

DB Digital Header:challenge

Digital Solutions for the DB

Till Plumbaum & Andreas Lommatzsch

 **DAI-Labor**
Distributed Artificial Intelligence Laboratory



DB Challenge

Innovate DB: Enter DB mindbox

The image shows three horizontal cards below a large red banner. The first card is yellow and labeled 'DB Accelerator' with an icon of a person speaking. The second card is green and labeled 'DB Challenges' with an icon of a person with a solution. The third card is blue and labeled 'DB Open Data' with an icon of a cloud. Each card has a small downward arrow pointing towards the red banner.

Innovating railway infrastructure

Pitch your idea

Your solution for our tasks

Dive into our data

DB Accelerator

DB Challenges

DB Open Data



One 4.0 initiative of DB

Three major instruments

One fabulous space for coworking and exchange

DB Challenge

We want you contributing new solutions

DB Challenges

About DB Challenges

Challenges are well-defined problems for which we do not have a suitable solution yet. Therefore we challenge you to compete for the best solution.

We are looking for

We seek to find both technical – mechanical and ICT solutions, as well as new approaches, for services and business models in the context of digitalization. We invited everyone to compete.

Your benefit

The winner of a challenge will receive a monetary price. In addition we are always interested in cooperating with people which have proven to solve problems fast and professionally.

1. DB Information:challenge
Dec. 2015 – Feb. 2016
2. DB Information UX:challenge
March 2016 – April 2016
3. DB Digital Header:challenge
March 2016 - April 2016

<https://www.mindboxberlin.com/index.php/challenges.html>

DB Digital Header:challenge

The challenge

Supporting the DB to enter the age of digitalization

- The Deutsche Bahn has large numbers of scanned, but not digitalized, documents, describing various technical systems
- Transferring the information in these documents into a searchable, digitalized structure will ease the daily work of DB employees
- The process of transferring the information currently means manual work and is very time consuming
- Goal of the challenge is to find and develop automated (or semi-automated) approaches to do this

<http://gmedia.de/dbdigitalheader/>



What exactly is the challenge all about?

What is a digital header?

The image shows a comparison between a physical paper-based cablesplice plan and its digital representation.

Physical Plan (Left):

- A:** Kabelspleißplan für GSM-R Stich. It includes a table for Streckenkabel F 4299 26" (6/20), a table for Stichkabel F 20" (-/20), and a table for Endverschluß TEV 68. Below these are two connection diagrams: "Richtung Ehingen" (Str. 4540) and "Richtung Sigmaringen" (Str. 4540). A legend indicates "GSM-R".
- B:** A detailed table for Stichkabel F 20" (-/20) showing individual conductor assignments (KLP-Nr., Art und Größe).
- C:** A table for Lagenaufbau Stichkabel F 20" (-/20) showing the arrangement of conductors in the cable.
- D:** A table for Stichkableinführungen listing connection points (BTS numbers, locations, and km values).
- E:** A table for Änderung (Change) showing a row for "Änderung" with columns for Gez. (Approved), Datum (Date), Gepr. (Inspected), Datum (Date), Überre. (Delivered), and Datum (Date).
- F:** A table for Maßstab (Scale) showing a row for "Maßstab" with columns for Format (Format), DIN A4, Überre. (Delivered), 29.01.2004, Stefan Mäder, and Urspr. Dlks sp 20 1 GSM R.

Digital Header (Right):

- H1:** Änderung (Change) table with columns for Gez. (Approved), Datum (Date), Gepr. (Inspected), Datum (Date), Überre. (Delivered), and Datum (Date).
- Header Information:**
 - Verwaltung: DB Dienstleistungen Systel Regionalbereich Südwest
 - Eigentümer: DB Infrastruktur Netz Regionalbereich Südwest
 - Stichkableinführung F 20" (4x10x0,9)
 - Strecke 4540 Ulm Hbf - Sigmaringen
 - Streckenfernmeldekabel F 4299 26" (6/20)
 - Ehingen - Sigmaringen
 - alte Bezeichnung: F 44.4540 -
- Project Description:** Kabelspleißplan TK
- Blatt:** Blatt 6457009120_Tkks_Spleißplan_Fs20
- Table Headers:**
 - Tkks
 - DB Systel - intern
 - Bestand
 - Format: DIN A4
 - Maßstab: ohne
- Table Data:**
 - Gez. 18.08.2003 Schäfer/Natterer
 - Gepr. 18.08.2003 Schäfer
 - Übere. 29.01.2004 Stefan Mäder
 - Ers. f.: 6457009120_Tkks_Spleißplan_Fs20

Extracting information

Detecting the document type

List of document types

Aufstellungsplan TK
Übersichtsplan
Systemübersichtsplan
<u>Grundrißplan</u>
Kabellängen- und <u>Pupinisierungsplan</u>
Kabellängenplan TK
Kabelzubehörplan
Kabelübersichtsplan TK
Rangierfunkanlage
Stromlaufplan TK

Document header

Nr.	Änderung	Gez.	Datum	Gepr.	Datum	Übere.	Datum
DIS-Beschreibung:				Maßstab: ohne Format: 297x594			
TKKZ	2057010379	Eigentümer: DB Netz AG DB Niederlassung Süd	Planzustand: Bestand				
		Bearbeitet durch: DB Telematik DB Region Süd	Kabelzubehörplan Bf Rosenheim				
			LWL-Kabel F771226 24' BASA Rosenheim - POP Rosenheim				
Datum Name				Dokument-ID: Blatt 01			
Gez. 09.11.06 Gauert				01 Bl.			
Bearb. 09.11.06 Horwitz							
Gepr.							
Aenderung	Datum	Name	Urspr.	Ers.f.: Zubehoer_F771226.dwg			

Extracting Information

1 Nr. Änderung		2 3 4 5 6 7 Gez. Datum Gepr. Datum Übere. Datum					
DIS-Beschreibung: 8							
28 TKKZ	29 2057010379	Eigenümer: 9 DB Netz AG DB Niederlassung Süd		Maßstab: ohne 11 Format: 297x594 12			
Bearbeitet durch: 10 DB Telematik DB Region Süd							
Kabelzubehörplan Bf Rosenheim LWL-Kabel F771226 24' BASA Rosenheim - POP Rosenheim							
		Datum	Name				
		Gez. 09.11.06 14	Gauert 15				
		Bearb. 09.11.06 16	Horwitz 17				
		Gepr. 18	19				
Änderung	Datum	Name	Urspr. 26	Dokument-ID: 20		Blatt 21 01	... 01 Bl 22
Ers.f.: Zubehoer_F771226.dwg 27							
23	24	25					

Document header

1 - 8: empty field

9 Owner: - DB Netz AG Niederlassung Süd

10 Processed by: - DB Telematik Region Süd

11 Scale: - none

12 Format: - 297x594

...

21 Page - 01

22 Pages in total: 01

...

29 ID: 2057010379

...

Overall 52 different fields for the different plan types



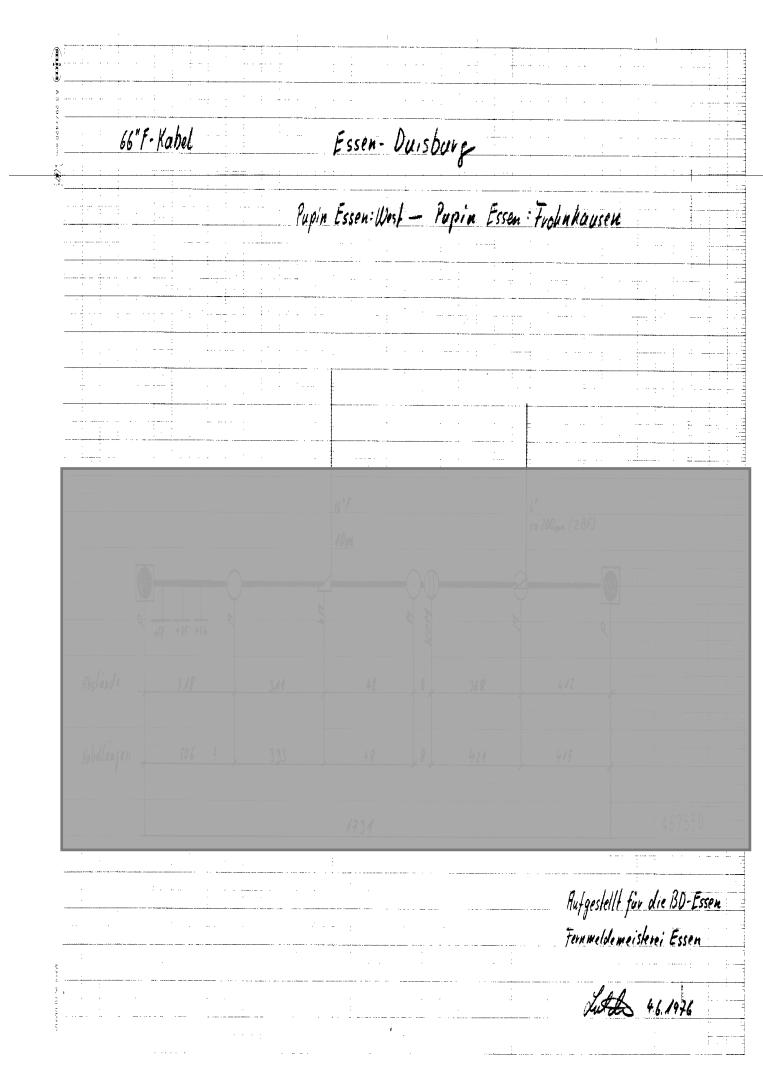
Is it really that simple?

Complicated headers

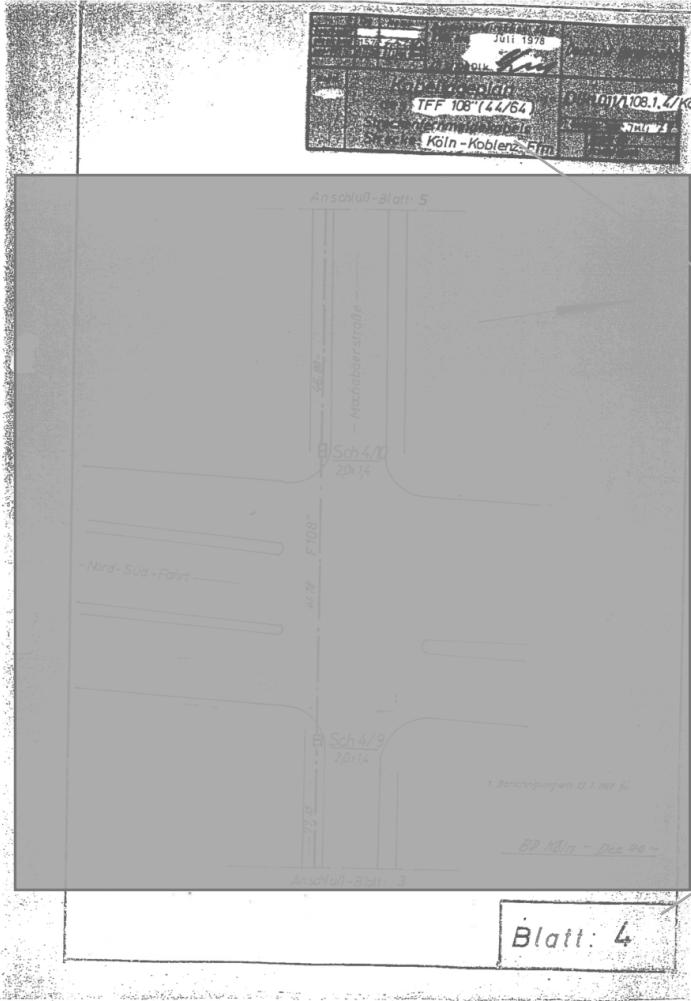
NO!

The document collection includes different types from different times

Some are hand-drawn / hand-written



Complicated headers



In some documents the information is blacked out

Köln-Koblenz and Page 4

Different header formats

- ▶ The good news – the most complicated or unreadable documents were removed
 - ▶ But there is still a wide variety of documents

	DA 1 ... 8	1. Lage 0,9
	DA 9 ... 18	
	DA 19 ... DL	2. Lage 0,9
	DA 20 ... 28	
	DA 29 ... 36	
	DA 37 G11 A → B	
	DA 38 G12 A → B	
	DA 39 G13 A → B	
	DA 40 G14 A → B	
	DA 41 ... 48	
	DA 49 G11 A ← B	3. Lage 1,4
	DA 50 G12 A ← B	
	DA 51 G13 A ← B	
	DA 52 G14 A ← B	

Abzweigverbindung

Deutsche Post

The Challenge

We are looking for an algorithm which will allow us to automatically extract and organize documents such as blueprints and plans. These are our main goals:



Auto detection of various header types



Header extraction & transfer into data structures



Portability of the algorithm onto unknown headers

What you have to do to win

Challenge consist of three parts:

- ▶ Automatized identification of different headers and sorting these into different stacks – known, unknown, indecipherable.
- ▶ Extraction of information from known headers and text blocks and subsequent transfer into an organized data structure.
- ▶ Portability of the algorithm onto unknown headers.

The winner is selected by the combined results for the three tasks.

1. Detect document type 20%
2. Extract information 50%
3. Portability 30%

End of April, a jury of DB representatives and external experts announce the winners and award the prizes.

Why should you participate?

The three winner(-teams) will get :

- ▶ a great opportunity to solve a real-world problem
- ▶ Price money of 14.000 Euro overall
(8.000 for the first, 4.000 for the second, and 2.000 for the third)
- ▶ A great reference for your CV

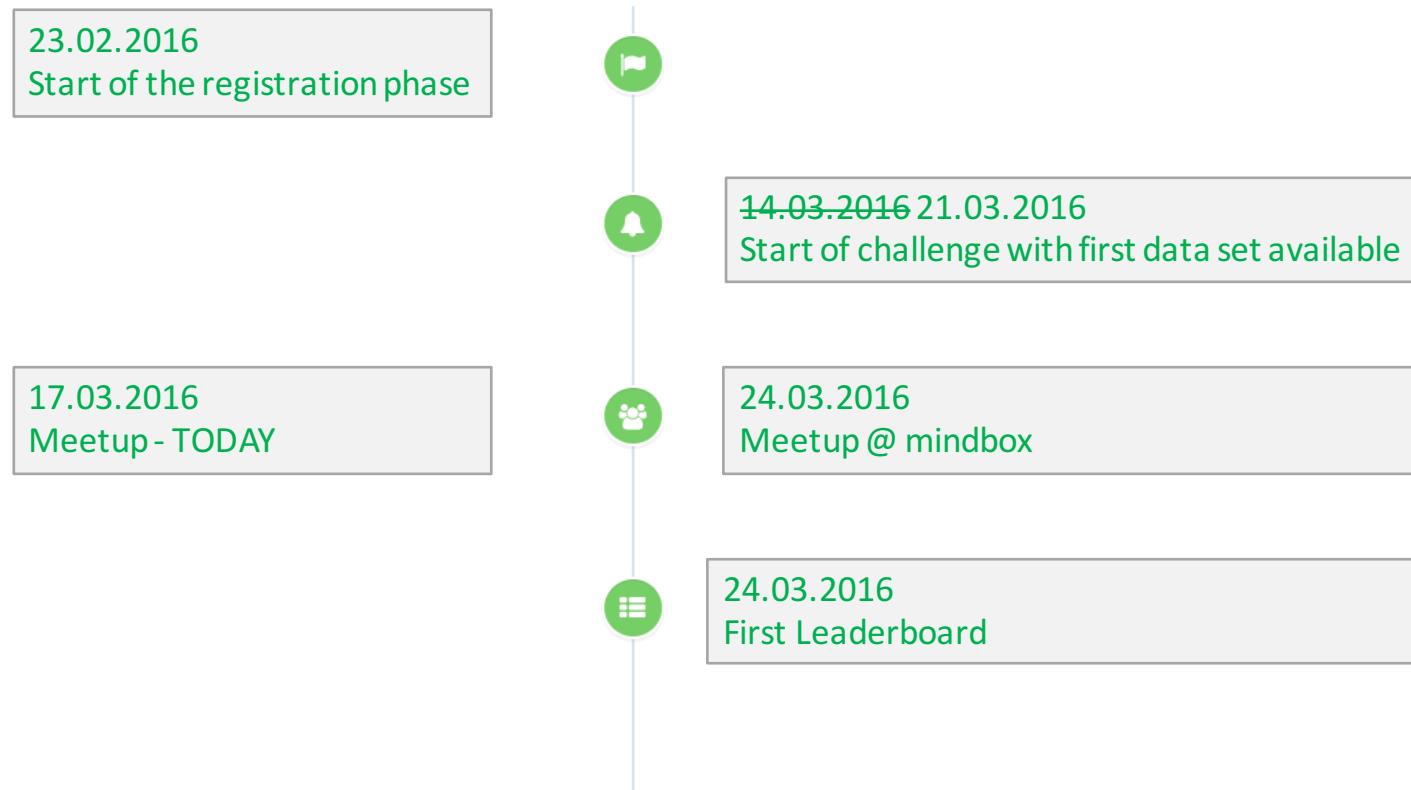


What data you get?

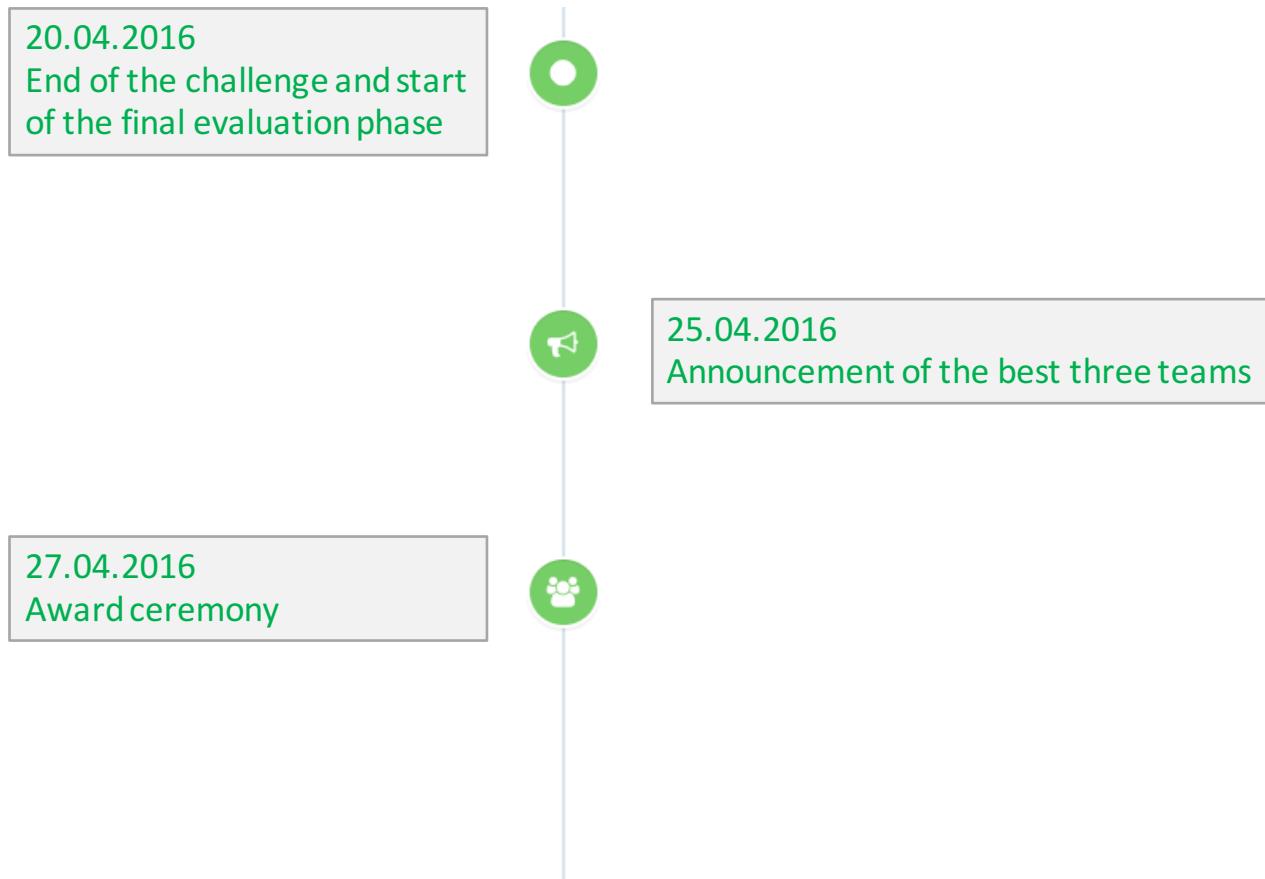
- ▶ First data set – available on Monday
 - Around 900 documents
 - Sample solutions for X different plan types
 - A list of all possible document types
 - A structured solution file for the answers
- ▶ Second data set – for the final scoring
 - Around 900 documents
 - No sample solutions
 - Limited time to solve

The Challenge: What to do and when

Timeline



The Challenge: What to do and when



Ways to approach the challenge

- ▶ Character recognition (OCR)
 - Most likely the first step...
- ▶ If you have text
 - Error correction
 - robust matching of words
 - Ontologies or/and dictionaries (for names, types, etc.)
- ▶ If you work with images
 - Robust matching of parts of the image
- ▶ Detect Pattern and known segments
- ▶ small number of documents impedes methods such as Deep Learning

0	1	1	1	0	1	0	0	0	1	0	1	0	HOME	FORUM	REGISTER	LOGIN	0	1	0	1	0	1	0	1	0	1
1	0	1	0	1	{	1	0	0	//	1	1	{	0	1	0	1	1	1	0	1	0	1	0	1	0	1
0	1	*/	0	0	1	0	1	0	0	0	1	0	#	1	&	1	0]	0	1	0	1	0	1	0	1
1	0	1	0	1	:	1	{	0	0	1	1	1	0	1	[1	!	1	-	1	0	1	0	1	0	1
0	0	0	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	;	1	0	1	0	1	0	1	1
1	0	1	0	#	[1	1	1)	1	0	0	1	0	1	0	1	*	1	0	1	0	1	0	1	0
0	1	/*	1	0	1	0	1	?	1	0	1	<	0	>	0	1	0	1	0	1	0	1	0	1	0	1
1	0	1	0	1	0	1	>	1	0	1	0	0	JETZT REGISTRIEREN	1	0	1	0	1	0	?	0	1	0	1	0	1
0	[0	1	0	1	0	1	0	1	0	1	0	-	0	1	1	0	<	0	>	±	1	0	1	0	1
1	0	1	&	1	1	1	0	1	0	%	1	0	1	1	1	1	1	0	1	0	1	0	1	0	1	0
0	1	?	1	++	1	0	:	0	1	1	1	0	1	0	1	0	1	0	1	1	0	1	0	1	0	1
1	0	1	0	0	1	0	≤	1	1	0	0	1	0	1	0	1	0	-	1	0	1	0	1	0	1	0
0	1	0	//	0	1	0	1	0	1	0	1	0	*/	0	1	+	1	0	0	1	0	1	0	1	0	1

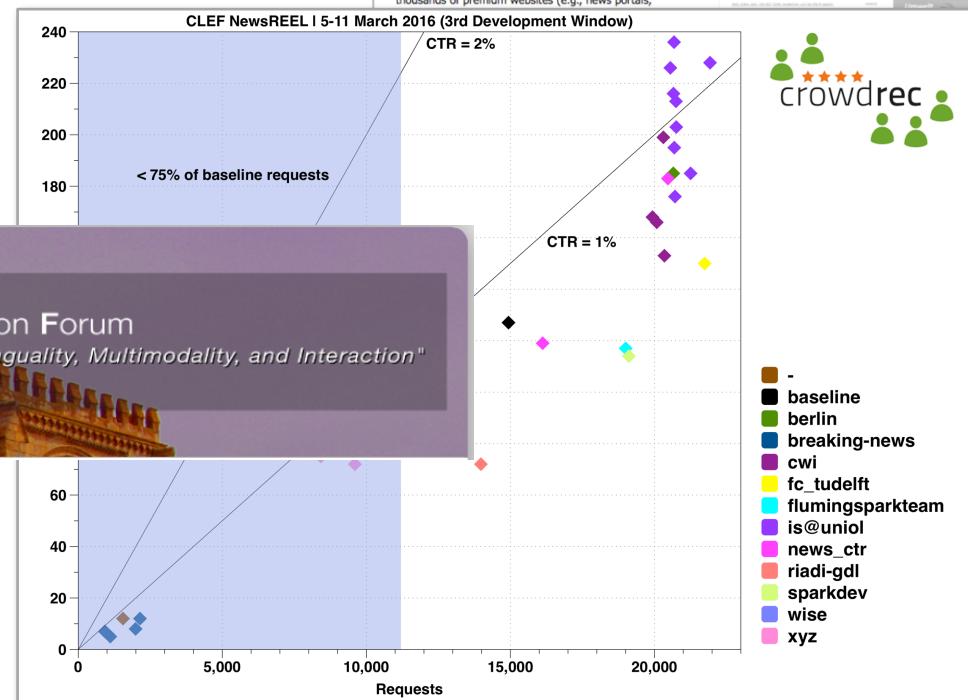
Digital Header Challenge

Digitale Lösungen
für die DB

<http://gmedia.de/dbdigitalheader/>

- News Recommendation
- Task 1:
 - receive recommendation requests from visitors
 - evaluation from 2-29 April 2016
- Task 2: predict how visitors read news with recorded data

Results will be presented at



CLEF-NEWSREEL NEWS RECOMMENDATION EVALUATION LAB

[HOME](#) | [TASKS](#) | [HOW TO PARTICIPATE](#) | [PUBLICATIONS](#) | [ORGANISATION](#) | [PREVIOUS CAMPAIGNS](#) | [CLEF 2016](#) | [TUTORIALS](#)

Overview

Many online news publishers display on the bottom of their articles a small widget box labelled "You might also be interested in", "Recommended articles", or similar where users can find a list of recommended news articles. Dependent on the actual content provider, these recommendations often consist of a small picture and accompanying text snippets.

While some publishers provide their own recommendations, more and more providers rely on the expertise of external companies such as plista, a data-driven media company which provides content and advertising recommendations for thousands of premium websites (e.g., news portals,



Important Dates

Labs registration opens: 30 October 2015
 Registration Closes: 22 April 2016
 End Evaluation Cycle: 4 May 2016
 Submission of Participant Papers [CEUR-WS]: 25 May 2016
 Submission of Lab Overviews [LNCS]: 3 June 2016
 Notification of Acceptance
 Lab Overviews [LNCS]: 10 June 2016
 Camera Ready Copy of Lab Overviews [LNCS] due: 17 June 2016
 Notification of Acceptance
 Participant Papers [CEUR-WS]: 17 June 2016
 Camera Ready Copy of Participant Papers and Extended Lab Overviews [CEUR-WS] due: 1 July 2016
 CEUR-WS Working Notes Preview for checking: 22 July 2016
 CLEF 2016: 5-8 September 2016

@clefnewsreel

Having any questions before the evaluation round starts? Get in touch with us in time!
 #recsys #clef2016 01:05:37 PM March 14, 2016 from Twitter Web Client
 ReplyRetweetFavorite
 3rd and final dev phase finished. 2 weeks left to prepare until 2 April when the evaluation starts #recsys #clef2016 https://t.co/nbGRge7wN 01:03:51 PM March 14, 2016 from Twitter Web Client
 ReplyRetweetFavorite
 Results of the 2nd Dev

Close proximity to research institutes



TEL Building at Ernst-Reuter-Platz

- Deutsche Telekom Innovation Laboratories (T-Labs)
- Connected Living Innovation Center
- DAI-Labor / TU Berlin
- German Turkish Advanced Research Centre for ICT (GT-ARC)
- EIT Digital (Berlin Node)
- Daimler Center for Automotive IT Innovations



Telekom Innovation Laboratories



DAI-Labor
Distributed Artificial Intelligence Laboratory



DAIMLER

Expert centers:



Spin-Offs:



Staff & Partners

15 Post-Doctoral Researchers

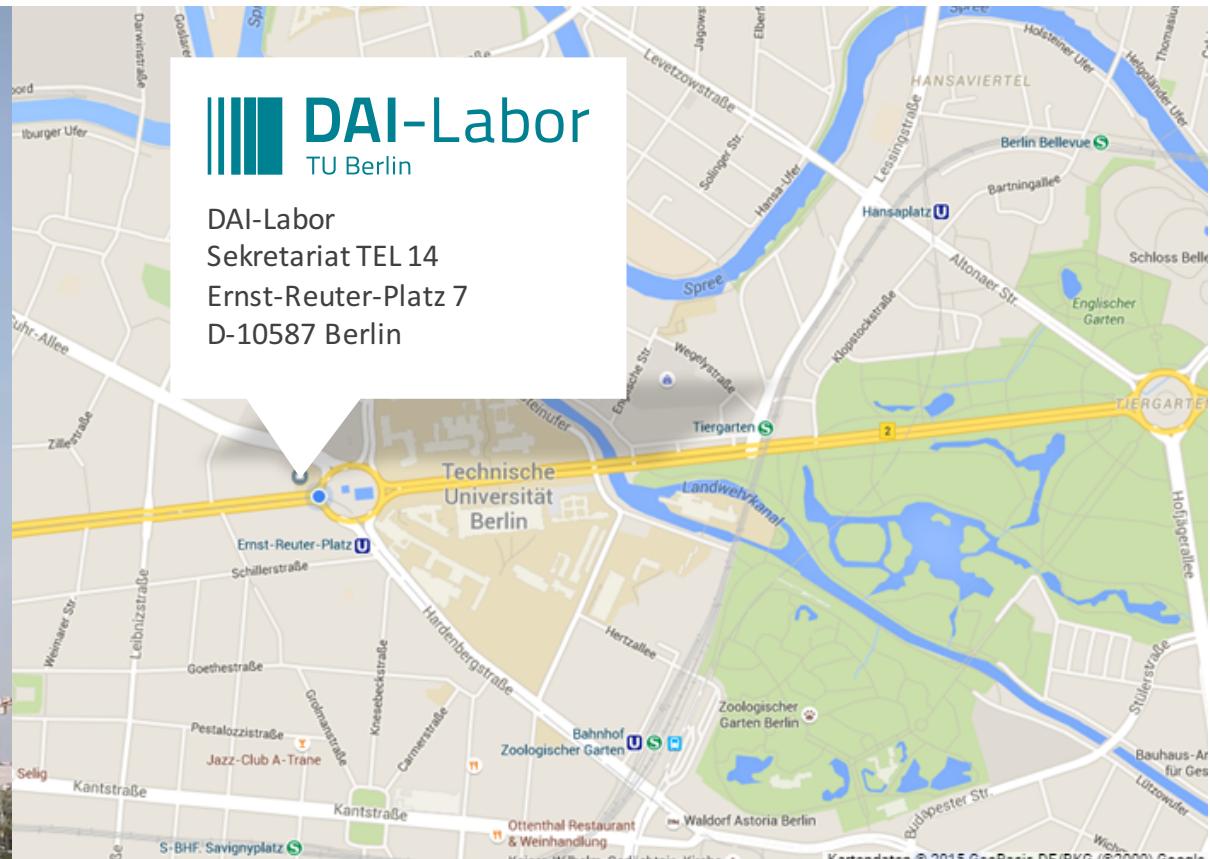
50+ Full-time Doctoral Researchers (Ph.D. Candidates)

50+ Half-time Student Research Assistants

8 Support Staff: Secretaries, System Admins & Graphics Designers

Strong Cooperation with Industry and Research Institutions





Get In Touch



sahin.albayrak@dai-labor.de

Prof. Dr. Dr. h.c. Sahin Albayrak



+49 30 - 314 74000