{!mlt v=$user maxqt=200 mintf=1 mindf=1 qf='docdocviews'}) OR
  ({!wkmltplus minqt=10 mm=10 mmp=70 score=false}{!mlt v=$user maxqt=200 mintf=1 mindf=1 qf='publicationdocviews'}) OR
  ({!wkmltplus minqt=10 mm=20 mmp=85 score=false}{!mlt v=$user maxqt=200 mintf=1 mindf=1 qf='palevel3'})
)
AND
({!filters param=$scope})
```

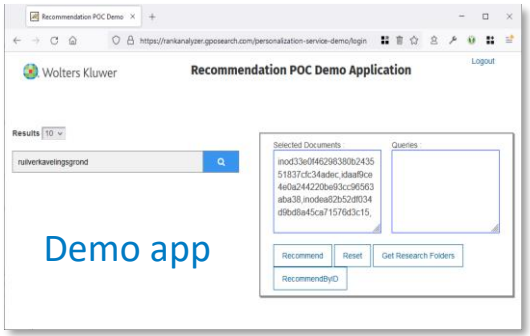


# Design recommendation engine

Source: Wolters Kluwer Navigator 2021 customer usage data.



# Personalization prototype




user's (last) year ID
*long-term personalization*

user's (last) session ID
*short-term personalization*

user's current actions
*Very short-term personalization*

content of personal folder

Recommended
- documents
- publications
- practice areas
- document types
- queries

**Personalization Service**

**Recommendation engine**

88.208 users

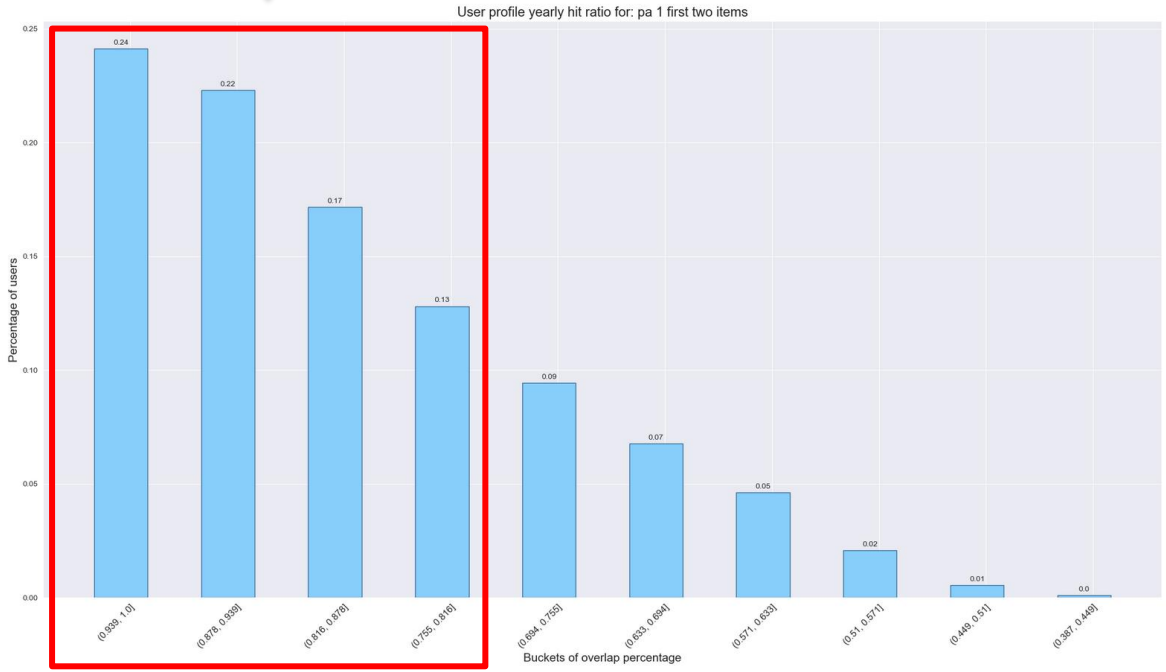
2.582.900 sessions

Solr

# What is the need for personalization based on a year history ?

## How consistent is the user's interest through the year?

32% users > 20 actions / year | 68% users < 20 actions / year

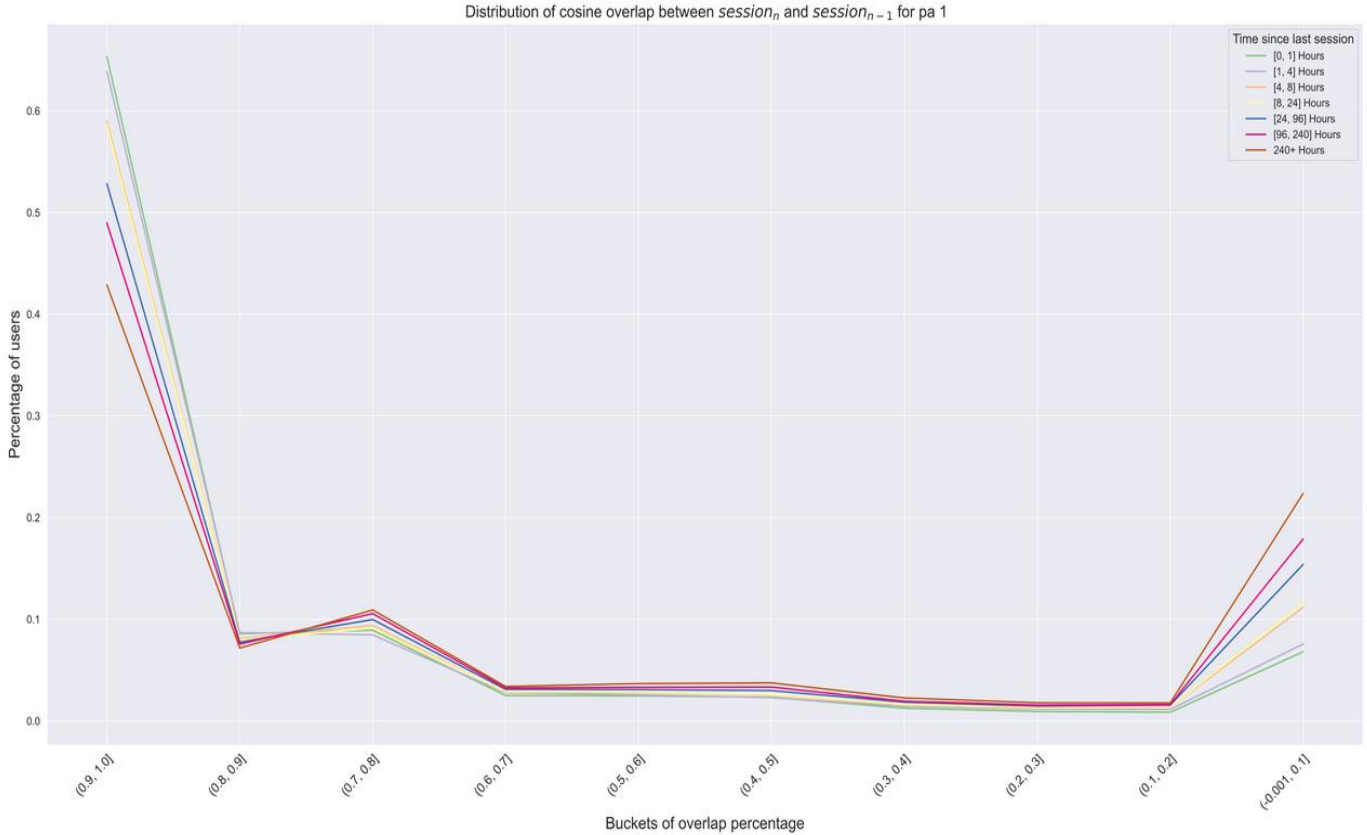



From the 32% most active users: 66% have viewed (relatively) consistently documents of the same 2 practice areas.

So for 21% of all users recommending PAs based on the activity of past year *might* make sense.

# What is the need for personalization based on the last session ?

## How much does the last session predict the interest of the next one?




For 42% of the users the previous session has a 0.9-1.0 consine overlap with the current one w.r.t. PA interest.

This rises to 66% when the time between sessions get shorter.

## How to measure success?

- Offline:
  - Manual tests by a domain expert
  - Build a model that predicts the likelihood that the previous session predicts the interest of the current one.
  - Replay user sessions and guestimate how much shorter they could have been in case personalization would have been applied.

- Online:
  - Measure clicks on explicit recommendations
  - A/B test (in case recommendations applied as boosts/filters)

# Lessons learned

## Yes we can

We can indeed infer the **user's interest via collaborative filtering** of user actions, using Solr and using usage data from a large Wolters Kluwer product

## What works best

Short-term personalization is applicable to **more users** than long-term personalization (may differ per business case)

## Quality logs needed

The accuracy of the recommendations depends on the **purity/completeness** of the **usage logs**. Preferred: front-end usage logging

## Challenge: testing

Offline testing of personalization is **hard**, but slightly easier for the short-term variant. Online testing **takes time**.

## We're not done yet

Improvement opportunities:

- Use **frond-end logs**
- Better data **cleaning**
- Query **normalization**
- Add **more signals** like filters & autocomplete clicks
- Testing strategies