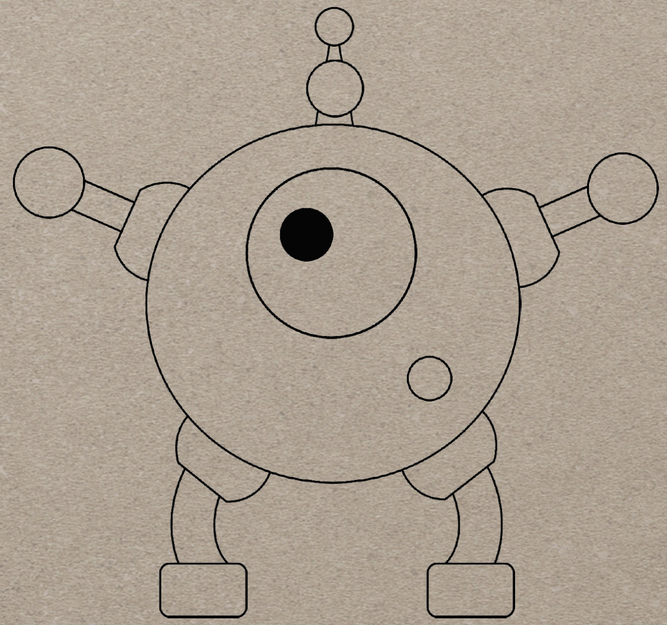
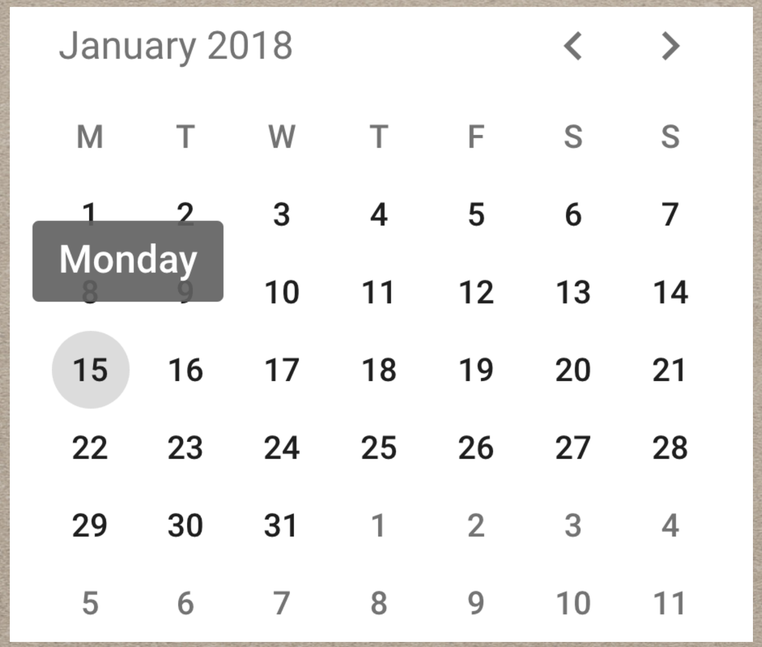
FIVE THINGS I LEARNED WHILE PROTOTYPING ML PAPERS

ELLEN KÖNIG / @ELLEN_KOENIG

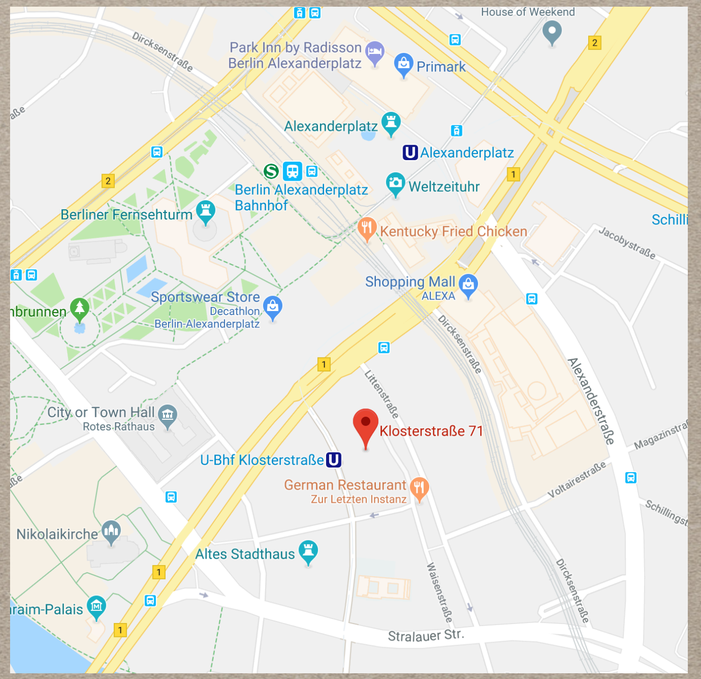
SENIOR DATA ENGINEER
THOUGHTWORKS

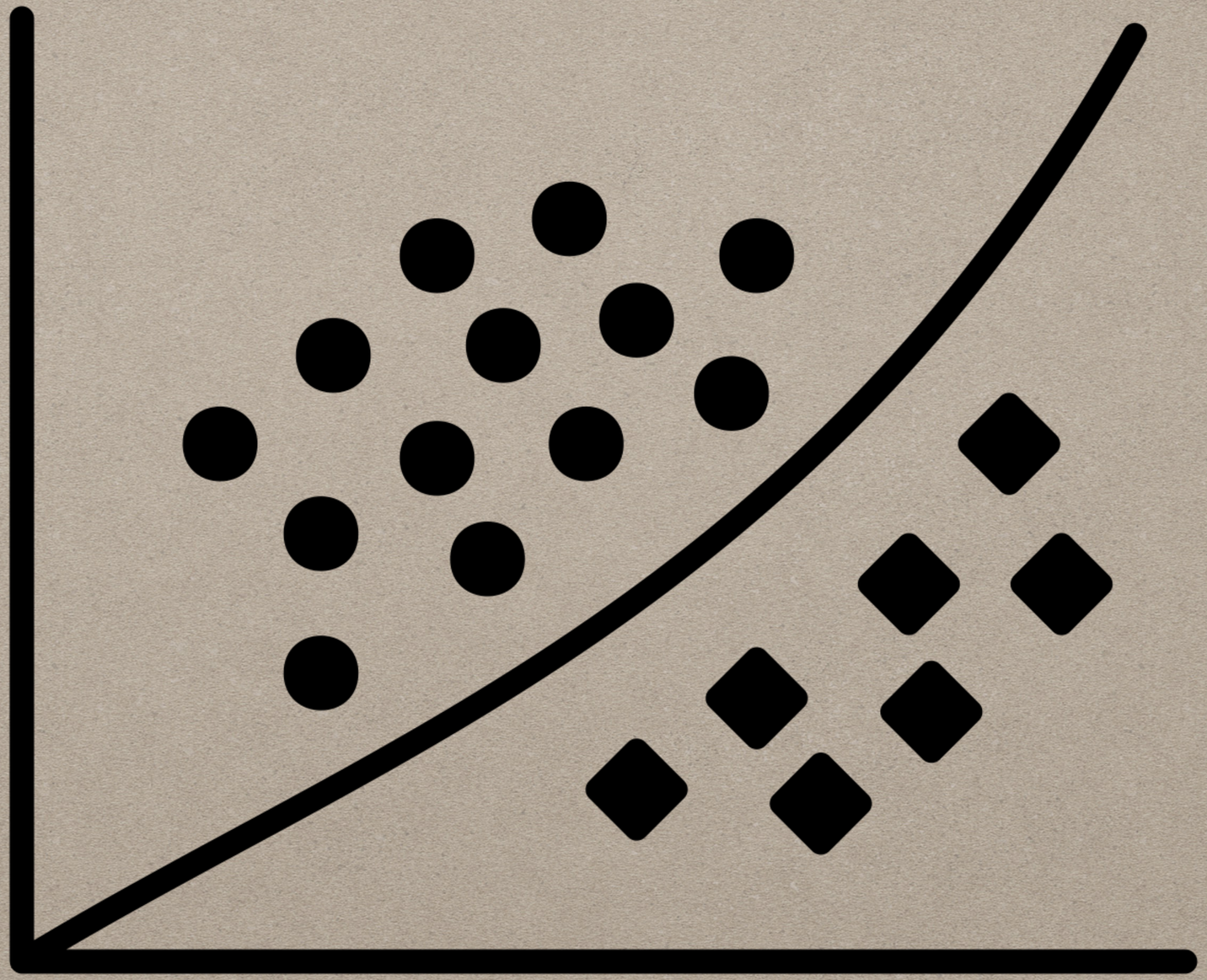


A LONG, LONG TIME AGO...



IN AN OFFICE FAR AWAY...





AND ONE DAY MY TEAM FACED A CHALLENGE

Page Stream Segmentation with Convolutional Neural Nets Combining Textual and Visual Features

Gregor Wiedemann, Gerhard Heyer
Department of Computer Science
Leipzig University, Germany
gregor.wiedemann@uni-leipzig.de, heyer@informatik.uni-leipzig.de

Abstract

For digitization of paper files via OCR, preservation of document contexts of single scanned images is a major requirement. Page stream segmentation (PSS) is the task to automatically separate a stream of scanned images into multi-page documents. This can be immensely helpful in the context of 'digital mailrooms' or retro-digitization of large paper archives. In a digitization project together with a German federal archive, we developed a novel PSS approach based on convolutional neural networks (CNN). Our approach combines image and text features to achieve optimal document separation results. Evaluation shows that our approach achieves accuracies up to 93 % which can be regarded as a new state-of-the-art for this task.

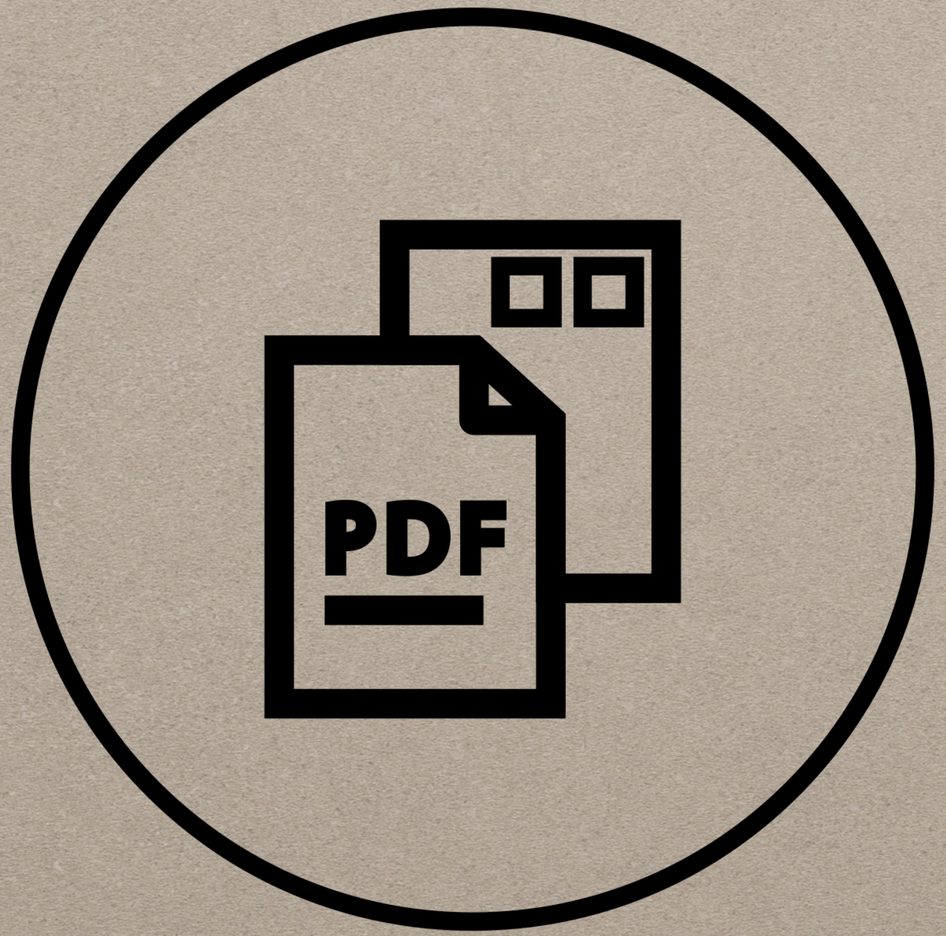
Keywords: page stream segmentation, convolutional neural nets, document image classification, document management, text classification

1. Introduction

For digitization of incoming mails in business contexts as well as for retro-digitizing archives, batch scanning of documents can be a major simplification of the processing workflow. In this scenario, scanned images of multi-page documents arrive at a document management system as an ordered stream of single pages lacking information on document boundaries. Page stream segmentation (PSS)

2. Related work

Page stream segmentation is related to a series of other tasks concerned with digital document management workflows. Table 1 gives an overview of recent related works. A common task is document image classification (DIC) in which typically visual features (pixels) are utilized to classify scanned document representations into categories such as "invoice", "letter", "certificate" etc. Category systems



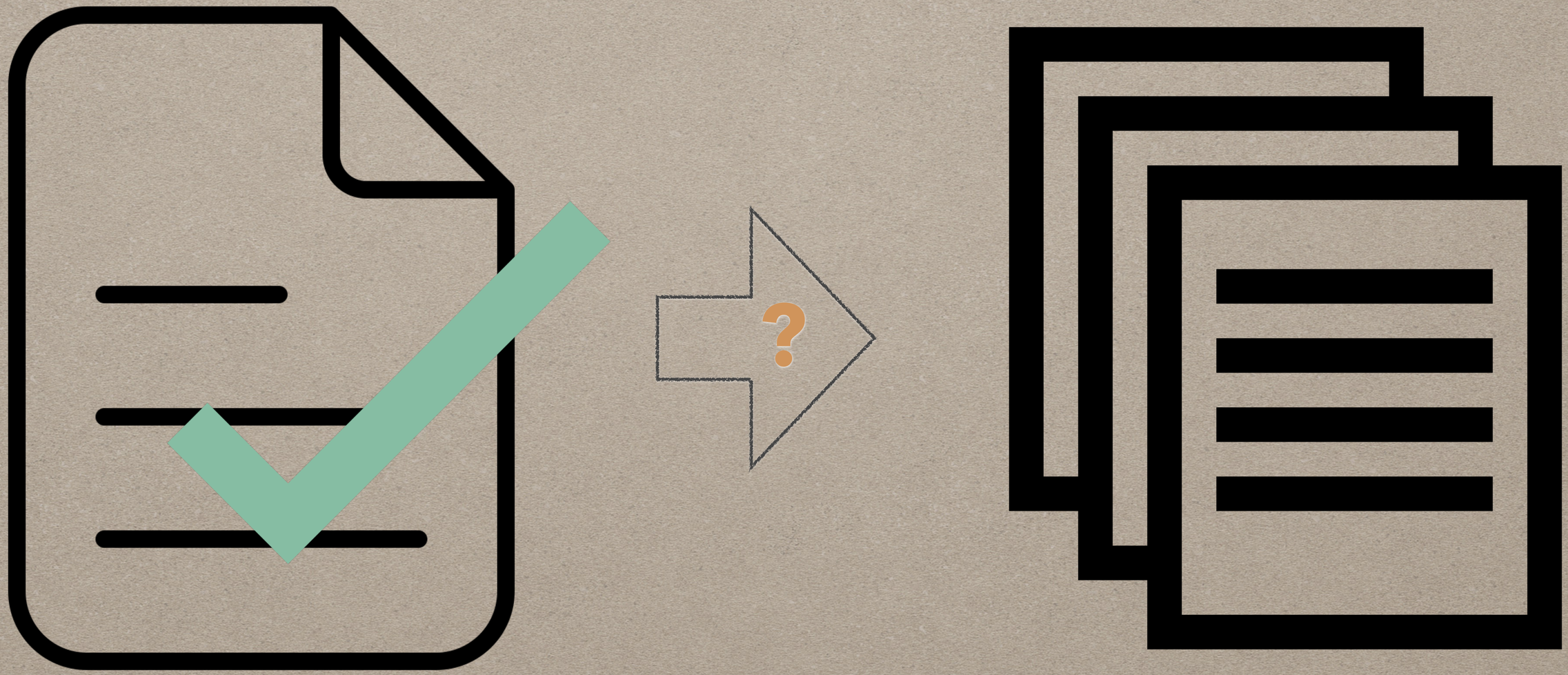
?

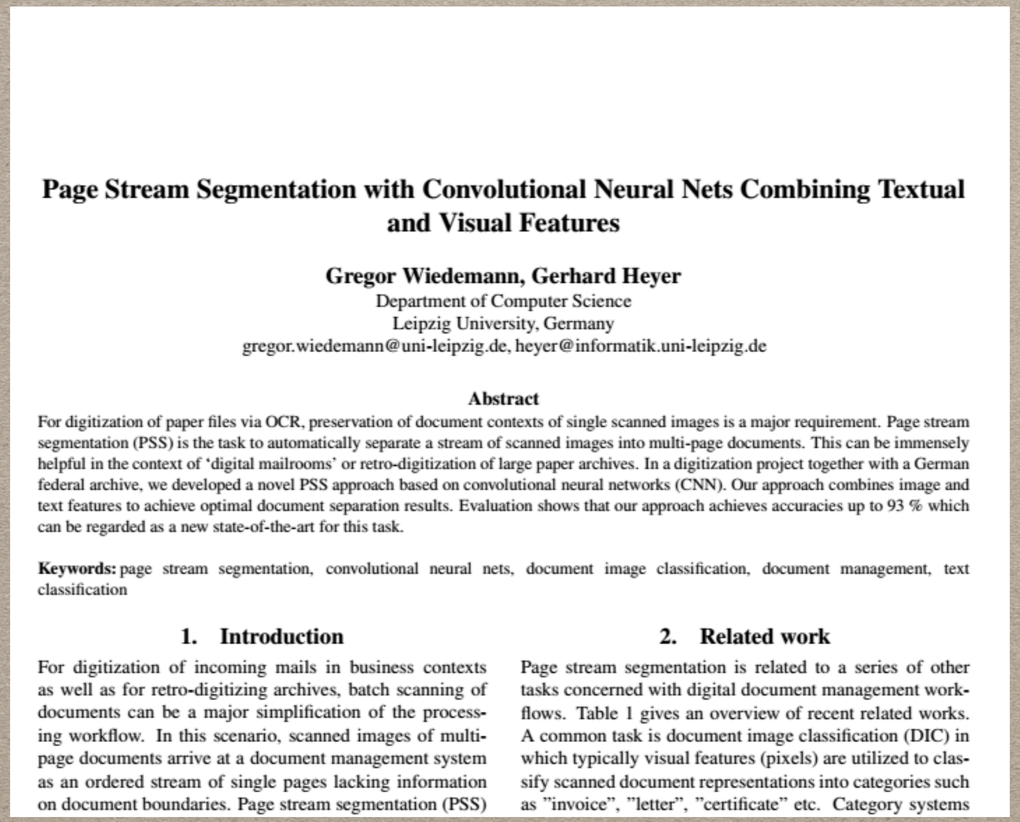
Bank statement

Identity document

Contract

...





Page Stream Segmentation with Convolutional Neural Nets Combining Textual and Visual Features

Gregor Wiedemann, Gerhard Heyer
Department of Computer Science
Leipzig University, Germany
gregor.wiedemann@uni-leipzig.de, heyer@informatik.uni-leipzig.de

Abstract

For digitization of paper files via OCR, preservation of document contexts of single scanned images is a major requirement. Page stream segmentation (PSS) is the task to automatically separate a stream of scanned images into multi-page documents. This can be immensely helpful in the context of 'digital mailrooms' or retro-digitization of large paper archives. In a digitization project together with a German federal archive, we developed a novel PSS approach based on convolutional neural networks (CNN). Our approach combines image and text features to achieve optimal document separation results. Evaluation shows that our approach achieves accuracies up to 93 % which can be regarded as a new state-of-the-art for this task.

Keywords: page stream segmentation, convolutional neural nets, document image classification, document management, text classification

1. Introduction

For digitization of incoming mails in business contexts as well as for retro-digitizing archives, batch scanning of documents can be a major simplification of the processing workflow. In this scenario, scanned images of multi-page documents arrive at a document management system as an ordered stream of single pages lacking information on document boundaries. Page stream segmentation (PSS)

2. Related work

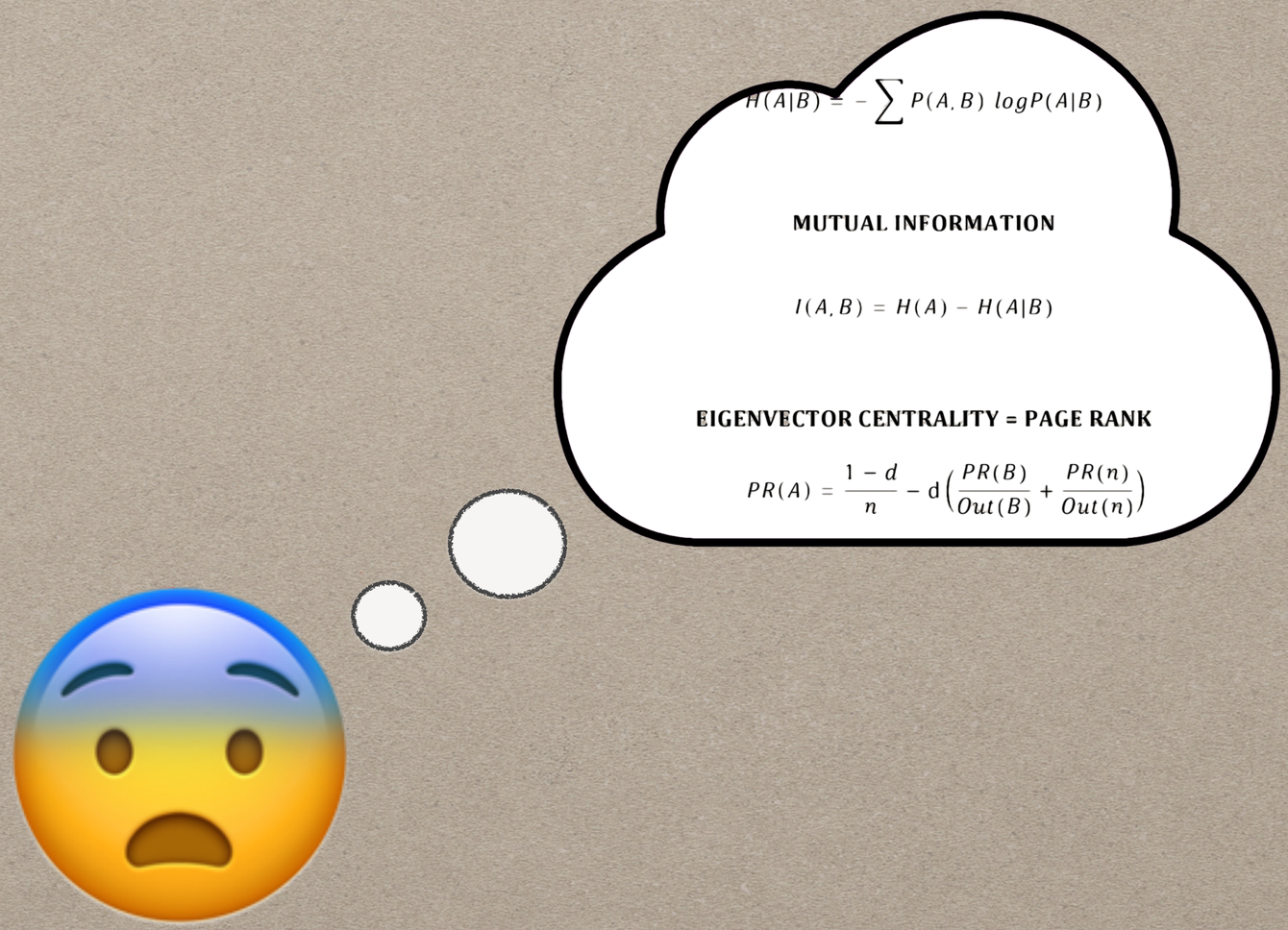
Page stream segmentation is related to a series of other tasks concerned with digital document management workflows. Table 1 gives an overview of recent related works. A common task is document image classification (DIC) in which typically visual features (pixels) are utilized to classify scanned document representations into categories such as "invoice", "letter", "certificate" etc. Category systems

WHY DID WE CONSIDER ML RESEARCH PAPERS?

- „Somebody must have solved this before!“

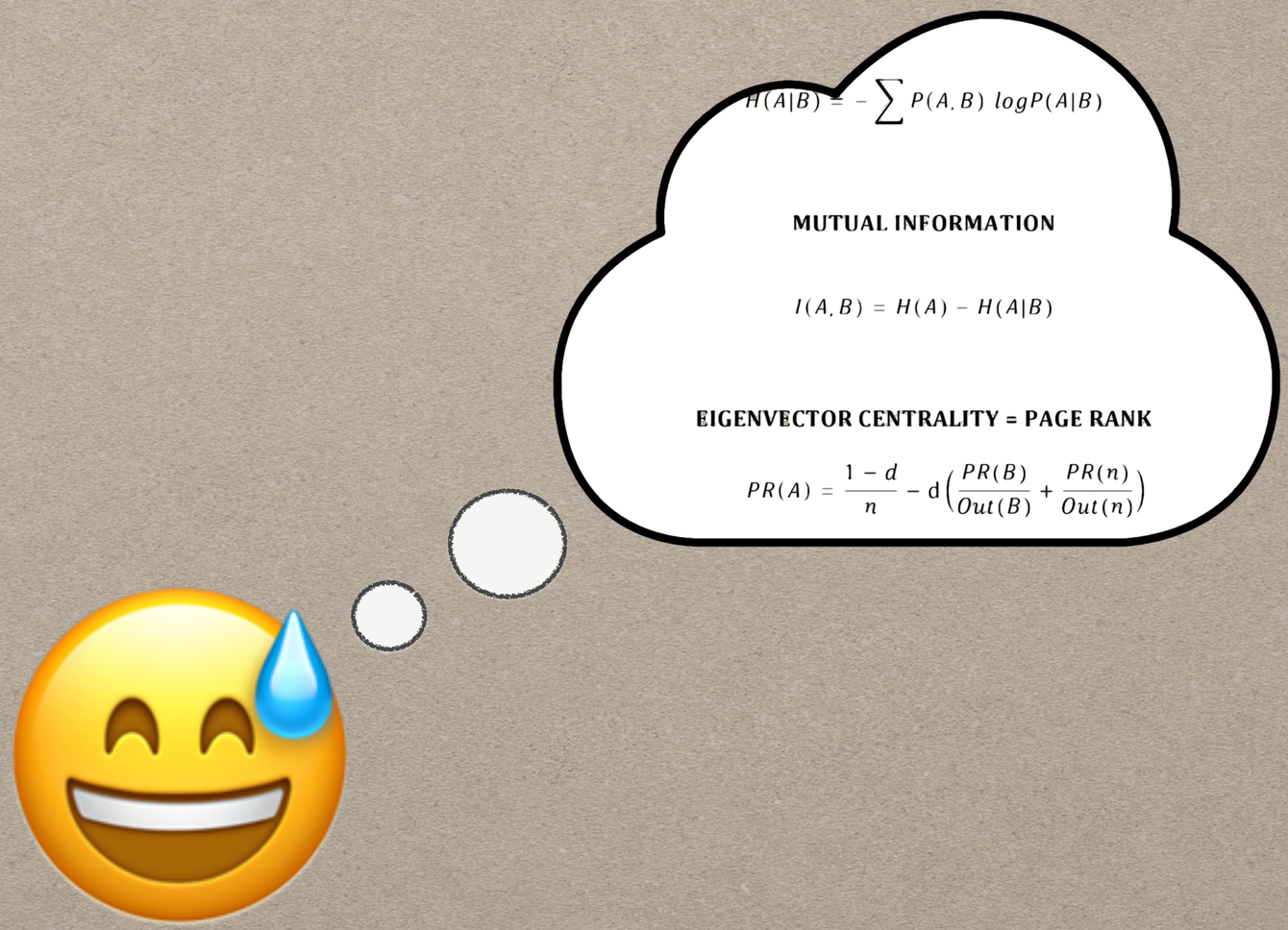
- No ready-to-use implementation

HOW MANY OF YOU CAN RELATE TO OUR PROBLEM?

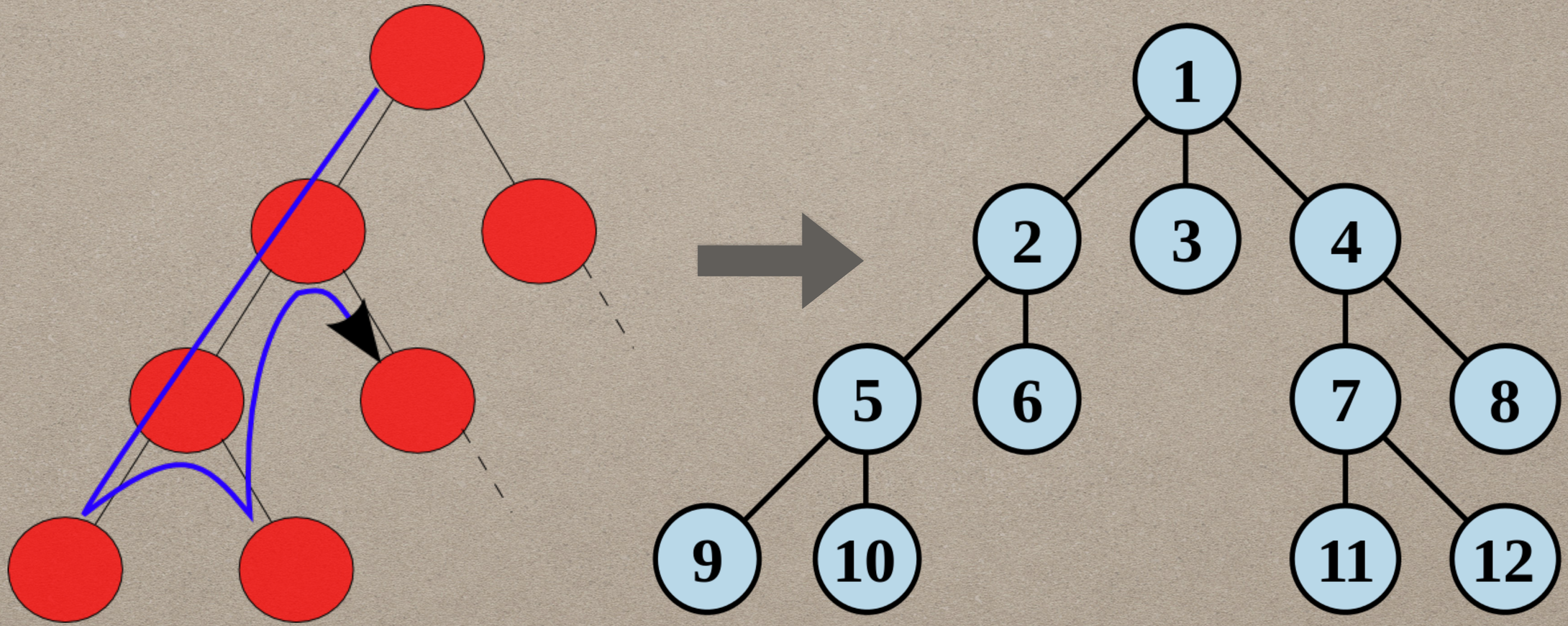

$H(A|B) = -\sum P(A,B)\ logP(A|B)$
MUTUAL INFORMATION
$I(A,B) = H(A) - H(A|B)$
EIGENVECTOR CENTRALITY = PAGE RANK
$PR(A) = \frac{1-d}{n} - d\left(\frac{PR(B)}{Out(B)} + \frac{PR(n)}{Out(n)}\right)$

BUT WORK IS ALL ABOUT GROWTH, RIGHT??

FORTUNATELY


$H(A|B) = -\sum P(A,B)\ logP(A|B)$
MUTUAL INFORMATION
$I(A,B) = H(A) - H(A|B)$
EIGENVECTOR CENTRALITY = PAGE RANK
$PR(A) = \frac{1-d}{n} - d\left(\frac{PR(B)}{Out(B)} + \frac{PR(n)}{Out(n)}\right)$

KEY INSIGHT:
BREADTH FIRST, NOT DEPTH FIRST

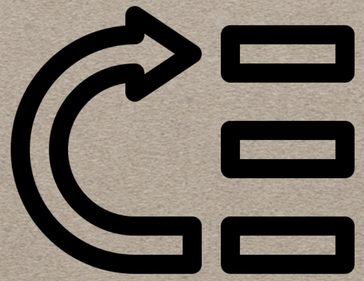


GOAL: FIND AND REPRODUCE THE BEST APPROACHES

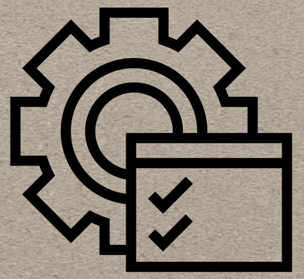
1. Search for research findings

3. Evaluate your papers

2. Decide on comparison criteria

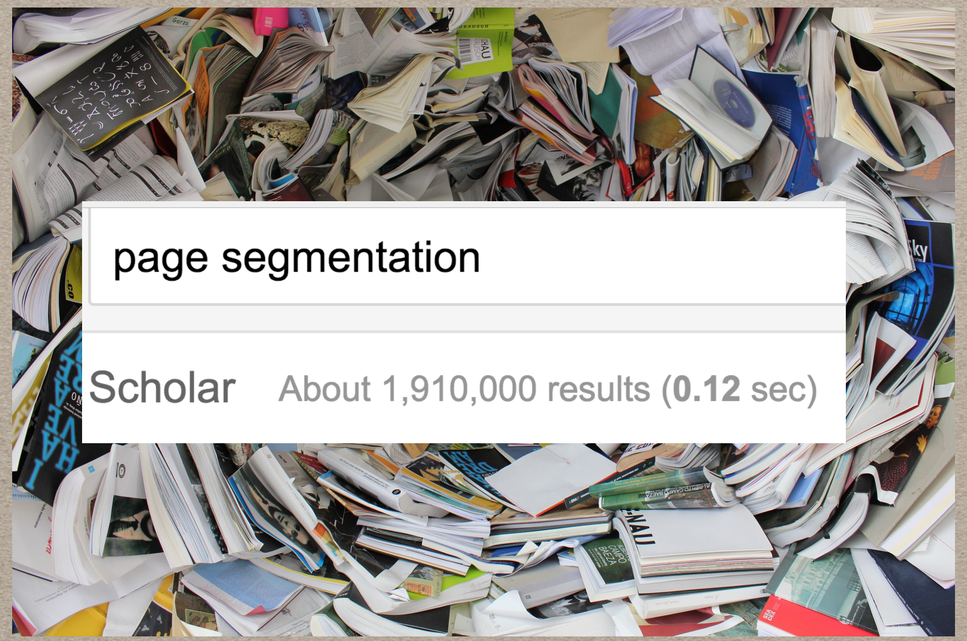


4. Prioritize approaches



5. Prototype approaches

STEP 1: SEARCH FOR RESEARCH FINDINGS

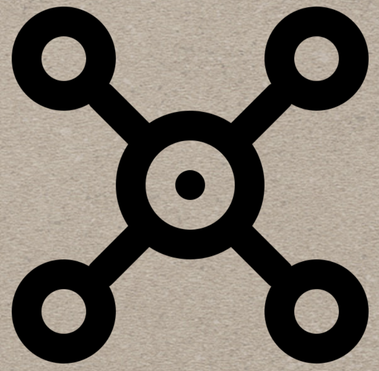


Needed:

An overview of the field

COMPILING AN OVERVIEW OF THE FIELD: BREADTH FIRST!

Start with survey papers, follow references



Common problems and approaches

Compile

Foundational and cutting edge papers

STEP 2: DECIDE ON YOUR COMPARISON CRITERIA

WHICH PAPERS ARE RIGHT FOR YOU?

Summarize common metrics and baselines

Pick simple metrics and baselines

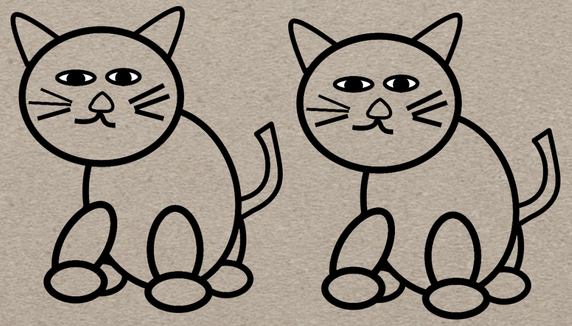
Minimally required metric targets?

Refresher on baselines: https://www.quora.com/What-does-baseline-mean-in-machine-learning

STEP 3: EVALUATE YOUR PAPERS

Groundbreaking?

Copycat?

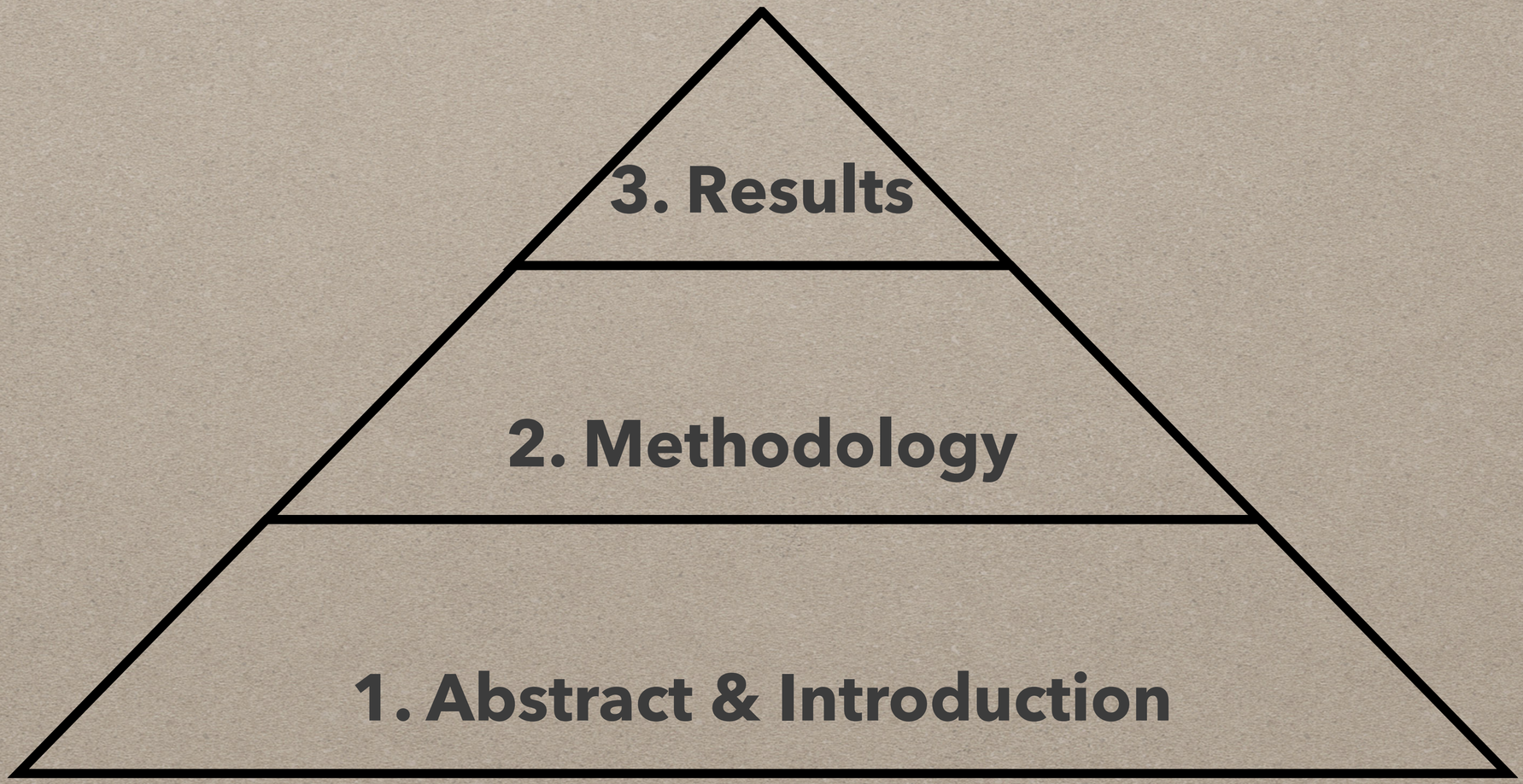


Garbage?

Journal / conference quality?

Team experience?

STEP 3: EVALUATE YOUR PAPERS – A CHECKLIST

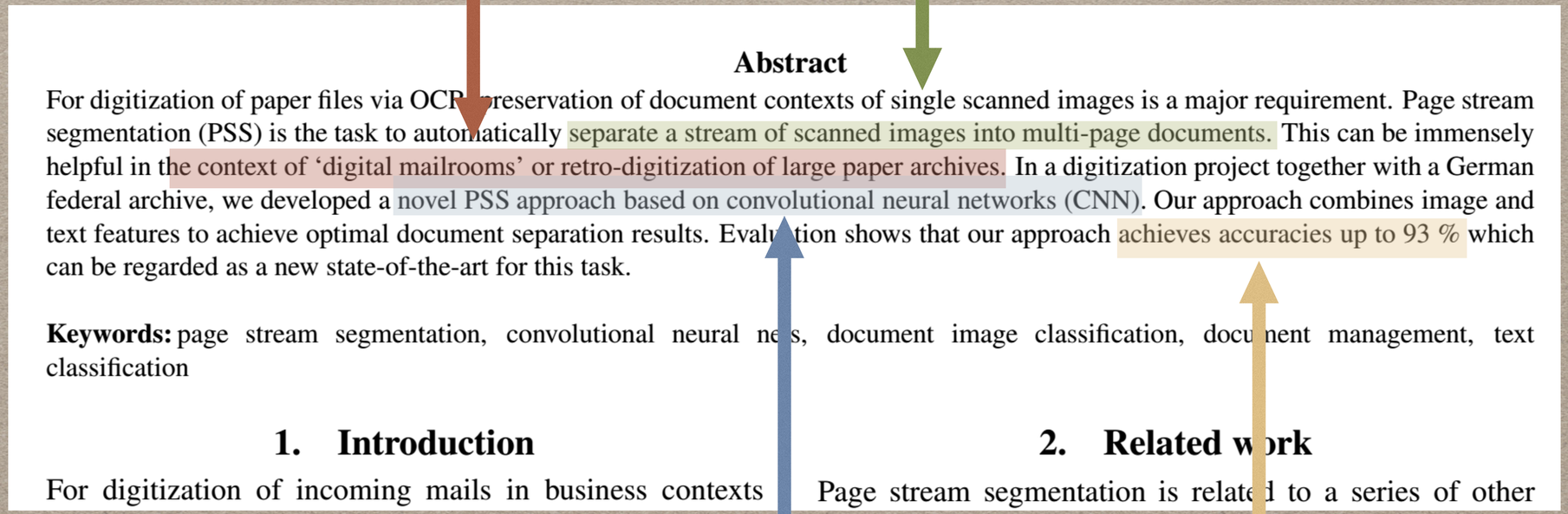


ABSTRACT & INTRODUCTION

Main question: Relevant to your problem?

Similar *context*?

Addresses your *problem*?



Abstract

For digitization of paper files via OCR, preservation of document contexts of single scanned images is a major requirement. Page stream segmentation (PSS) is the task to automatically separate a stream of scanned images into multi-page documents. This can be immensely helpful in the context of 'digital mailrooms' or retro-digitization of large paper archives. In a digitization project together with a German federal archive, we developed a novel PSS approach based on convolutional neural networks (CNN). Our approach combines image and text features to achieve optimal document separation results. Evaluation shows that our approach achieves accuracies up to 93 % which can be regarded as a new state-of-the-art for this task.

Keywords: page stream segmentation, convolutional neural networks, document image classification, document management, text classification

1. Introduction

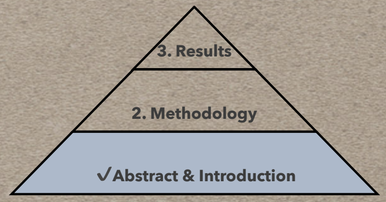
For digitization of incoming mails in business contexts

2. Related work

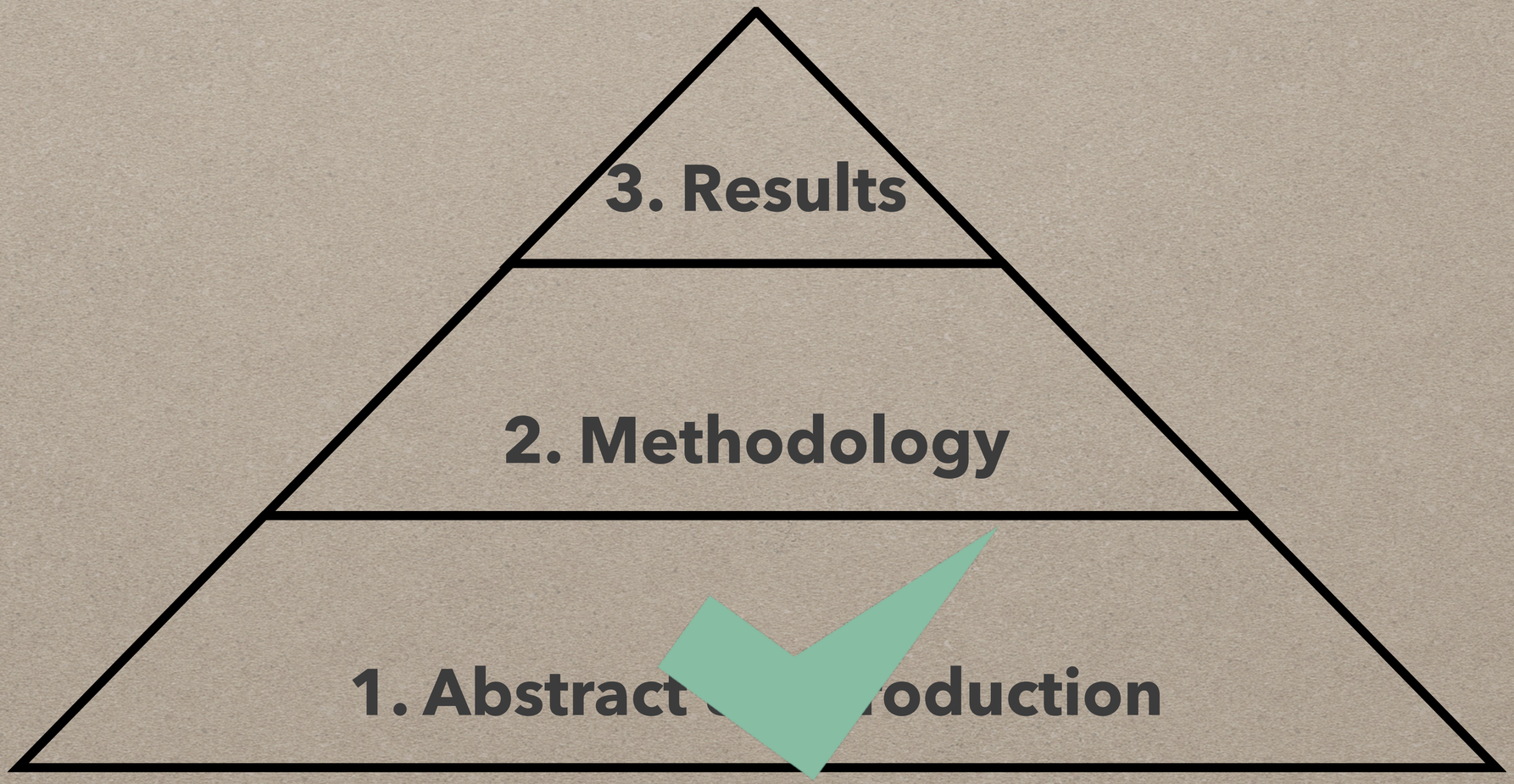
Page stream segmentation is related to a series of other

Approach: Groundbreaking or improvement?

Results: Better than targets & baseline?



STEP 3: EVALUATE YOUR PAPERS



METHODOLOGY SECTION

Main question: Approach reproducible?

1. **Solves similar problem?**

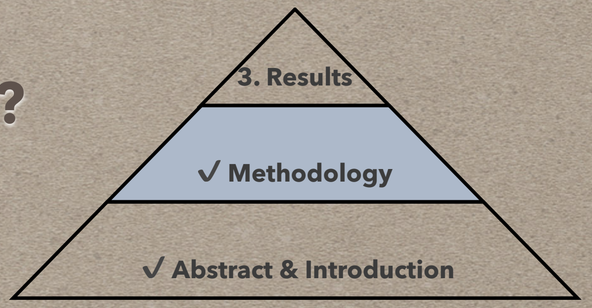
   Data set size and content similar?

2. **Description complete?**

   Entire process described?

   Pre-processing steps described completely?

   Well-known methods? Or completely described methods?



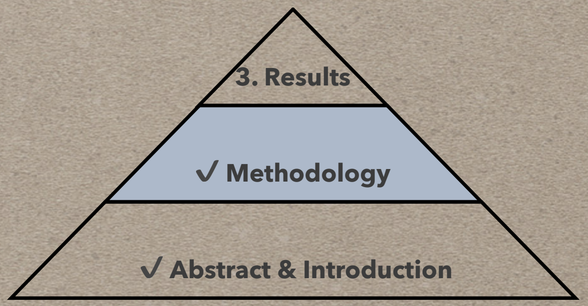
METHODOLOGY SECTION

Data set size and content similar?

✓22k black-and-white pages

✓German corpus

? Research documents rather than banking documents



METHODOLOGY SECTION

Entire process described?

✓Seems to be complete

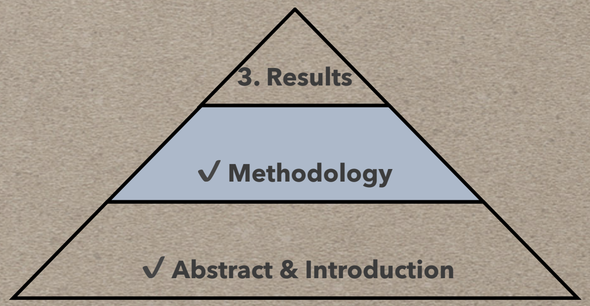
Pre-processing steps described completely?

✓Image conversion and scaling is described

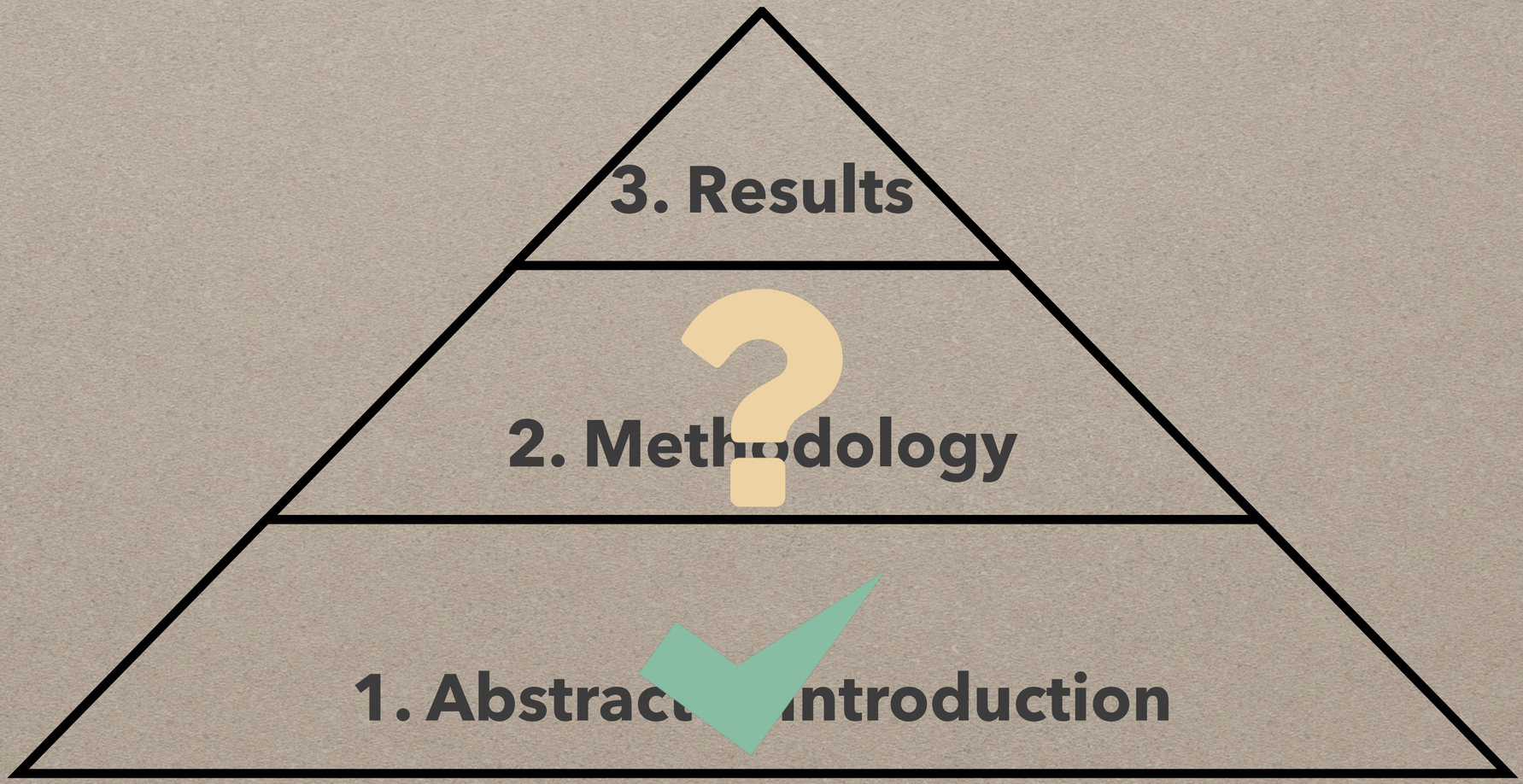
❓ OCR tool / approach is not mentioned

Well-known methods? Or completely described methods?

✓Neural network with descriptions of the configuration



STEP 3: EVALUATE YOUR PAPERS



RESULTS SECTION

Main question: Approach reproducible?

Evaluated with suitable metrics?

1.

Relevant metrics for your use case?

Metrics appropriate for the problem?

Metrics appropriate for the dataset?

Results good enough?

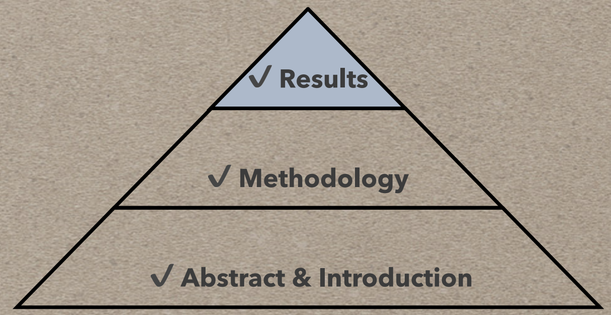
2.

Better than your baseline?

Better than the metrics target?

Any published review of the results?

Improvement analyzed with suitable statistical tests?



RESULTS SECTION

Main question: Results reliable?

Relevant metrics for your use case?

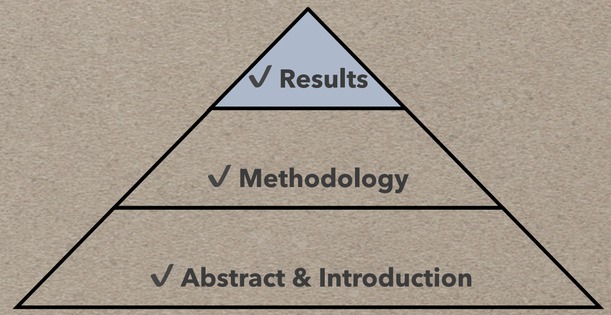
✓Accuracy

Metrics appropriate for the problem?

✓Common metric for classification

Metrics appropriate for the dataset?

✗ Not suitable for imbalanced classes



RESULTS SECTION

Better than your baseline?

✓Yes, by 0.23 over the baseline

Better than the metrics target?

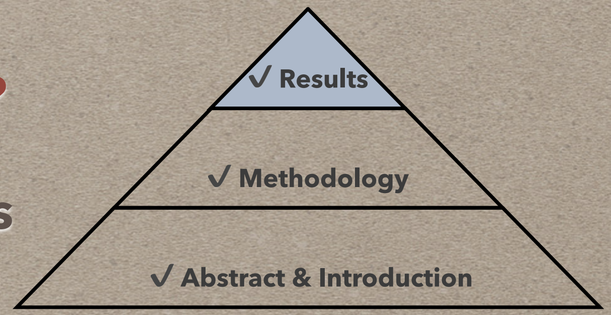
? They are close

Any published review of the results?

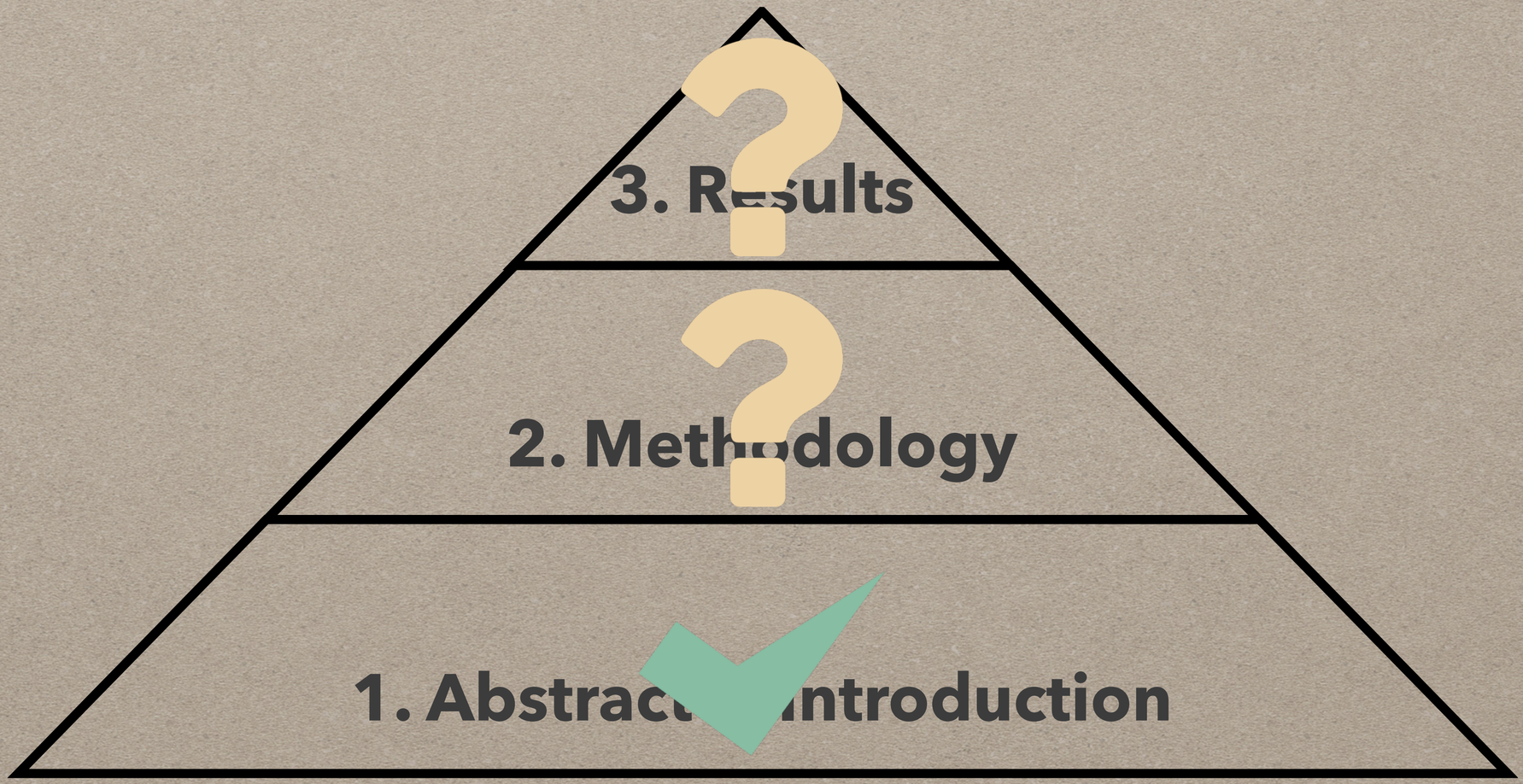
? Not yet

Improvement analyzed with suitable statistical tests?

X No statistical analysis, and reported measurements are not comparable

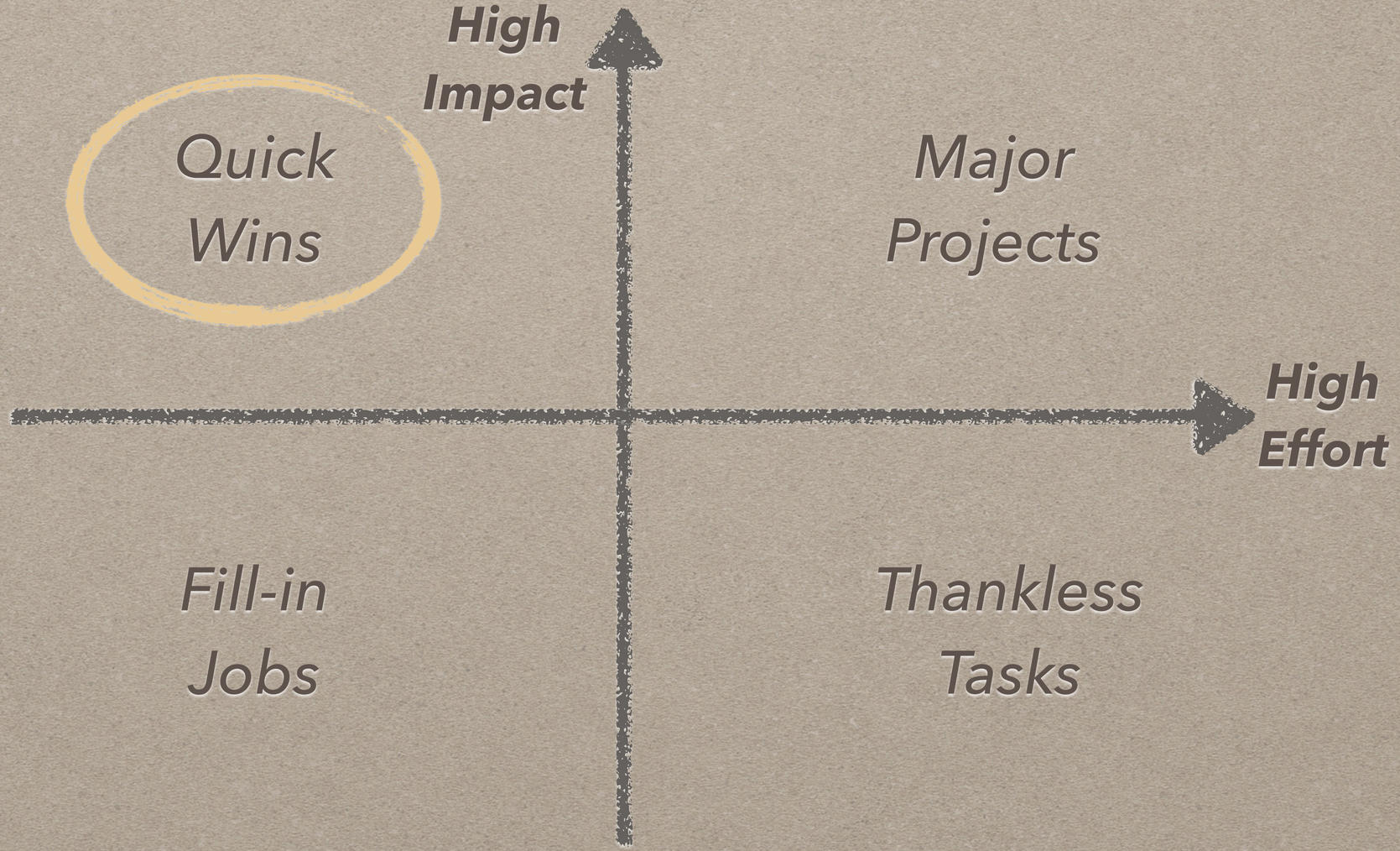


STEP 3: EVALUATE YOUR PAPERS



STEP 4: PRIORITIZE YOUR CHOSEN APPROACHES

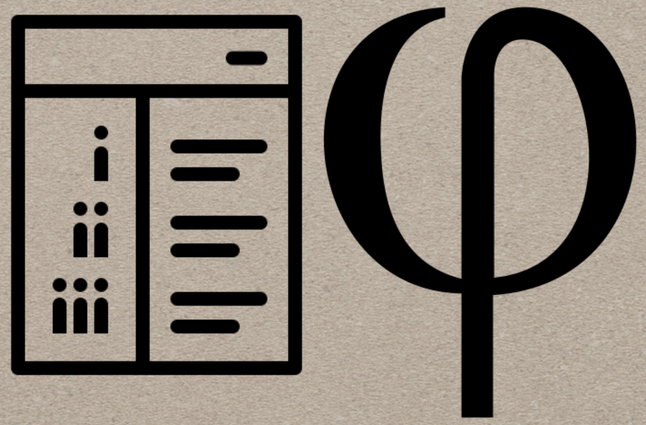
PRIORIZATION MATRIX



STEP 5: PROTOTYPE YOUR CHOSEN APPROACHES

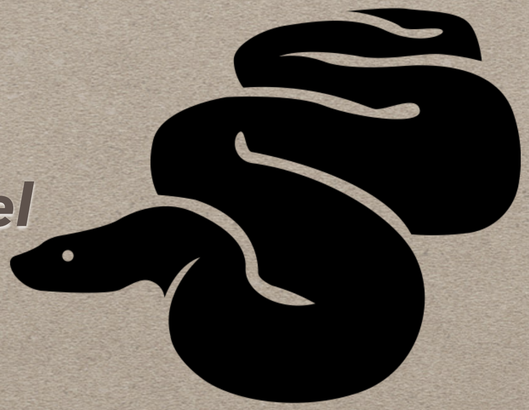
A FEW RECOMMENDATIONS

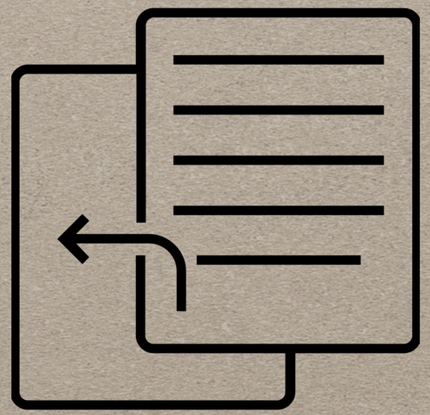
Compile a glossary



Understand all equations & code

Higher level language





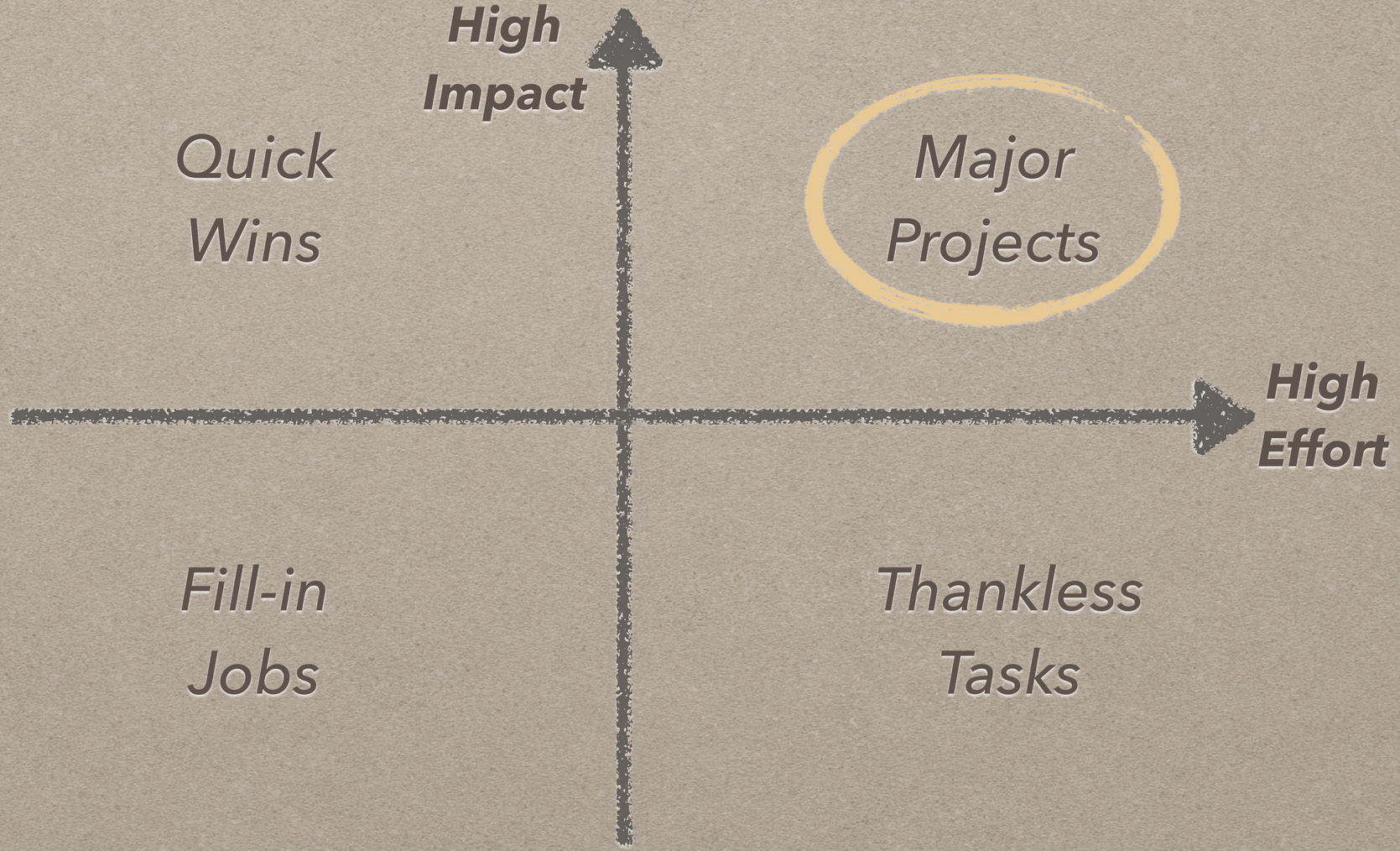
Reference sections of papers

MORE RECOMMENDATIONS

http://codecapsule.com/2012/01/18/how-to-implement-a-paper/

OUR FINAL RESULTS

PRIORIZATION MATRIX



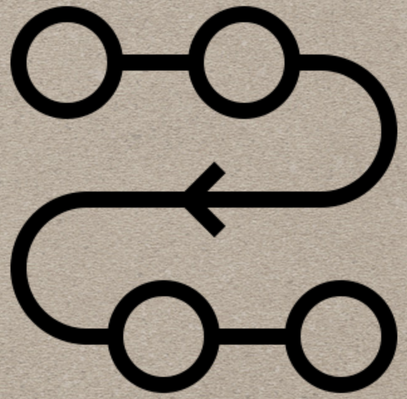
SUMMARY: WHEN SHOULD YOU LOOK FOR RESEARCH PAPERS?

- „Somebody must have solved this before!“

- No ready-to-use implementation

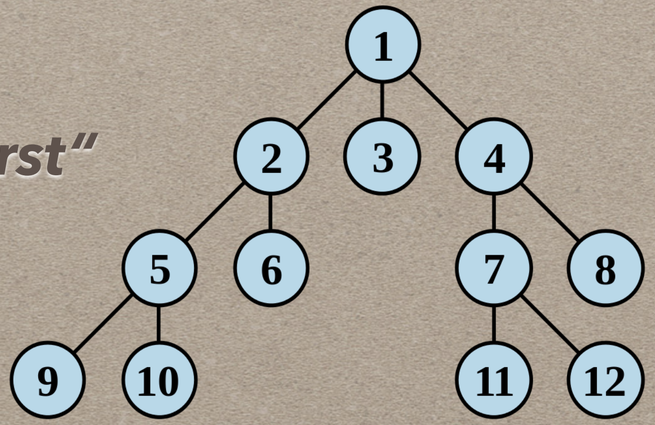
SUMMARY: OUR MAIN LESSONS

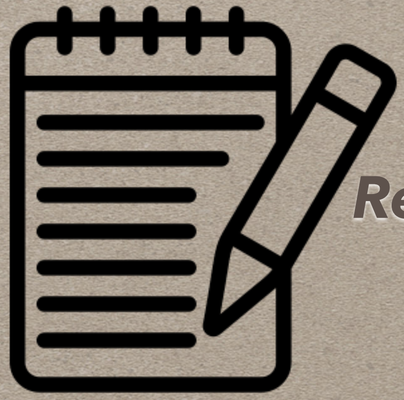
Pool your knowledge



Follow a strategy

Go „Breadth first“

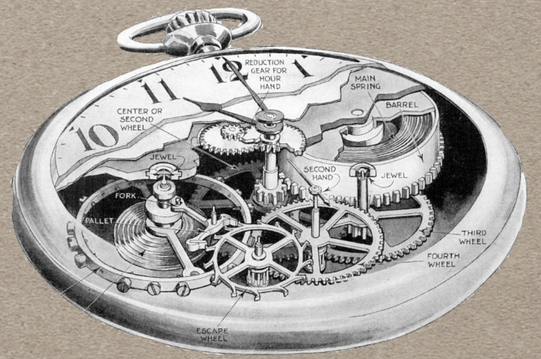




Record your insights

SUMMARY: A WORKFLOW FOR PROTOTYPING ML PAPERS

1. **Search** for research findings
2. **Decide** on your comparison criteria
3. **Evaluate** quality, relevance and reproducibility
4. **Prioritize** your chosen approaches
5. **Prototype** the best approaches



HAVE (MORE 🙃) FUN PROTOTYPING!

Slides will be tweeted from @ellen_koenig

IMAGE CREDITS

- Title slide: https://www.flickr.com/photos/vblibrary/6671465981
- Slide 2: Google calendar & maps
- Slide 10: https://www.datasciencecentral.com/profiles/blogs/140-machine-learning-formulas
- Slide 12: https://pixabay.com/de/bremer-stadtmusikanten-skulptur-2444326/
- Slide 14: https://commons.wikimedia.org/wiki/File:Breadth-first_tree.svg
- Slide 29: https://commons.wikimedia.org/wiki/File:Pocketwatch_cutaway_drawing.jpg

IMAGE CREDITS CONT.

- Slide 14 https://en.wikipedia.org/wiki/Map#/media/File:World_Map_1689.JPG
- Slide 26: https://pxhere.com/en/photo/109282
- Slide 27: Adapted from: http://www.sixsigmadaily.com/impact-effort-matrix/
- Slide 28: https://pixnio.com/objects/computer/programming-code-programmer-coding-coffee-cup-computer-copy-hands-computer-keyboard