Three perspectives on a collaborative attempt to use computer vision techniques to automatically classify historical newspaper images

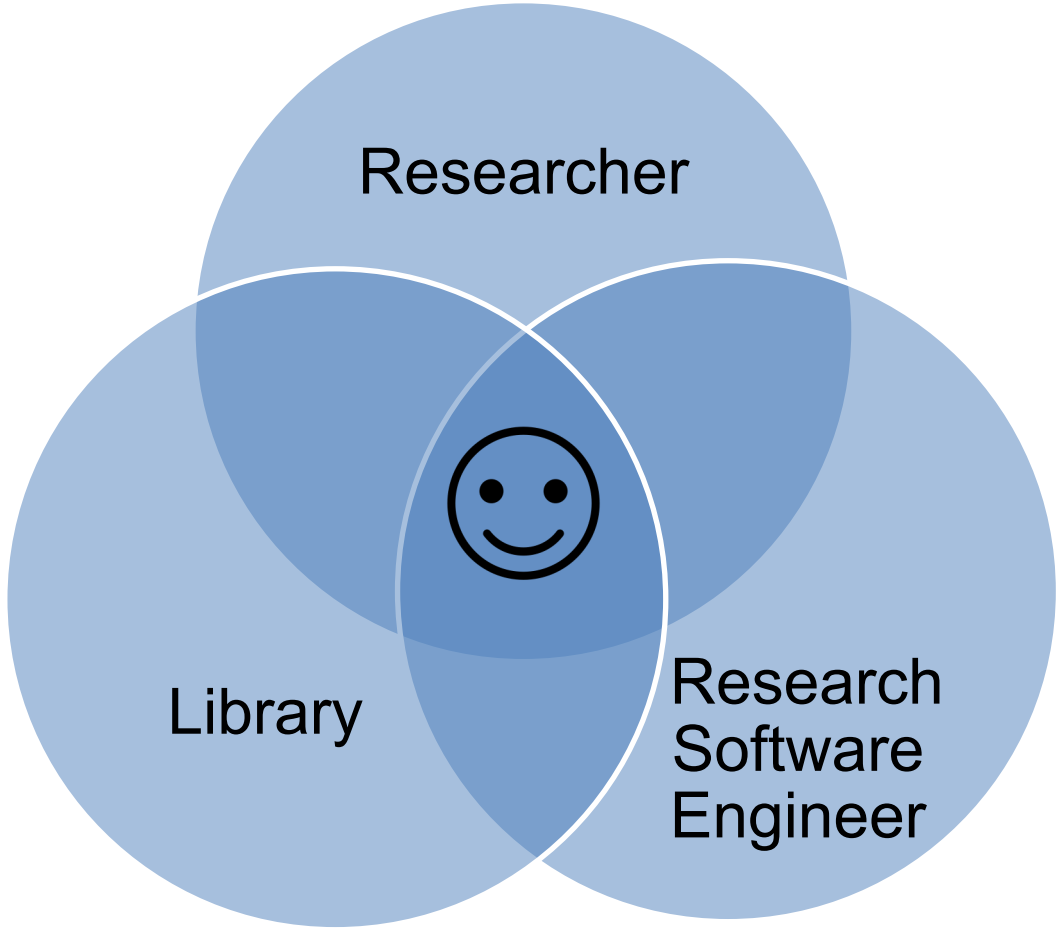
Martijn Kleppe, National Library of the Netherlands (KB)
Thomas Smits, Utrecht University
Willem Jan Faber, National Library of the Netherlands (KB)



DHN2019 workshop
Twin Talks: Understanding Collaboration in DH at DHN 2019
5 March 2019

Outline

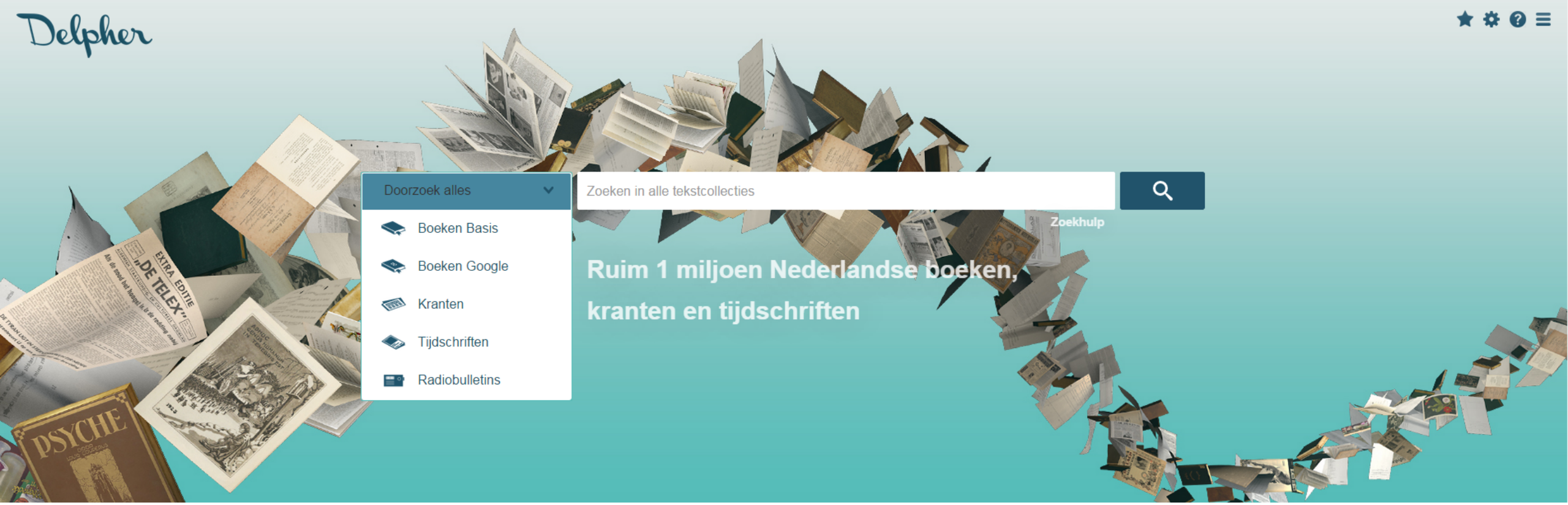
- **Introduction**:
  - Humanities perspective
  - KB perspective
  - Technical perspective
- **Results**
- **Discussion**



I. Intro - Humanities Perspective

- Historian
- Dissertation on ‘European illustrated press and the emergence of a transnational **visual culture** of the news, 1842-1870’
- Plus: Interested in **tipping point dominance photographs** in newspapers
- RQs: When did Dutch newspapers switch to **using photographs as the primary visual medium**?

Utrecht University



Full text (OCR) access to:

- ✓467.000 books (1486 – 2013)
- ✓15 million newspaper pages (1618 – 1995)
- ✓4,4 million magazine pages (1840 – 1940)
- ✓1,5 miljoen ANP-radiobulletins (1937 – 1984)

www.delpher.nl/

Delpher

Kranten

Kopenhagen

Uitgebreid zoeken

Menu

859.354 krantenartikelen gevonden

Zoekterm Kopenhagen ×

Sorteer op relevantie

Weergave

Periode

- 18e eeuw (1047)
- 19e eeuw (107498)
- 20e eeuw (750809)

Kies periode...

Soort bericht

- Advertentie (28475)
- Artikel (825361)
- Familiebericht (1552)
- Illustratie met onderschrift (3966)

DEENSCHE ELFTAL SAMENGESTELD. Sterkste deel is de verdediging.

DEENSCHE ELFTAL SAMENGESTELD. Sterkste deel is de verdediging. KOPENHA-GEN, 17 Oct (Eigen lel.) — De keaze-conimissie, van den Deenschea voetbalbond heeft hedenavond bet elftal samengesteld, dat …

Krantentitel De Telegraaf

Datum 18-10-1938

Meer details

VOETBAL. HET DEENSCHE ELFTAL. Tegen Oranje.

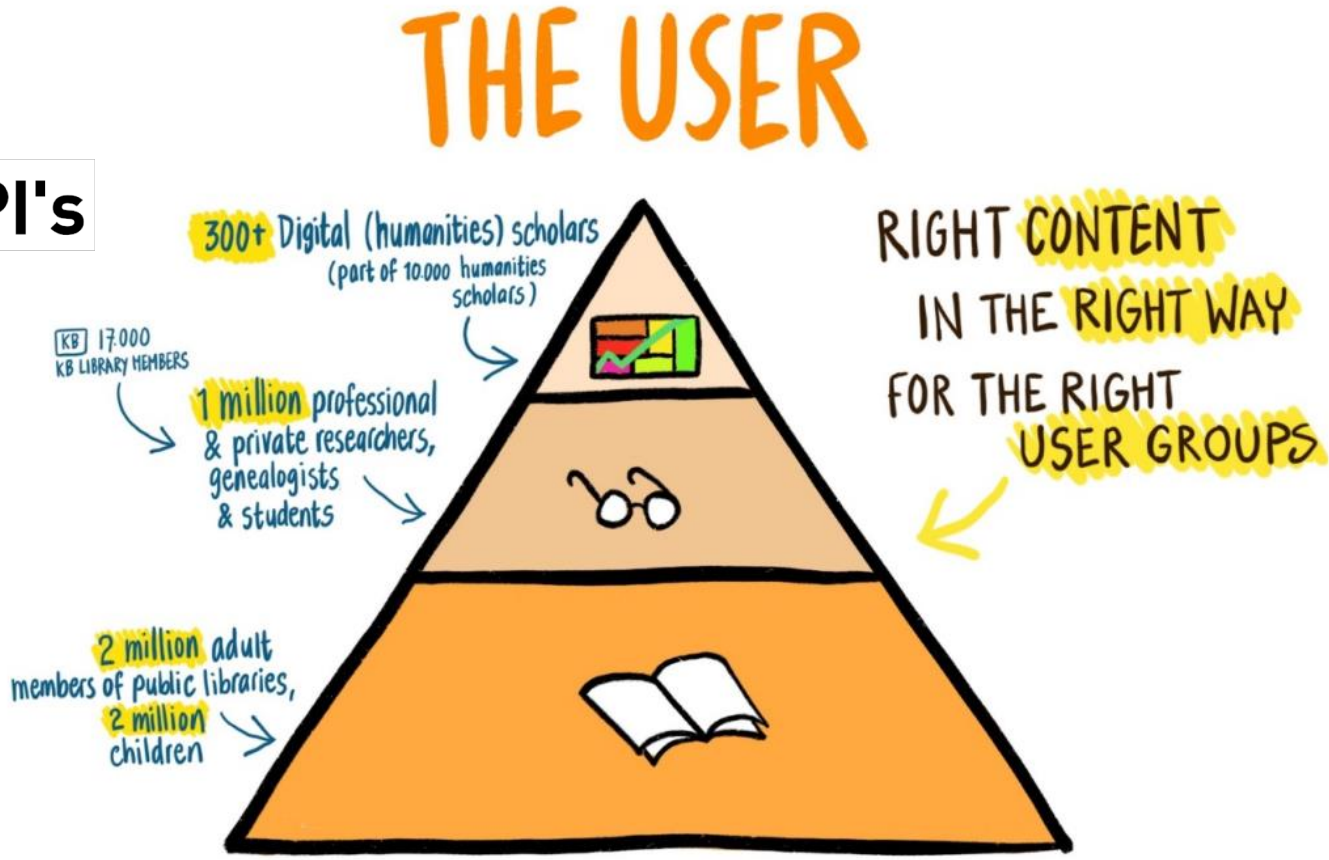
rno Nielsen (B. 93 Kopenhagen), Albreohlsen (A.B. Kopenhagen), Knudsen (Vejen) en Knud Larsen (Frem Kopenhagen). Slechts één debutant, n.l. de jonge middenvoor Rein-hold Nielsen. Zooals b …

Krantentitel Nieuwsblad van Friesland : Hepkema's courant

I. Intro - KB Perspective

Dataservices en API's

Delpher



https://www.kb.nl/en/resources-research-guides/data-services-apis

DENKSCHETS.NL

nl

I. Intro - KB Perspective

Researcher-in-residence program

- Understand user needs
- Co-develop new tools to open up digital collections
- (Create KB ambassadors
- Academic network & follow-up projects)

https://www.kb.nl/en/organisation/research-expertise/researcher-in-residence

I. Intro - KB Perspective

Researcher-in-residence program

- 2 early talented early career researchers per year
- Open call & review committee
- 6 months, 0,5 fte
- Technical & research support from KB Lab
- Work at KB Office
- Tools and/or data always available for others

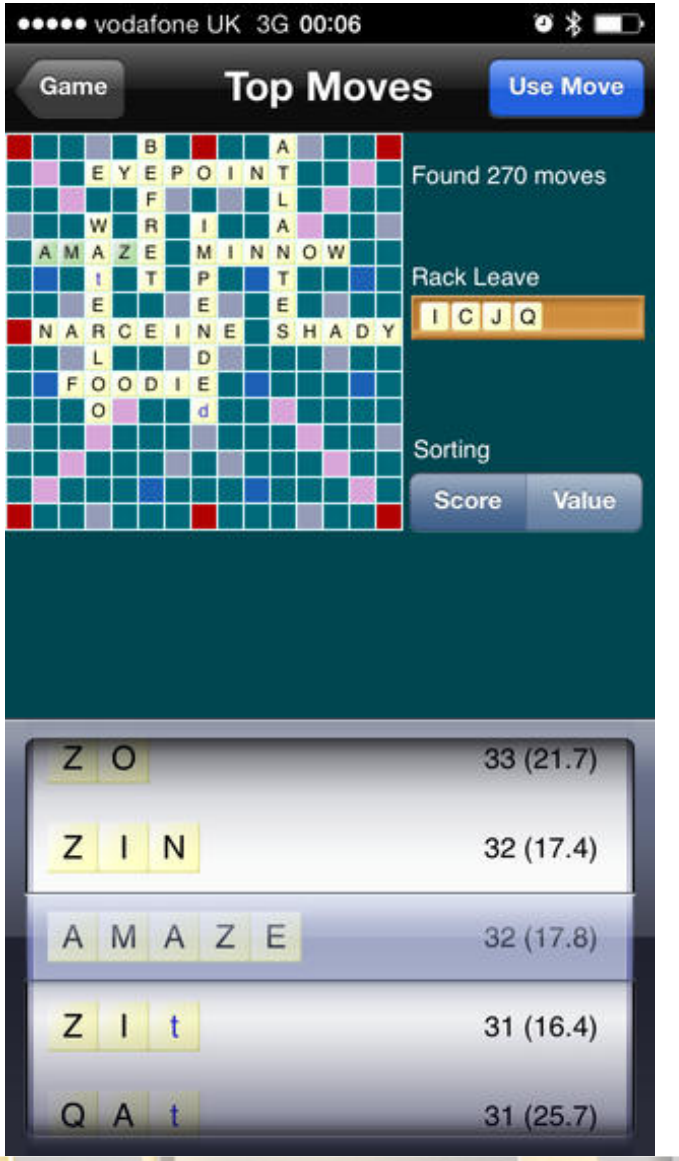
http://lab.kb.nl/

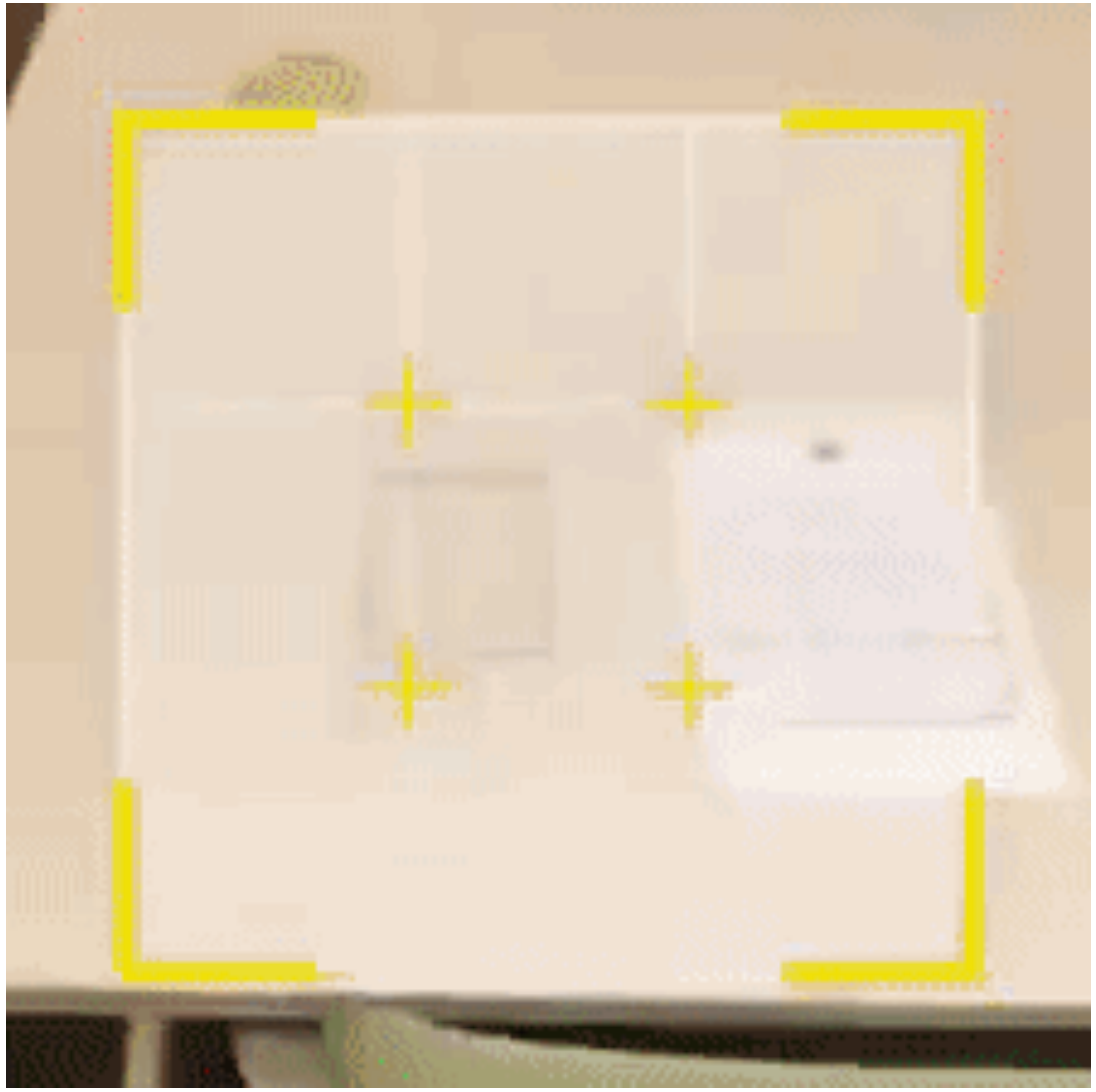

ARCADE

I. Intro - Technical Perspective

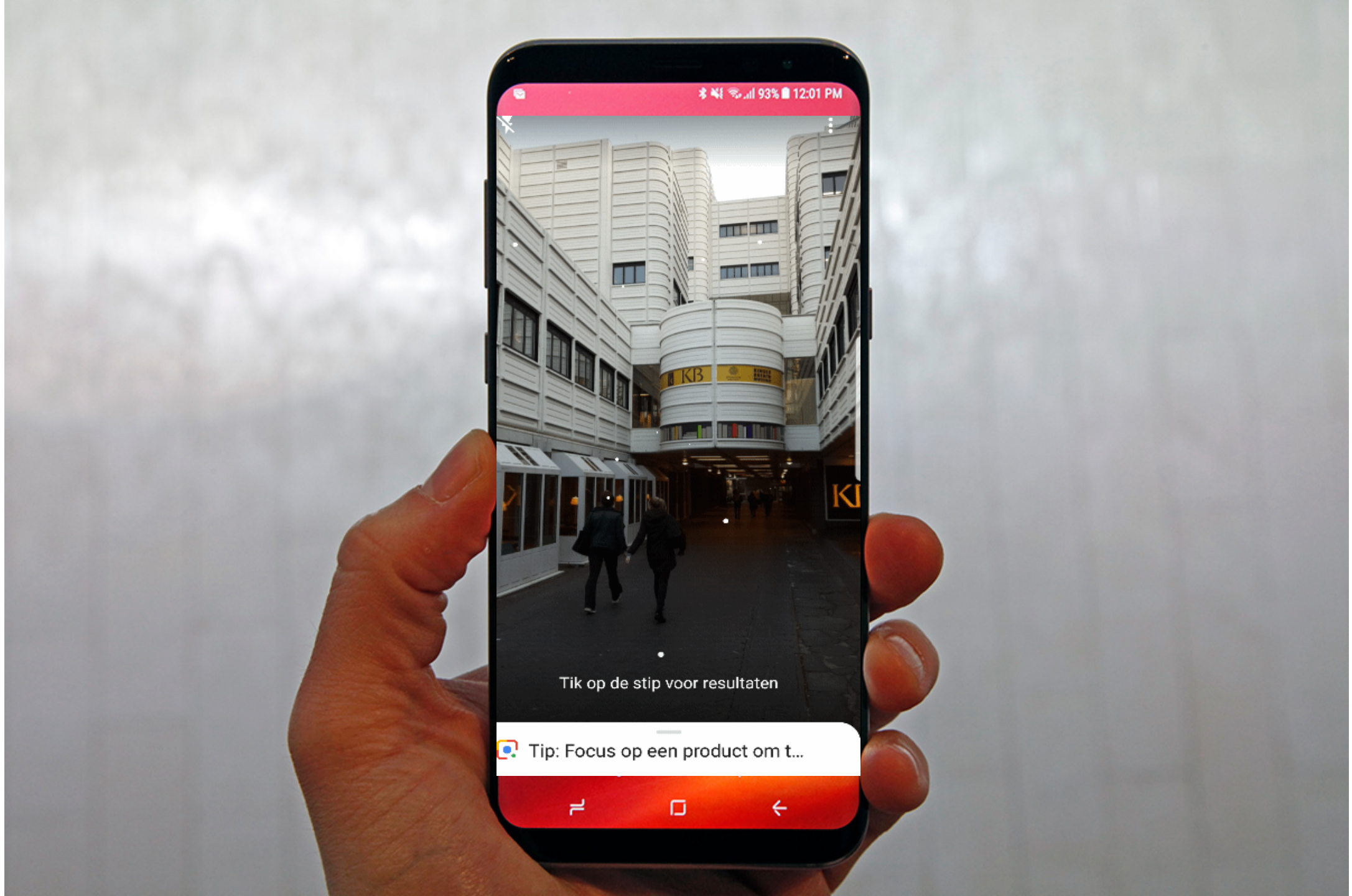
- Project Thomas: When did Dutch newspapers **switch to using photographs** as the primary visual medium?
- As a library mainly used to working with **texts that have been OCR'ed**
- But: newspapers, books & magazines **contain loads of images that are not described**, except for the caption
- Let's experiment with **computer vision**!

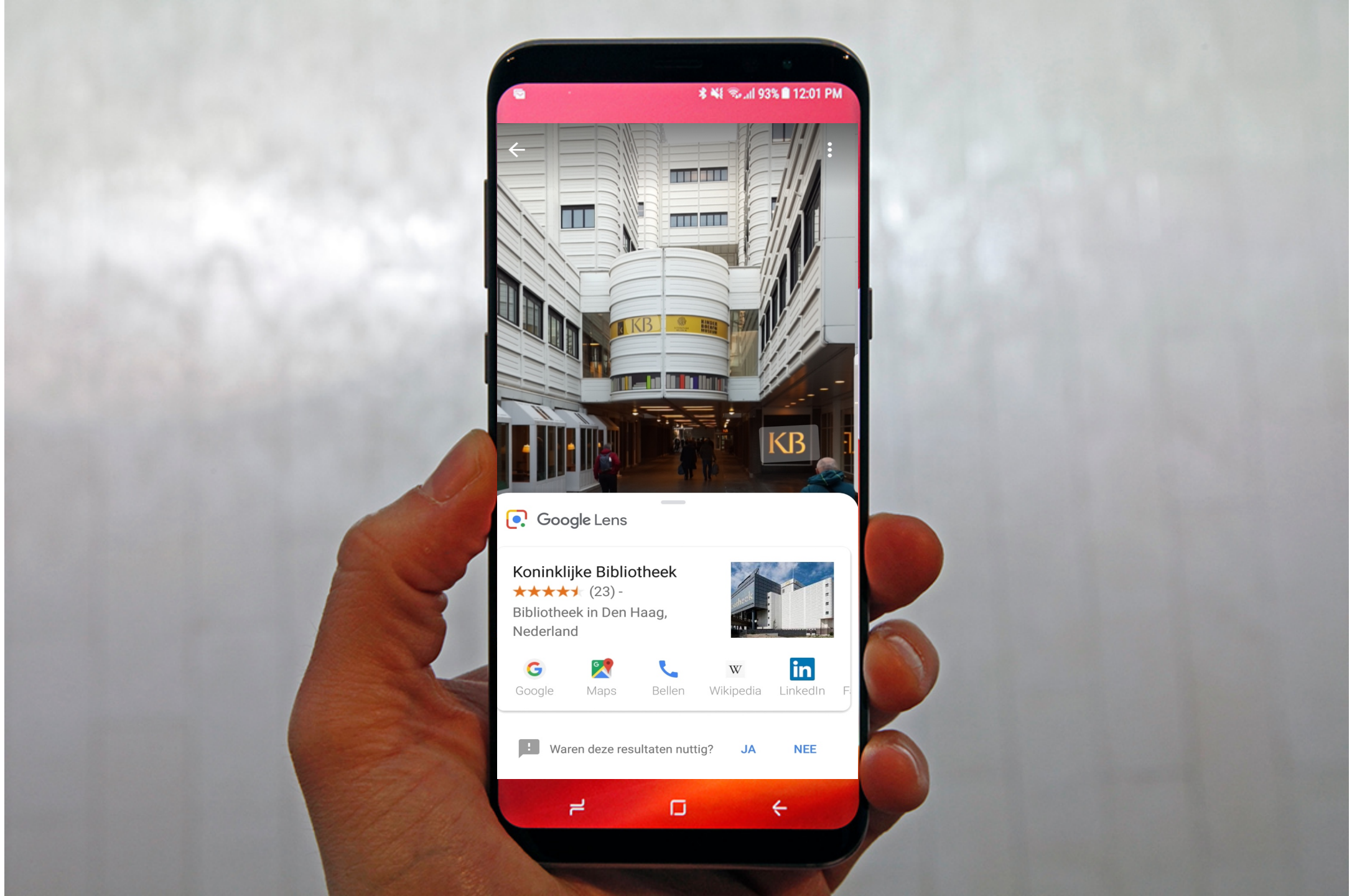

vodafone UK 00:04
New
Unsaved Game
Solve
I Z C J M A Q


vodafone UK 3G 00:06
Game
Top Moves
Use Move
Found 270 moves
Rack Leave
I C J Q
Sorting
Score
Value
Z O 33 (21.7)
Z I N 32 (17.4)
A M A Z E 32 (17.8)
Z I t 31 (16.4)
Q A t 31 (25.7)



https://blog.prototypr.io/behind-the-magic-how-we-built-the-arkit-sudoku-solver-e586e5b685b0


93% 12:01 PM
KB
Tik op de stip voor resultaten
Tip: Focus op een product om t...


93% 12:01 PM
KB
KB
Google Lens
Koninklijke Bibliotheek
(23) -
Bibliotheek in Den Haag, Nederland
Google
Maps
Bellen
Wikipedia
LinkedIn
Waren deze resultaten nuttig?
JA
NEE

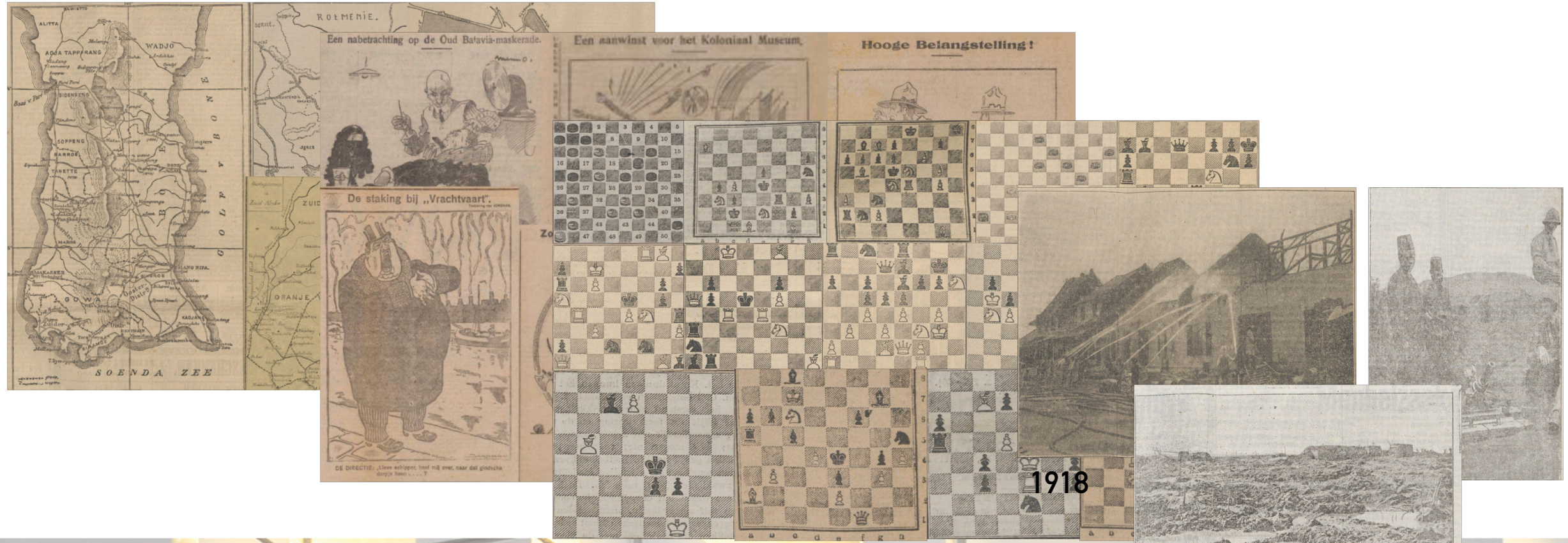
I. Intro - Technical Perspective

Steps:

1. Set up **Hardware**: Images demand more computing power than text > scale up & GPUs needed: Super computing SURF SARA
2. **Collect data** from Delpher
3. Create **pipeline**:
    1. Binary classifier drawing - non drawing
    > Train: **Thomas** created labelled set of images

II. Results

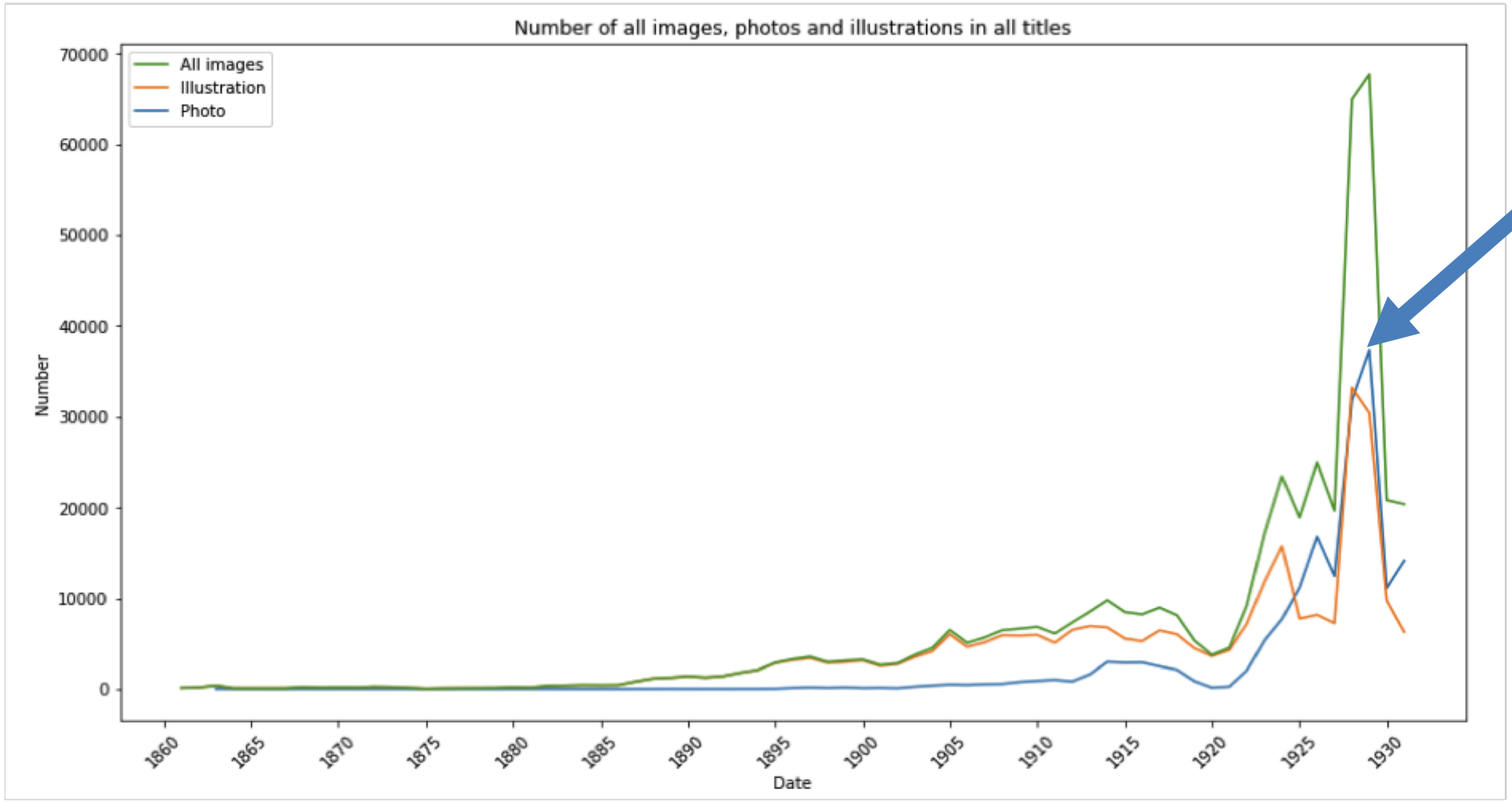
- Technical perspective: **Managed to classify** images



II. Results

- Technical perspective: Managed to classify images

- Humanities perspective: We **pinpointed** moment in time photographs were predominantly being used


Number of all images, photos and illustrations in all titles
All images
Illustration
Photo
Number
Date
70000
60000
50000
40000
30000
20000
10000
0
1860
1865
1870
1875
1880
1885
1890
1895
1900
1905
1910
1915
1920
1925
1930

1927

II. Results

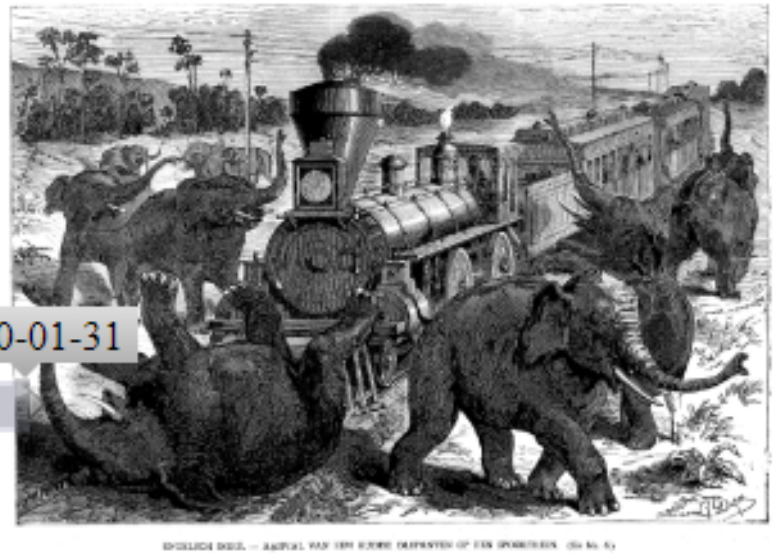
- Technical perspective: Managed to classify images
- Humanities perspective: We pinpointed moment in time photographs were predominantly being used
- Bonus: **experimental content analyses** of images

I. Intro - Technical Perspective

Steps:

3. Create pipeline:
    1. Binary classifier drawing - non drawing
    > Train: Thomas/Humanities reseacher created labelled set of images
    > Binary classifier **face – no face**
    No training needed, but existing tooling
    > Train: Thomas iteratively created **extra labelled set of images**: cartoon, crowd, chess, face, logo, schematic, sheetmusic, weather, other

CHRONReader: Query Classified Historical Newspaper Images



Select date-range:

1900-03-19 1920-01-31

Select image type:

(•) Photo ( ) Drawing ( ) Photo or drawing

Select image category:

Building ▾

- Any
- ✓ Building
- Cartoon
- Chess
- Crowds
- Face
- Logo
- Schematics
- Sheetmusic
- Weather
- Other

Image with face:

( ) No ( ) Yes (•) Yes or no

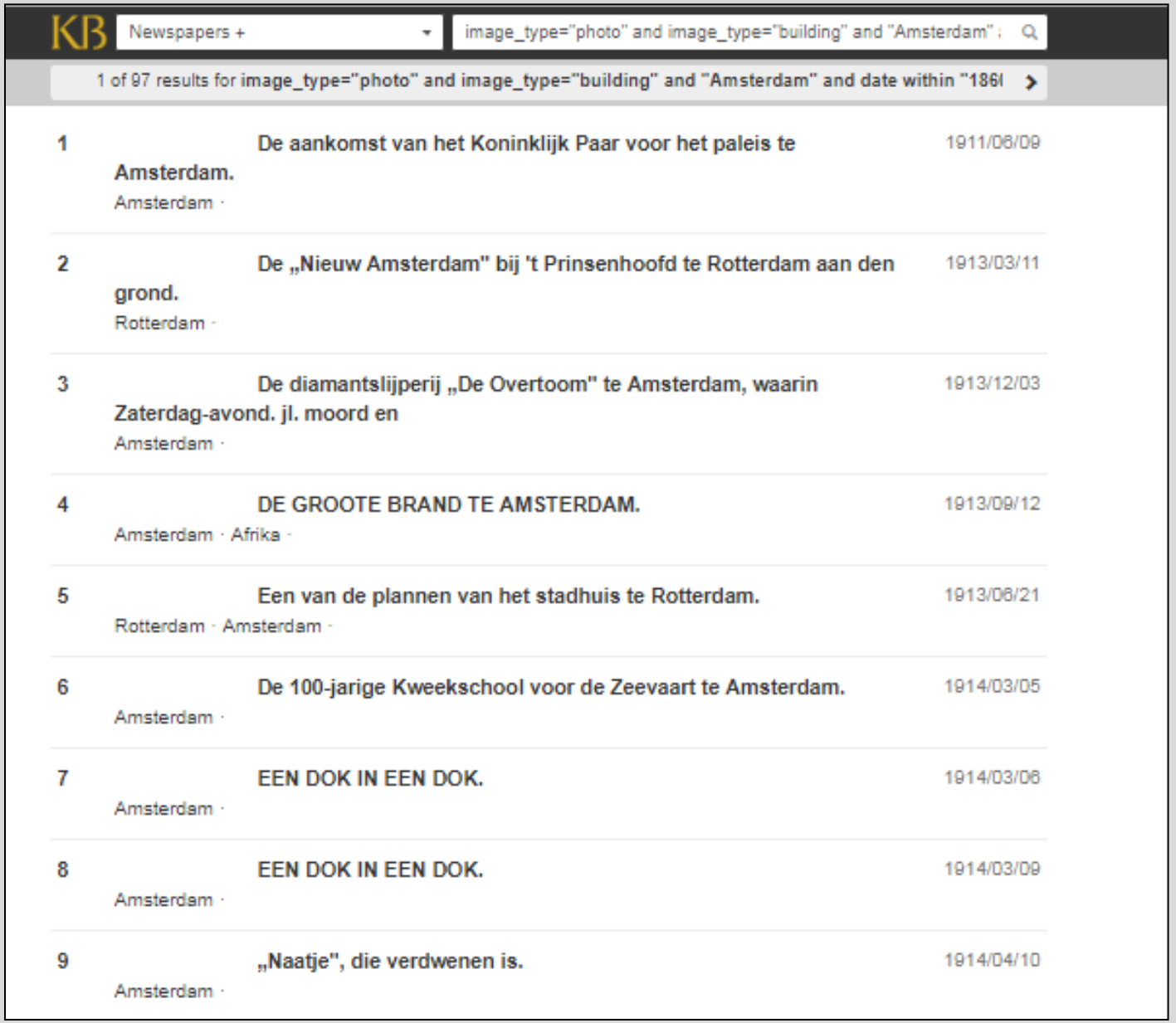
Text in article:

amsterdam

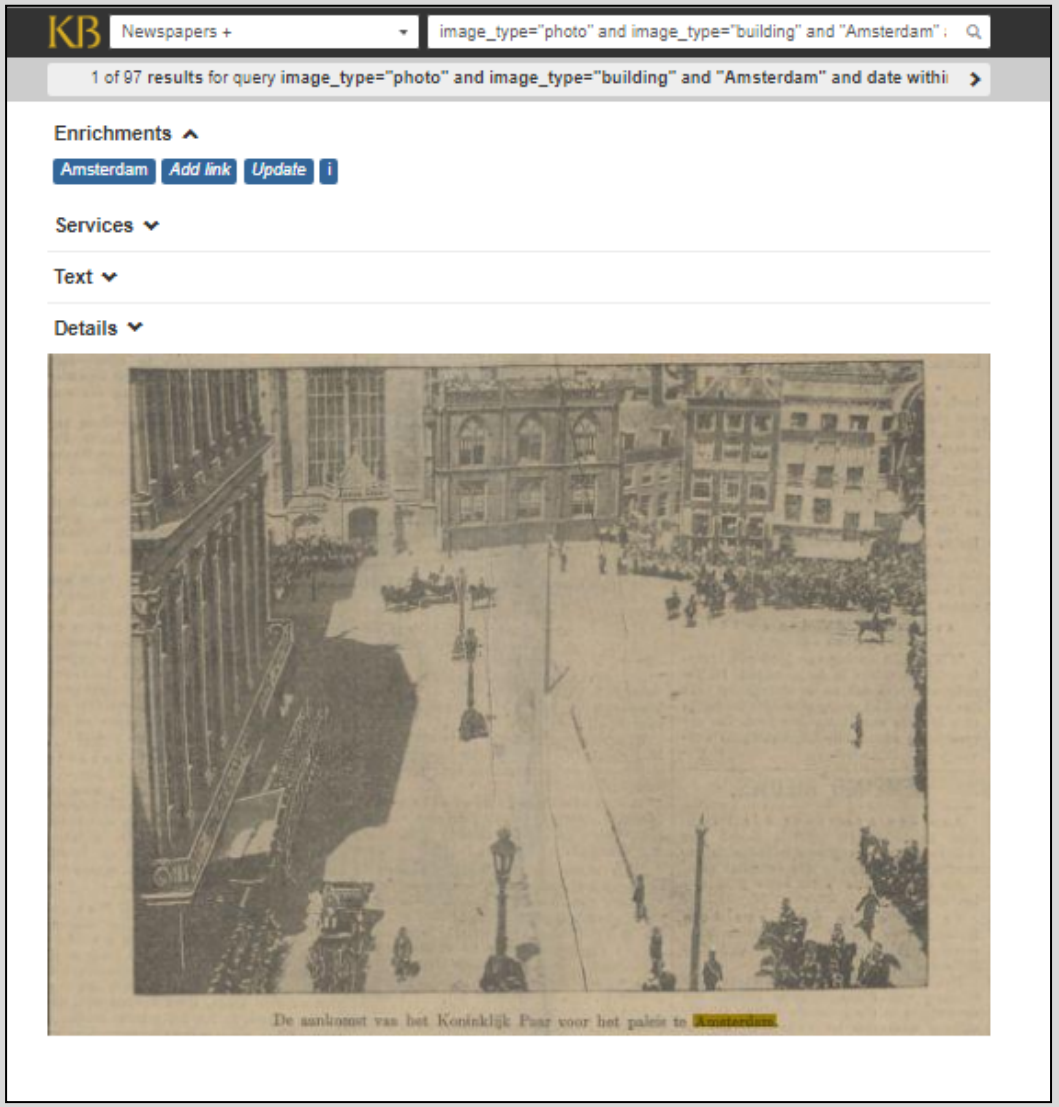
http://www.kbresearch.nl/xportal?image_type="phot amsterdam" and date within "1900-03-19T00:00:

go

http://lab.kb.nl/tool/chronreader



http://lab.kb.nl/tool/chronreader

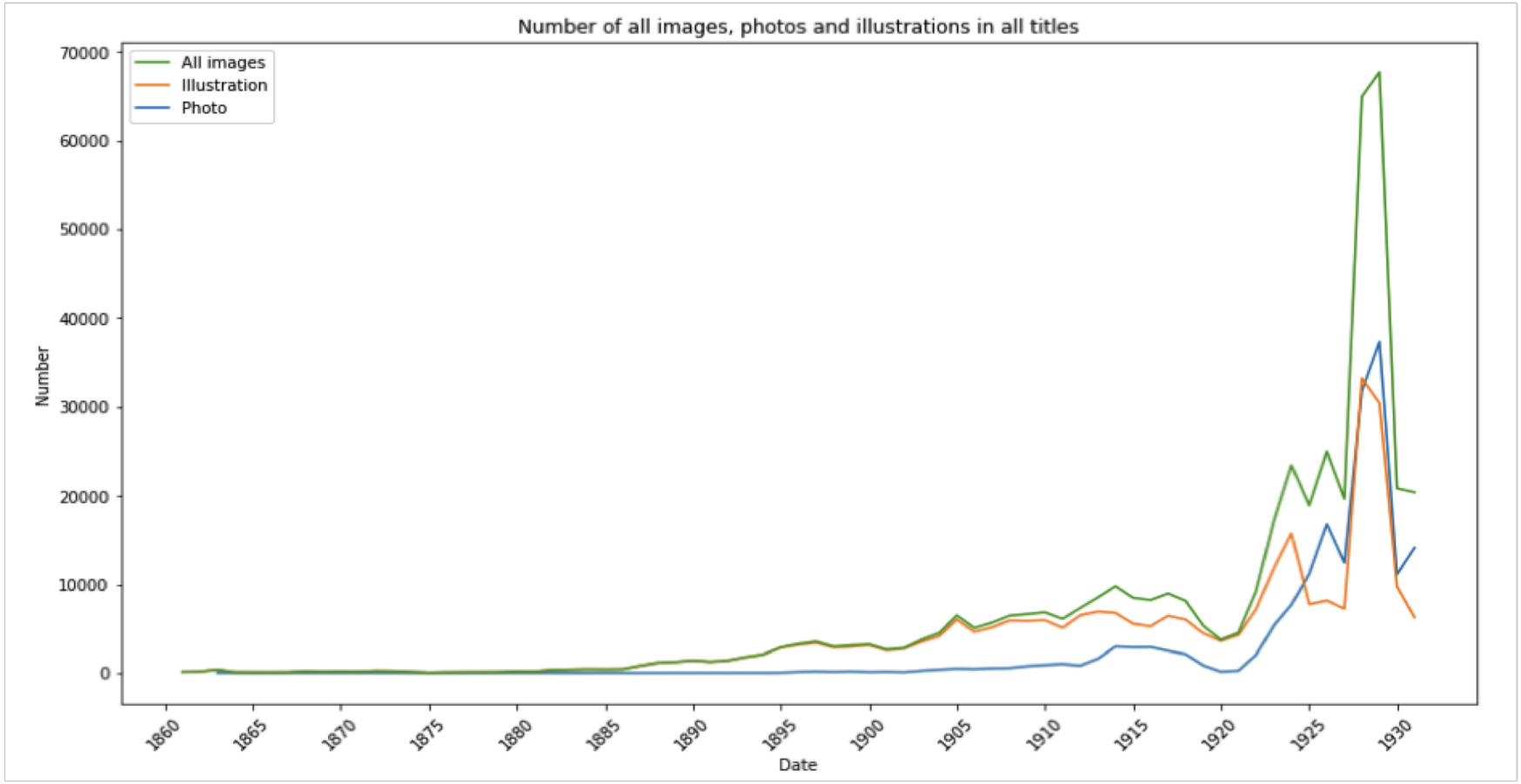


“De aankomst van het Koninklijk Paar voor het paleis in Amsterdam”
“Arrival of the Royal Couple at the palace in Amsterdam”

http://lab.kb.nl/tool/chronreader

II. Results

II. Results


Number of all images, photos and illustrations in all titles
All images
Illustration
Photo
Number
70000
60000
50000
40000
30000
20000
10000
0
1860
1865
1870
1875
1880
1885
1890
1895
1900
1905
1910
1915
1920
1925
1930
Date

II. Results

https://academic.oup.com/dsh/advance-article/doi/10.1093/llc/fqy085/5296356#.XHklANusYXc.twitter

DSH Digital Scholarship in the Humanities
eadh European Association for Digital Humanities
ADHO Alliance of Digital Humanities Organizations

Article Navigation

The visual digital turn: Using neural networks to study historical images

Melvin Wevers, Thomas Smits

Digital Scholarship in the Humanities, fqy085, https://doi.org/10.1093/llc/fqy085
Published: 18 January 2019

Abstract

Digital humanities research has focused primarily on the analysis of texts. This emphasis stems from the availability of technology to study digitized text. Optical character recognition allows researchers to use keywords to search and analyze digitized texts. However, archives of digitized sources also contain large numbers of images. This article shows how convolutional neural networks (CNNs) can be used to categorize and analyze digitized historical visual sources. We present three different approaches to using CNNs for gaining a deeper understanding of visual trends in an archive of digitized Dutch newspapers. These include detecting medium-specific features (separating photographs from illustrations), querying images based on abstract visual aspects (clustering visually similar advertisements), and training a neural network based on visual categories developed by domain experts. We argue that CNNs allow researchers to explore the visual side of the digital turn. They allow archivists and researchers to classify and spot trends in large collections of digitized visual sources in radically new ways.

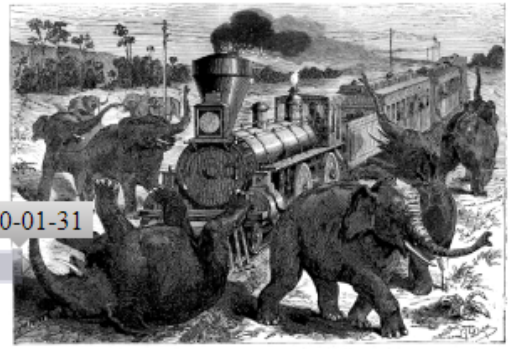
II. Results

Best Paper Award voor (voormalige) researchers in residence

https://www.kb.nl/nieuws/2018/onderzoek-met-kb-collectie-in-de-prijzen

II. Results

CHRONReader: Query Classified Historical Newspaper Images



Select date-range:

1900-03-19 1920-01-31

Select image type:

Photo Drawing Photo or drawing

Select image category:

Building

Image with face:

No Yes Yes or no

Text in article:

amsterdam

http://www.kbresearch.nl/xportal?image_type="photo" and image_type="building" and "amsterdam" and date within "1900-03-19T00:00:

go

II. Results

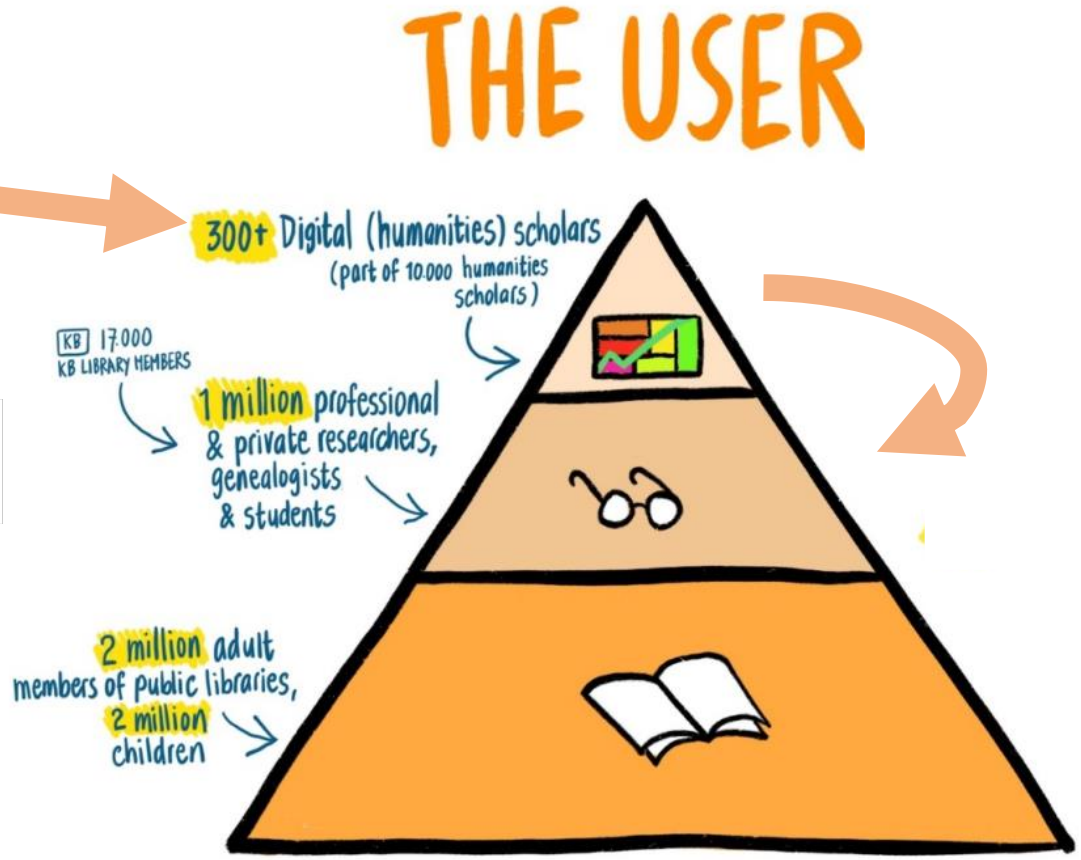
What worked?

- **Work in office**: access to internal expertise
- **Bi weekly meetings**: Researcher + Research Software Engineer + Digital Scholarship Advisor
- **Domain expertise researcher** in creating a training set

III. Discussion: One more result

- KB Perspective: possible **new services for larger group of customers**

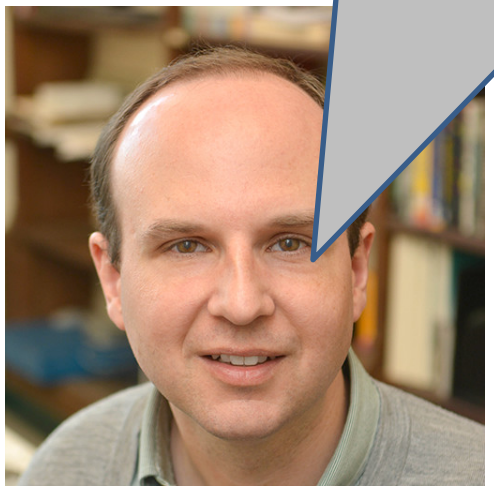
Delpher

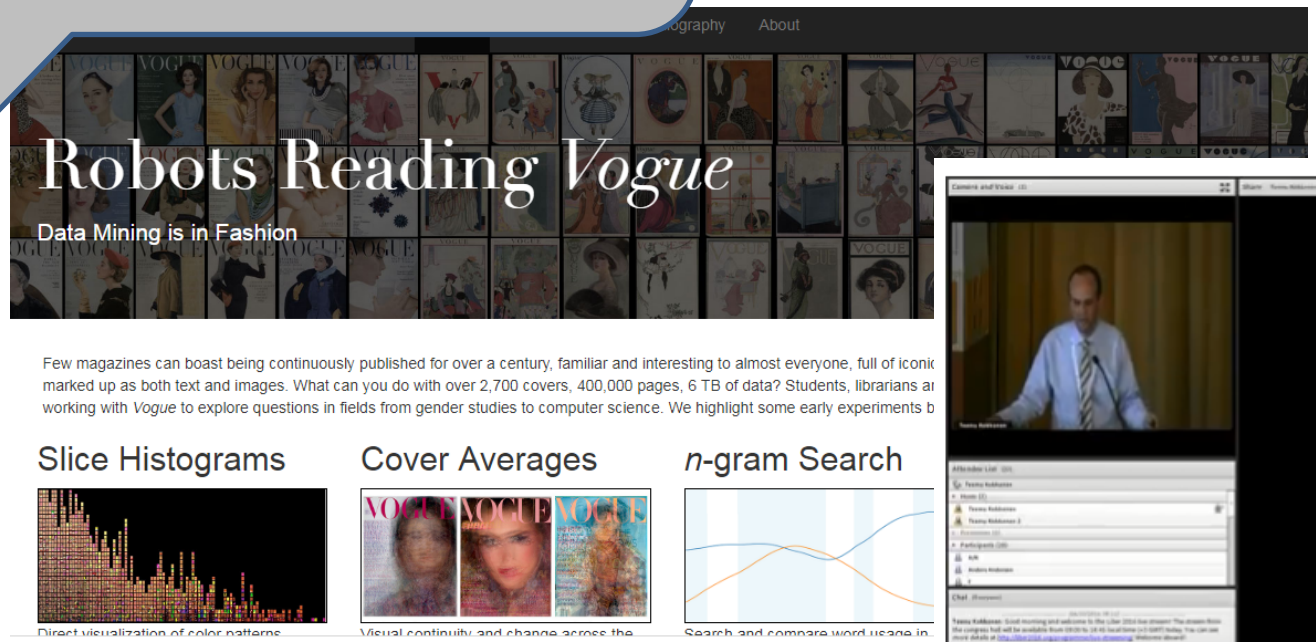

THE USER
300+ Digital (humanities) scholars
(part of 10.000 humanities scholars)
KB 17.000
KB LIBRARY MEMBERS
1 million professional & private researchers, genealogists & students
2 million adult members of public libraries,
2 million children

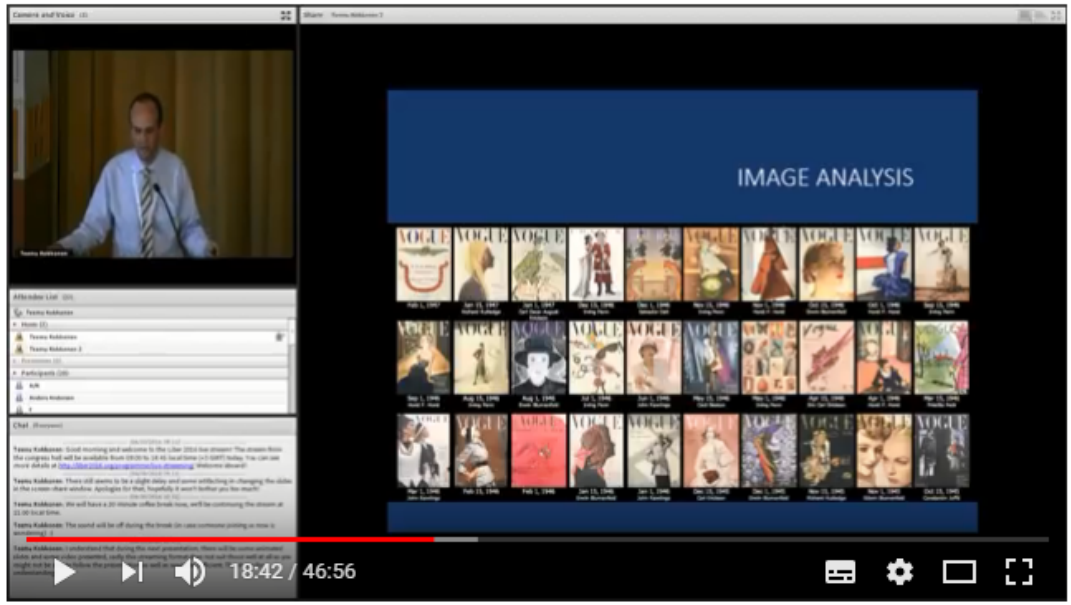
III. Discussion

- KB Perspective: possible new services for larger group of customers
- Peter Leonard & Alex Humpfreys: **Applied DH**

"Putting TDM in the Mainstream", i.e. search portals for bigger audience"







LIBER 2016 Keynote: Peter Leonard, Digital Humanities Lab, Yale University

http://dh.library.yale.edu/projects/vogue/
https://www.youtube.com/watch?v=yHi4TD4YfGQ
https://twitter.com/sclaeyssens/status/748047246722228228

JSTOR

Search

Login to My Account | Register

Advanced Search | Browse | About | Support

Text Analyzer BETA

BROUGHT TO YOU BY JSTOR LABS

Analyze Another Document | Help us make this better | About Text Analyzer

ANALYSIS

Prioritized terms

Adjust results by changing the weights for each term.

- Fairy tales
- Victorians
- Fantasy fiction
- History of science
- Philosophy of religion

Add your own term

Identified terms

Click to add to Prioritized Terms.

TOPICS

Agrology | Botany | Childrens literature | Creationism | Crime fiction | Fairy tales | Fantasy | Fantasy fiction | Higher criticism | History of science | Horror fiction | Information science | Irish folklore | Landscape paintings

RESULTS

Results with the prioritized terms: Fairy tales, Victorians, Fantasy fiction, History of science, Philosophy of religion

Search Filters: content I can access from 1900 - 2018

FREE ARTICLE

Nature's Invisibilia: The Victorian Microscope and the Miniature Fairy

Laura Forsberg

Victorian Studies, Vol. 57, No. 4 (Summer 2015), pp. 638-666

Download PDF | Save | Cite This Item

Prioritized Terms: Fairy tales, Victorians, Fantasy fiction, History of science, Philosophy of religion

Topics: Fairy tales, Victorians, Microscopes, Childrens literature, Imagination, History of science, Fantasy fiction, Philosophy of religion, Bonsai, Irish folklore,

ARTICLE

"But in a sense, what we do is: Applied Digital Humanities"

https://www.jstor.org/analyze/analyzer
https://www.slideshare.net/AlexHumphreys1/the-case-for-applied-digital-humanities-in-scholarly-communications
https://www.jstor.org/analyze/about

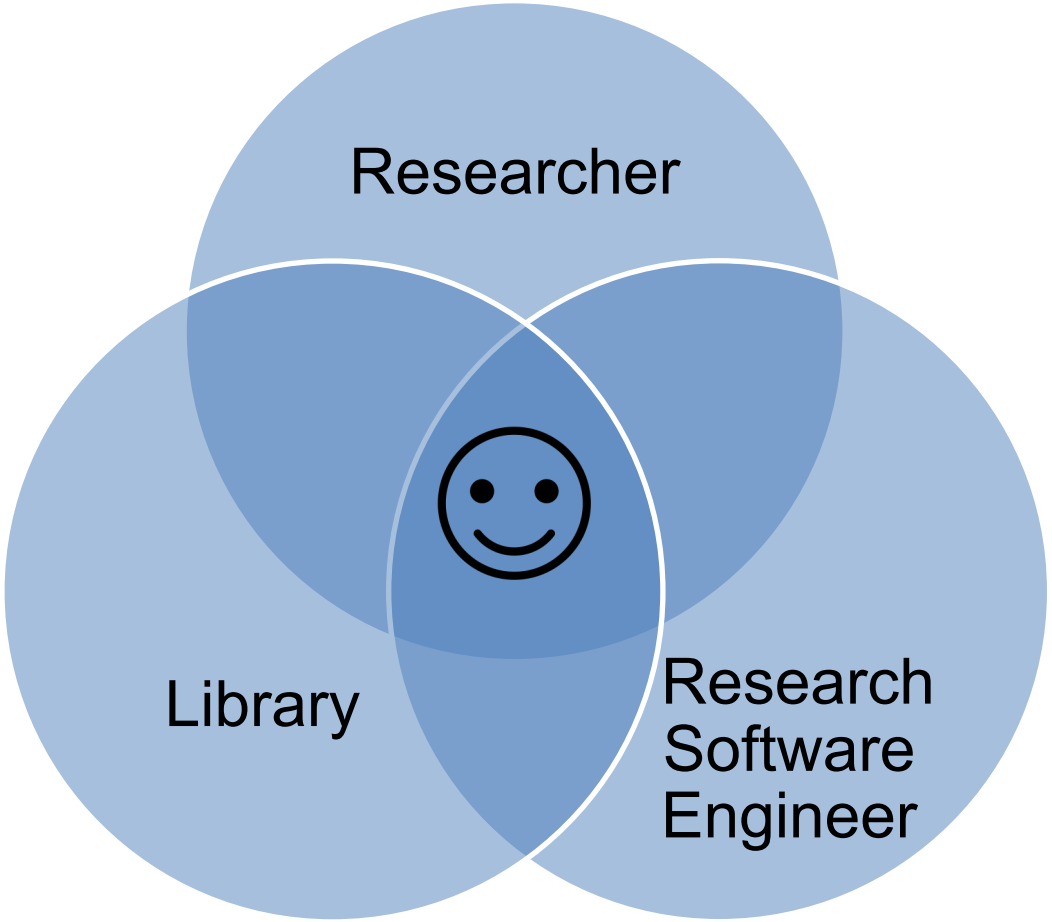
III. Discussion

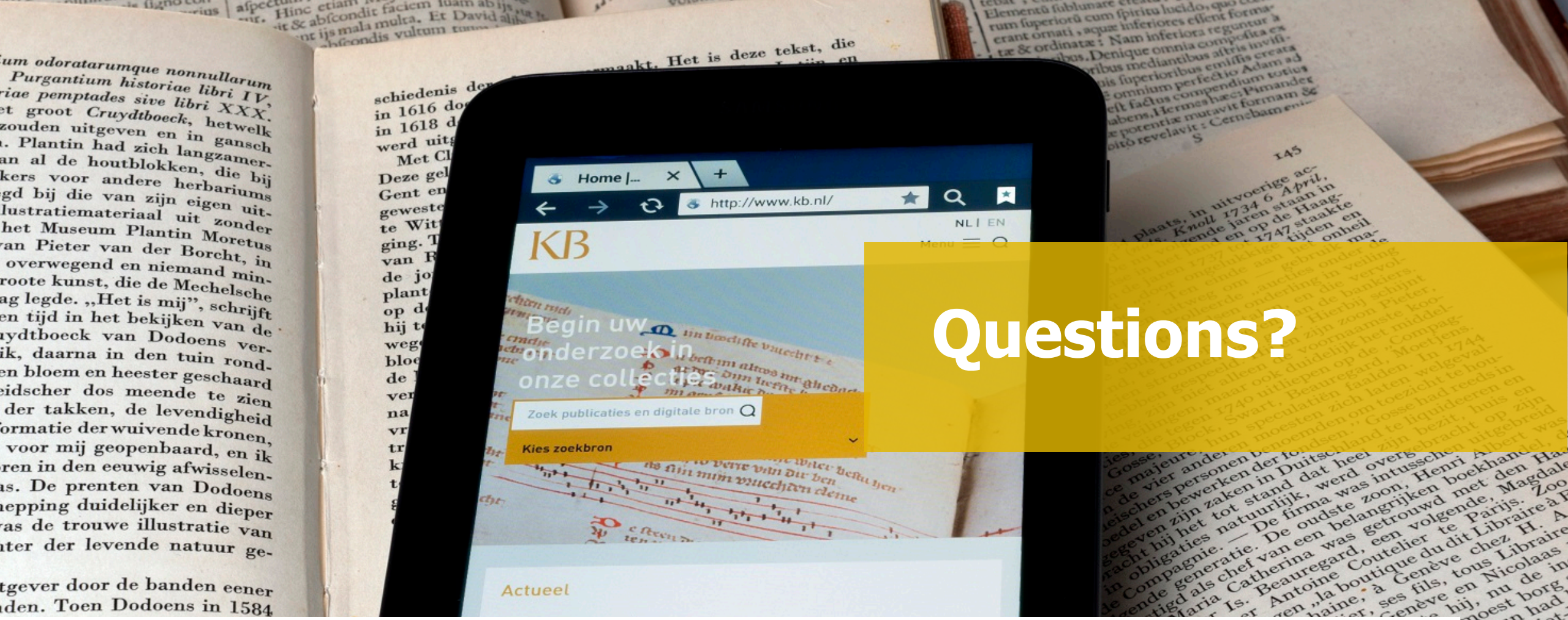
- KB Perspective: possible new services for larger group of customers
- Peter Leonard & Alex Humpfreys: Applied DH
- Libraries & Digital Collection Owners are **a third crucial partner** in DH projects

III. Discussion

Roles Collections holders:

- Data **provider**
- Full **Research partner**
- **Preserve results** by incorporating into services & products
- Perfect **Valorisation partner**





Questions?

Three perspectives on a collaborative attempt to use computer vision techniques to automatically classify historical newspaper images

dr. Martijn Kleppe – martijn.kleppe@kb.nl | @martijnkleppe | www.kb.nl/martijnkleppe
dr. Thomas Smits – t.p.smits@uu.nl | @thomassmits | visualnewsculture.tumblr.com/
Willem Jan Faber, MsC – willemjan.faber@kb.nl | lab.kb.nl/person/willem-jan-faber