DFG - Projekt

Optimierter Einsatz von OCR-Verfahren – Tesseract als Komponente im OCR-D-Workflow

Noah Metzger, Stefan Weil
Universitätsbibliothek Mannheim

19.09.2019

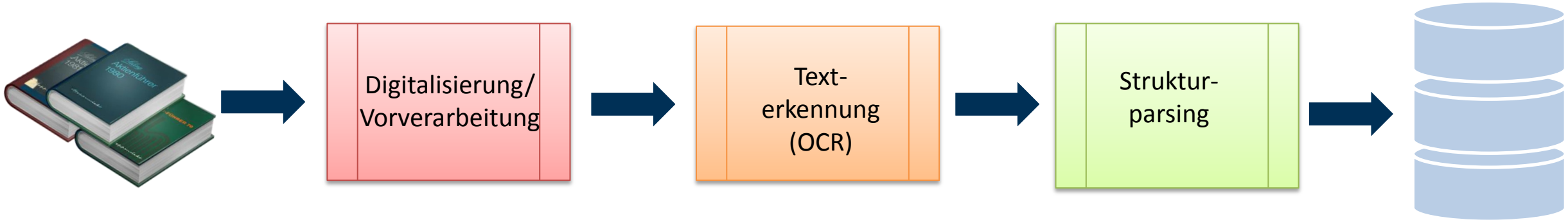
OCR-D
Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR)

Agenda

UB
UNIVERSITÄTSBIBLIOTHEK
MANNHEIM

- Was ist OCR?
- Tesseract
- Zielsetzung des OCR-D Projekts
- Stabilitätsverbesserung von Tesseract
- Performanceverbesserung von Tesseract
- Schnittstellen Entwicklung
- Positive Nebeneffekte
- Ausblick

Prozesskette
Forschungsdaten aus Digitalisaten



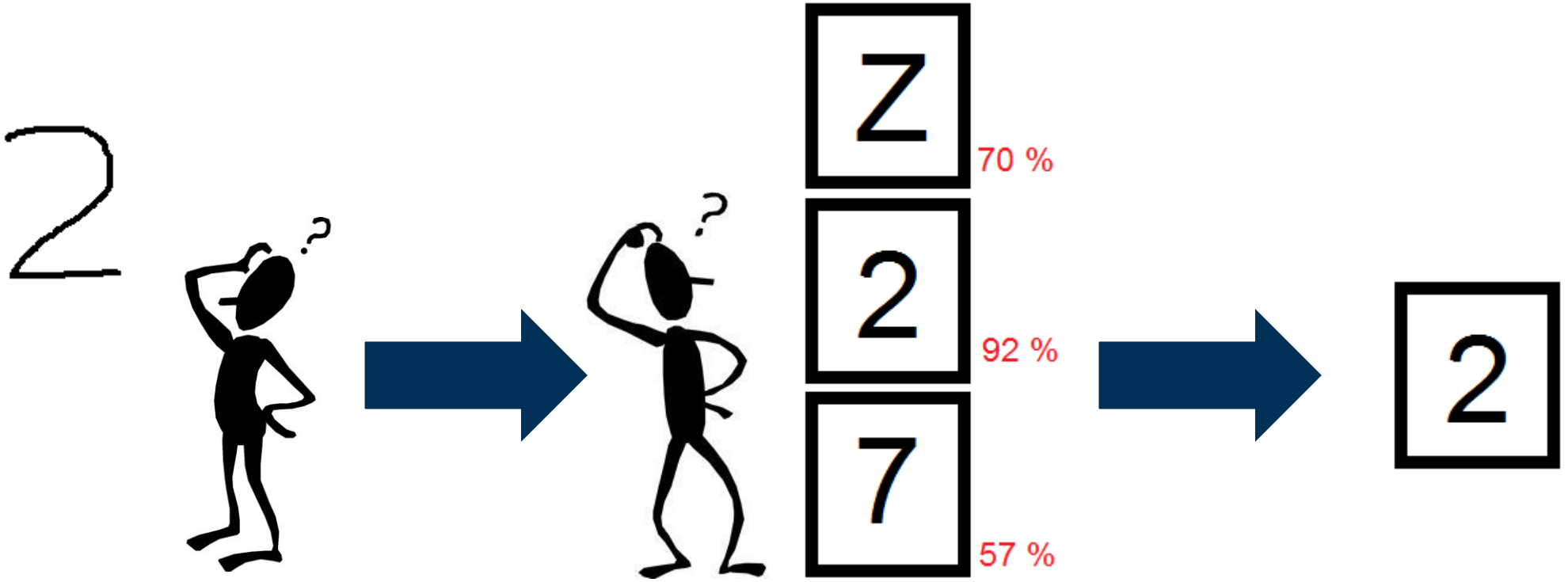
Bücher

Generierung
digitaler Ausgangsformate

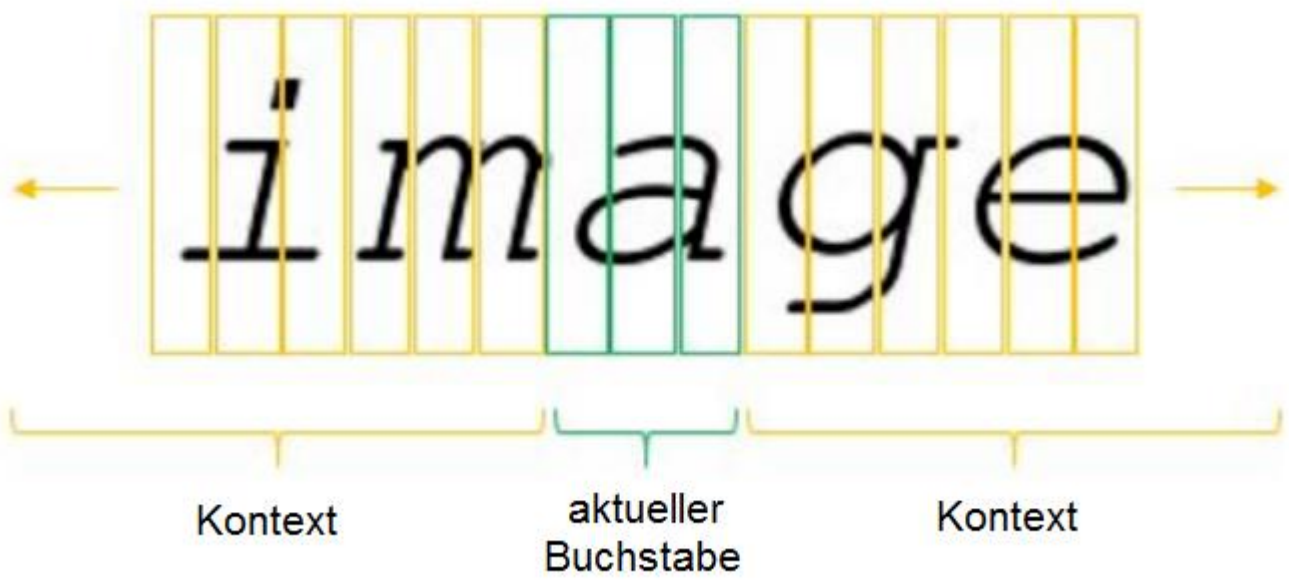
Generierung
der digitalen Inhalte

Strukturierung
der digitalen Inhalte
(Datenextraktion)

OCR-Verfahren – Pattern Matching



OCR-Verfahren – Neuronale Netzwerke



OCR-Software (Übersicht)

kommerzielle Software

fett = eingesetzt in Bibliotheken

ABBYY Finereader
BIT-Alpha
Readiris
OmniPage

Adobe Acrobat
CorelDraw
Microsoft OneNote
…

Tesseract
OCRopus / Kraken / Calamari
CuneiForm
…

freie Software

ABBYY Cloud OCR
Google Cloud Vision
Microsoft Azure Computer Vision
OCR.space Online OCR …

Cloud OCR

Tesseract OCR

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Open Source OCR Software
- Entwicklungsbeginn 1985 von HP
- 2005 als Open Source Projekt veröffentlicht
- 2006 von Google übernommen
- 2016 Umstellung von Pattern Matching auf neuronale Netzwerke

Tesseract OCR

Tesseract OCR

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Open Source
- Komplettlösung „All-in-1“
- Mehr als 100 Sprachen / mehr als 30 Schriften
- Liest Bilder in allen gängigen Formaten (nicht PDF!)
- Erzeugt Text, PDF, hOCR, ALTO, TSV
- Große, weltweite Anwender-Community
- Technologisch aktuell (Texterkennung mit neuronalem Netz)
- Aktive Weiterentwicklung u. a. im DFG-Projekt OCR-D

Tesseract an der UB Mannheim

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Verwendung im DFG-Projekt „Aktienführer“ https://digi.bib.uni-mannheim.de/aktienfuehrer/
- Volltexte für Deutscher Reichsanzeiger und Vorgänger https://digi.bib.uni-mannheim.de/periodika/reichsanzeiger

- DFG-Projekt „OCR-D“ http://www.ocr-d.de/, Modulprojekt „Optimierter Einsatz von OCR-Verfahren – Tesseract als Komponente im OCR-D-Workflow“: Schnittstellen, Stabilität, Performance und praktische Einsetzbarkeit, Erweiterungen wie z. B. Konfidenzen

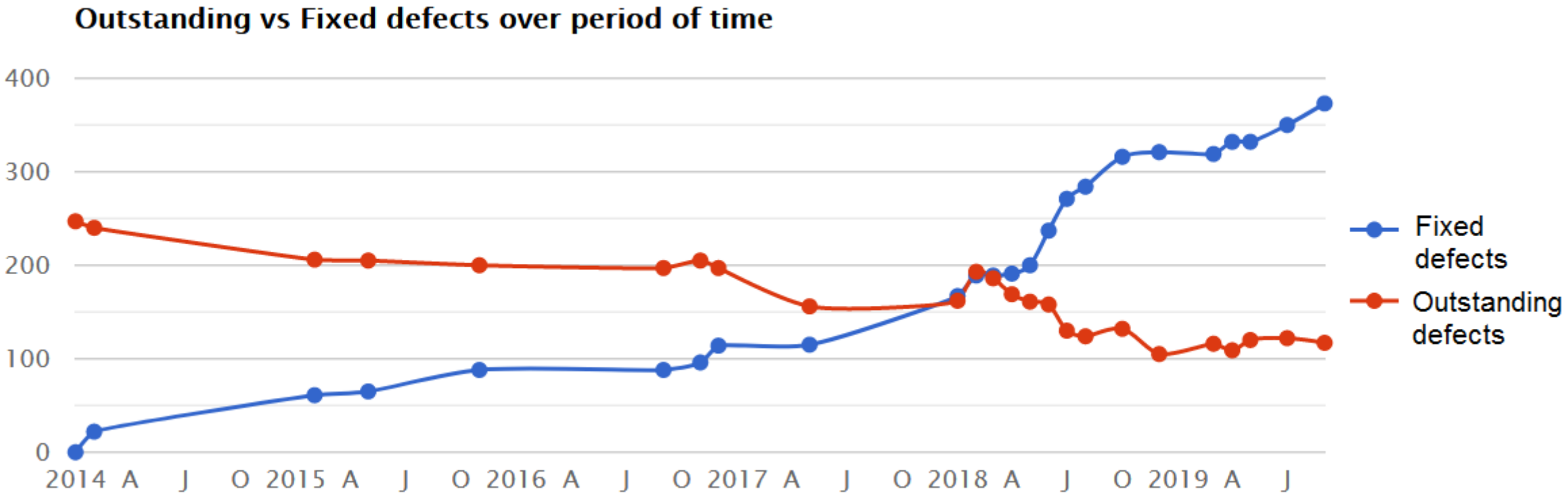
OCR-D
Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR)

DFG-Projekt „OCR-D“
Zielsetzung

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

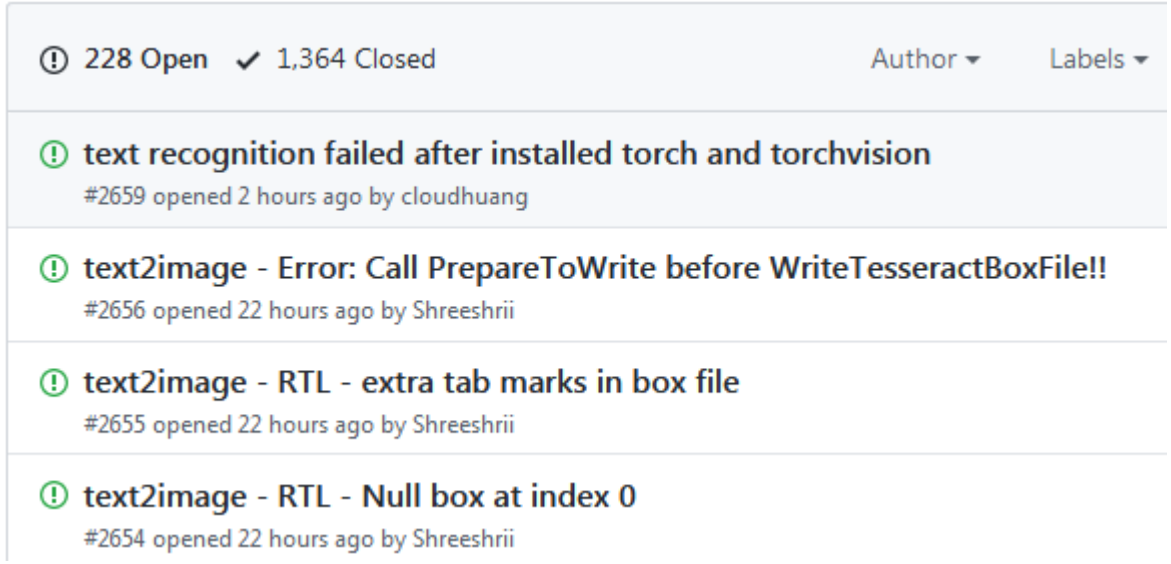
- Stabilitätsverbesserung
- Performanceverbesserung
- Bereitstellung von Schnittstellen für andere Modulprojekte

DFG-Projekt „OCR-D“
Stabilitätsverbesserung –Coverity Scan



DFG-Projekt „OCR-D“
Stabilitätsverbesserung – GitHub

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM



DFG-Projekt „OCR-D“
Stabilitätsverbesserung – andere Tools

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

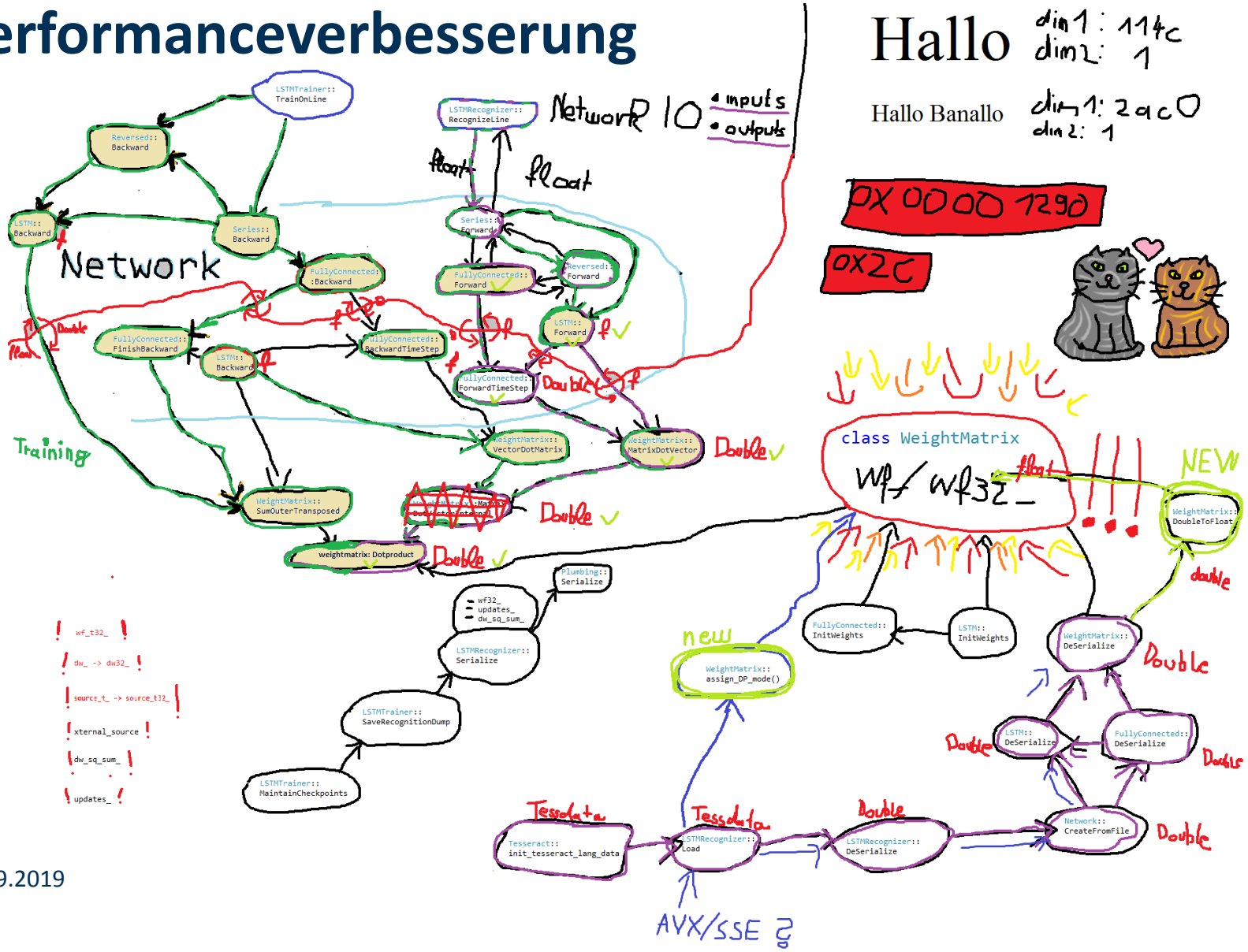
- C++ Compiler Warnungen — Mehr als 100 Korrekturen
- LGTM (Looks good to me) — Etwa 40 Korrekturen
- Google OSS – Fuzz

DFG-Projekt „OCR-D“
Performanceverbesserung

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Etwa **90 %** der verwendeten Rechenzeit wird für Skalarprodukte aufgewendet
- Verwendung von 32-bit Werten anstelle der ursprünglich verwendeten 64-bit Werten
- Nutzung des Kahan-Summations-Algorithmus, um den entstehenden Verlust an Genauigkeit zu kompensieren

DFG-Projekt „OCR-D“
Performanceverbesserung



DFG-Projekt „OCR-D“
Performanceverbesserung

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Durchschnittliche Zeitersparnis von **42,5 %**
- Durchschnittliche Performanceverbesserung von **74 %**
- Trotz geringerer Genauigkeit **keine** Abweichung der Ergebnisse

DFG-Projekt „OCR-D“
Schnittstellen für andere Modul Projekte

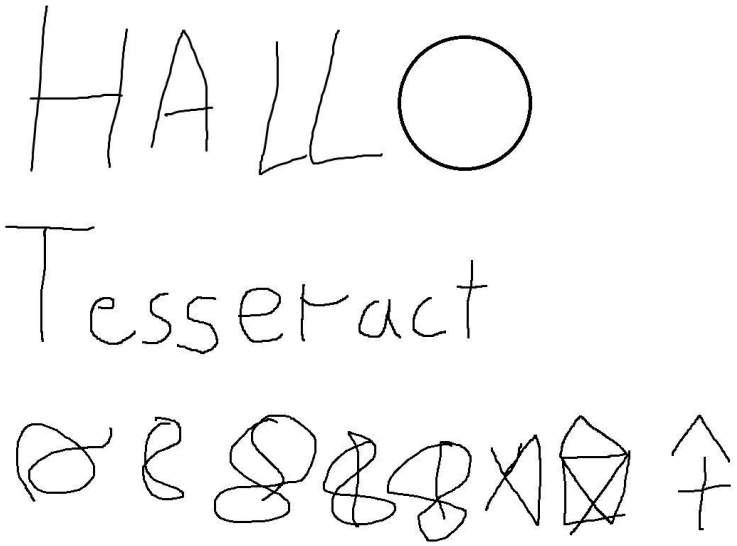
UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

UNIVERSITÄT LEIPZIG

- Geplante OCR post-Korrektur der Universität Leipzig

- Bereitstellung eines Ausgabemodus, welcher zusätzliche Zeichen-Alternativen zu dem bestehenden Ergebnis liefert.

DFG-Projekt „OCR-D“
Auslesen der LSTM Alternativen




```
>HALLO
='choice_1_1_1' title='x_confs 99'>H</span><span class='ocr_glyph' id='choice_1_1_2' title='x_confs 0'>A</sp
='choice_1_1_3' title='x_confs 91'>A</span><span class='ocr_glyph' id='choice_1_1_4' title='x_confs 4'> </sp
='choice_1_1_8' title='x_confs 53'>L</span><span class='ocr_glyph' id='choice_1_1_9' title='x_confs 32'>I</s
='choice_1_1_15' title='x_confs 94'>L</span><span class='ocr_glyph' id='choice_1_1_16' title='x_confs 1'> </
='choice_1_1_21' title='x_confs 100'>O</span></span></span>

07 -22; x_size 126.37594; x_descenders 16.37594; x_ascenders 44">
>Tesceract
='choice_1_2_1' title='x_confs 28'>T</span><span class='ocr_glyph' id='choice_1_2_2' title='x_confs 17'> </s
='choice_1_2_18' title='x_confs 88'>e</span><span class='ocr_glyph' id='choice_1_2_19' title='x_confs 6'>c</
='choice_1_2_23' title='x_confs 53'>s</span><span class='ocr_glyph' id='choice_1_2_24' title='x_confs 25'>c<
='choice_1_2_31' title='x_confs 59'>c</span><span class='ocr_glyph' id='choice_1_2_32' title='x_confs 35'>s<
='choice_1_2_35' title='x_confs 83'>e</span><span class='ocr_glyph' id='choice_1_2_36' title='x_confs 9'>t</
='choice_1_2_38' title='x_confs 93'>r</span><span class='ocr_glyph' id='choice_1_2_39' title='x_confs 6'>t</
='choice_1_2_40' title='x_confs 98'>a</span><span class='ocr_glyph' id='choice_1_2_41' title='x_confs 1'>u</
='choice_1_2_42' title='x_confs 100'>c</span></span>
='choice_1_2_43' title='x_confs 100'>t</span></span></span>
```

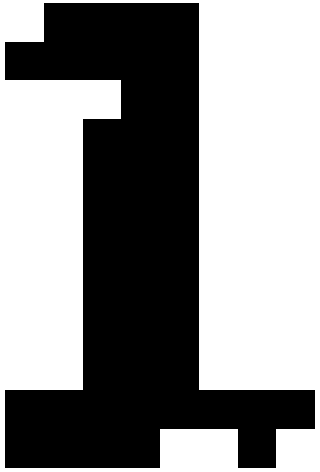
DFG-Projekt „OCR-D“
Auslesen der LSTM Alternativen – Nachteile

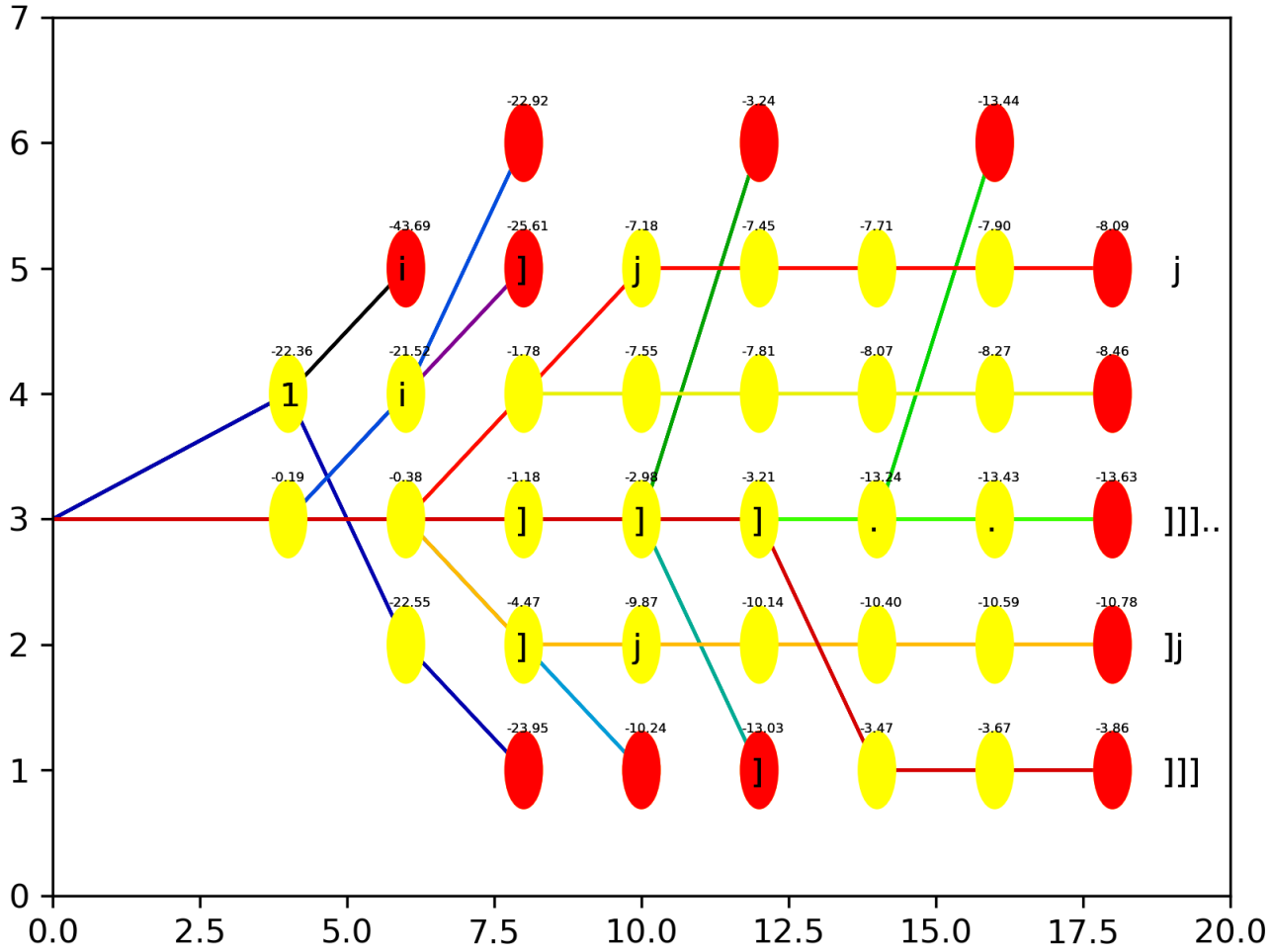
- Nur korrekte Ergebnisse für Sprachen, deren Zeichenzahl geringer ist, als die maximale Anzahl an Ausgabekanälen des LSTMs.

- Die punktuellen Wahrscheinlichkeiten sind nicht repräsentativ für die Entscheidungsfindung des LSTMs.

DFG-Projekt „OCR-D“
Auslesen der Strahlensuche

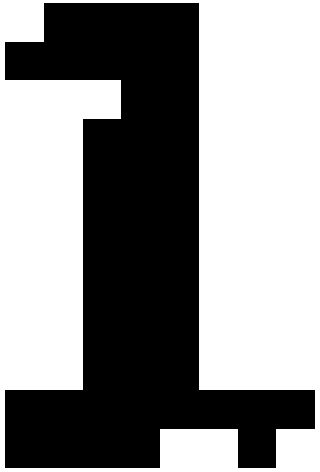
- Nur wenige Zeichenalternativen erreichen das Ende der Strahlensuche

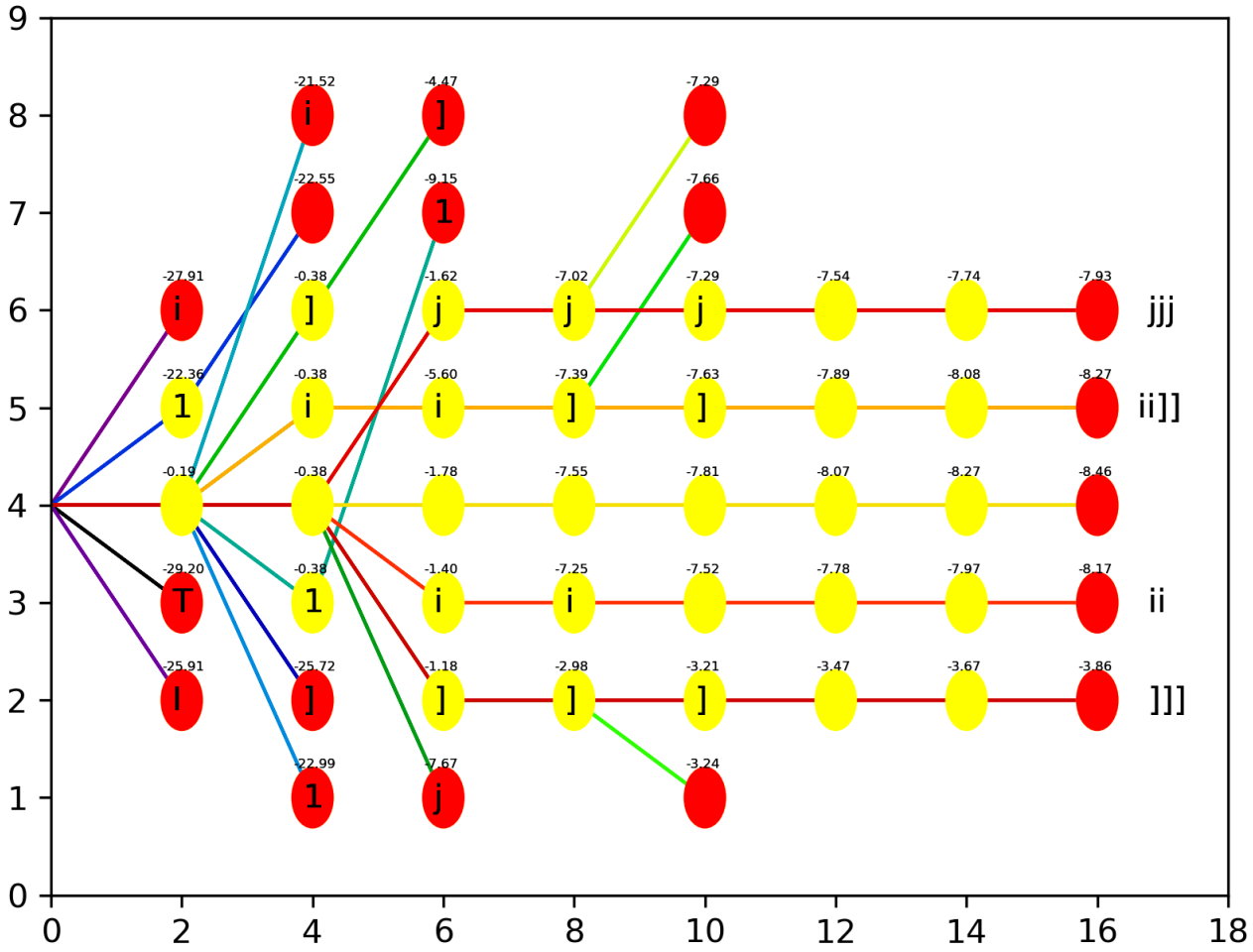




DFG-Projekt „OCR-D“
Auslesen der Strahlensuche – Erweiterte Kandidaten

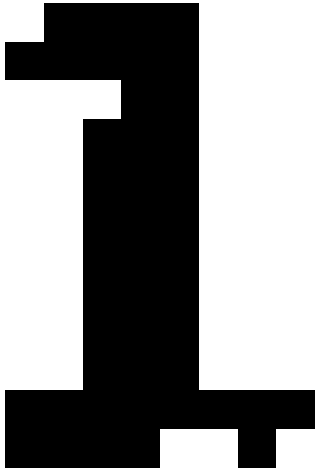
- Bei Erhöhung der Kandidaten kein Unterschied in den finalen Zuständen

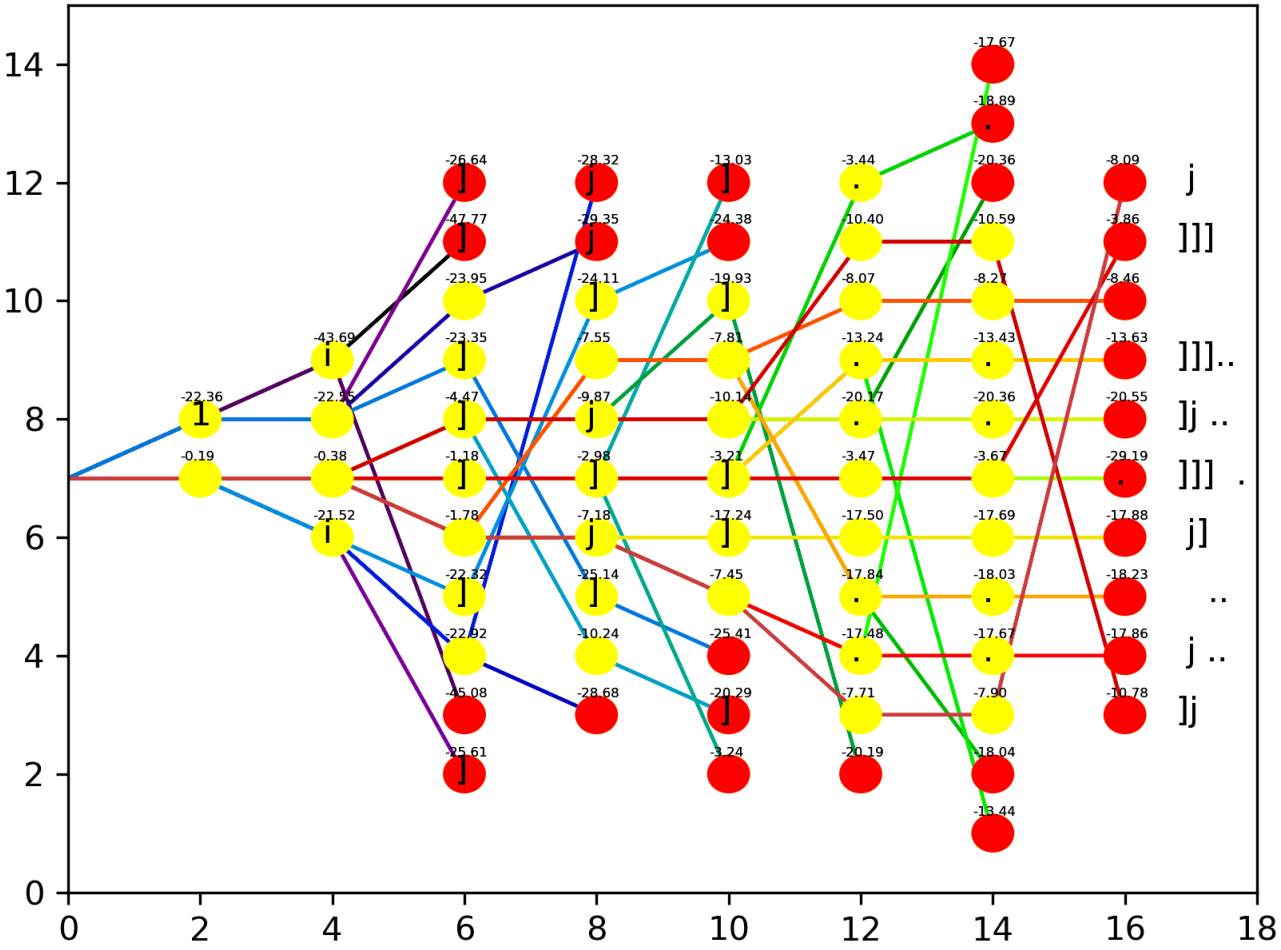




DFG-Projekt „OCR-D“
Auslesen der Strahlensuche – Erweiterter Strahl

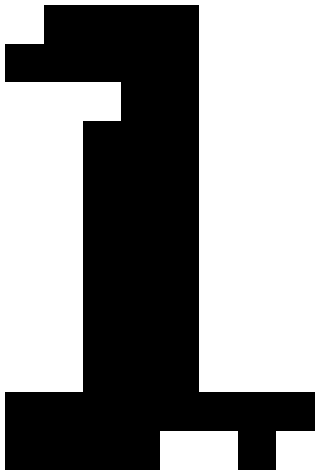
- Eine Erweiterung des Strahls führt zu mehr Duplikaten in der Ausgabe

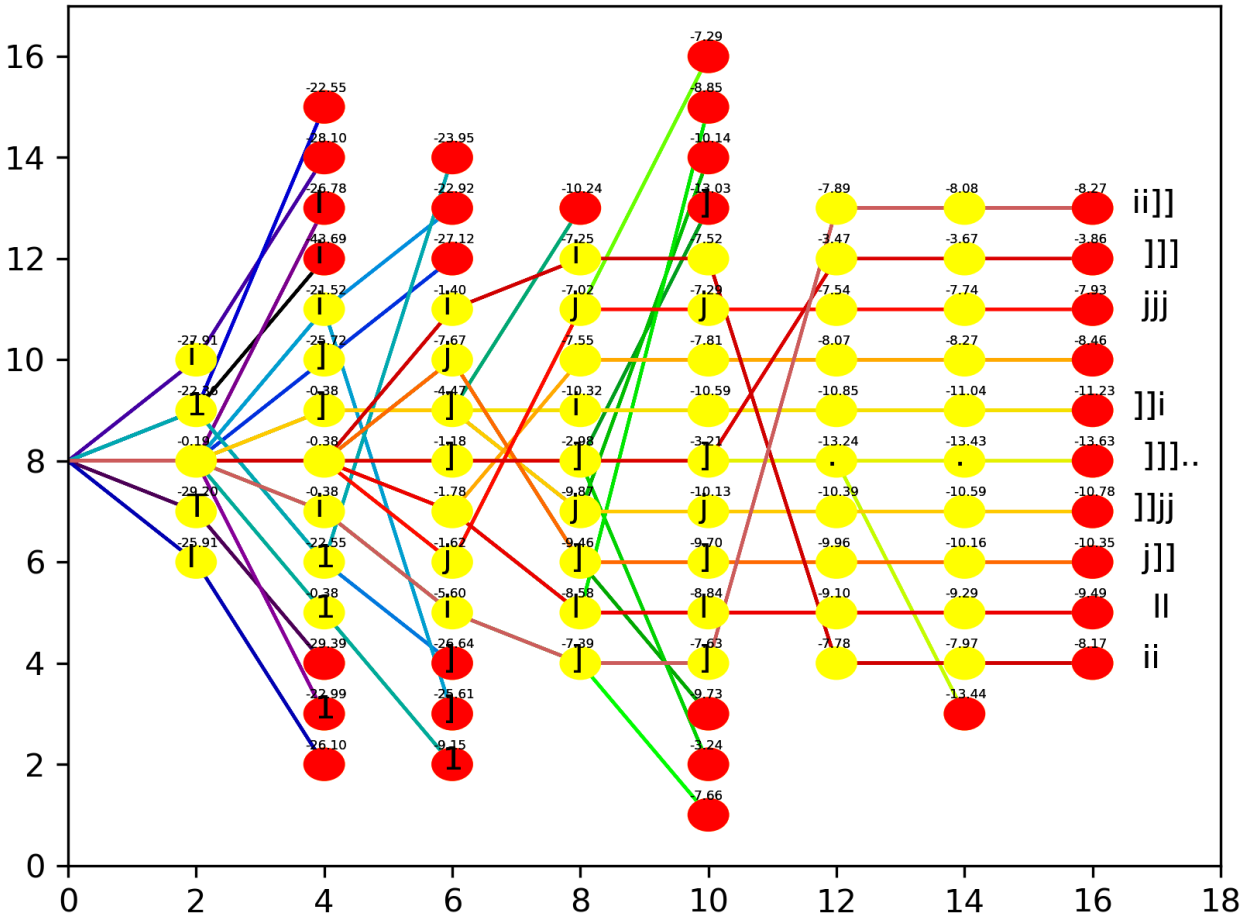




DFG-Projekt „OCR-D“
Auslesen der Strahlensuche – Kombinierte Erweiterung

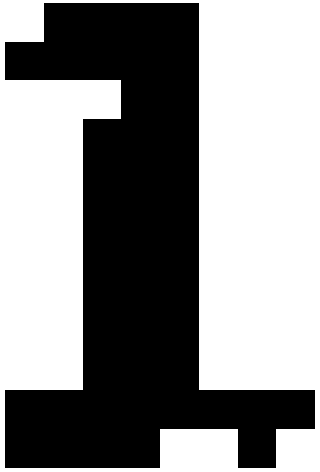
- Eine Erweiterung beider Faktoren führte ebenfalls nicht zu besseren Ergebnissen

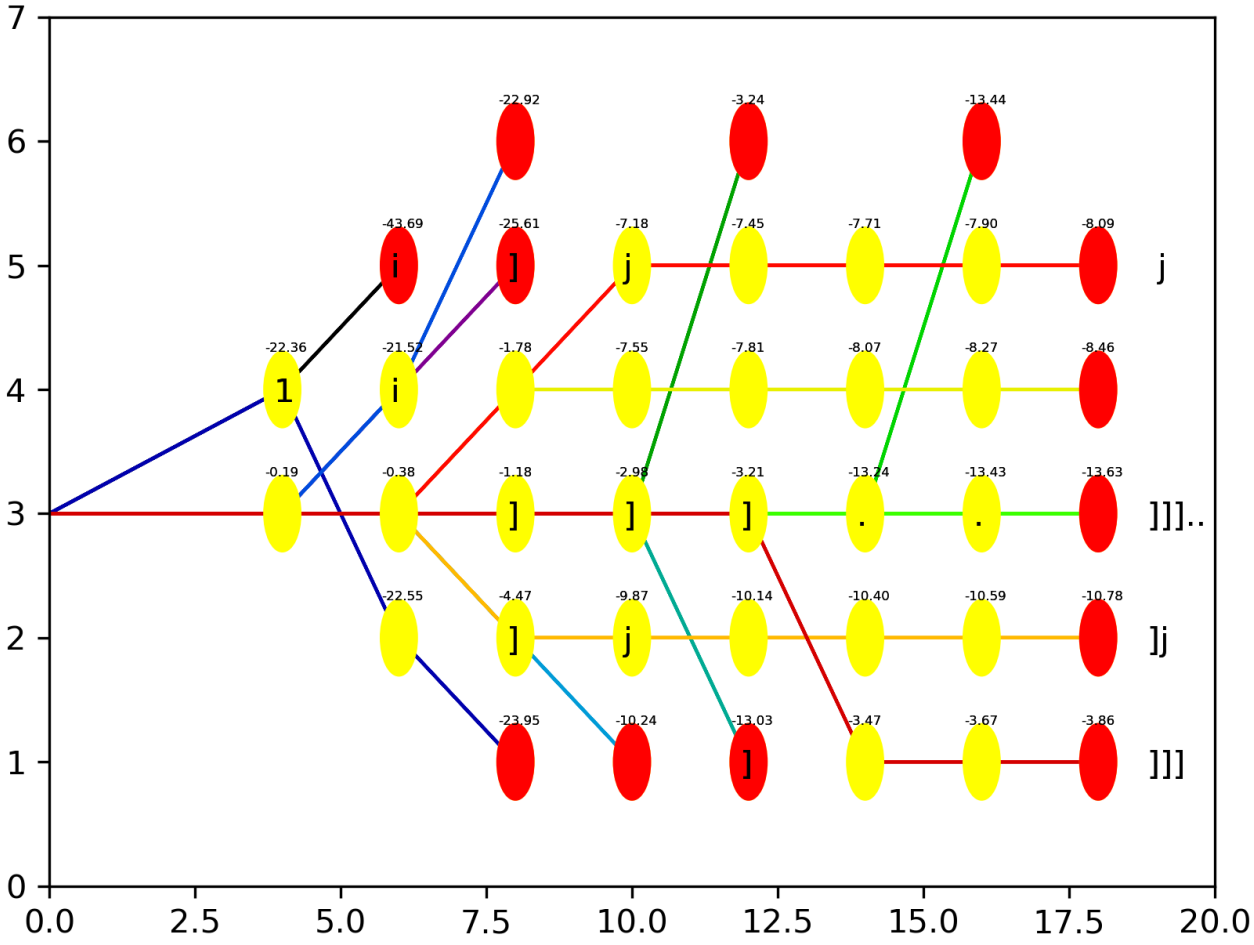




DFG-Projekt „OCR-D“
Auslesen der Strahlensuche – Kaskadierende Suche

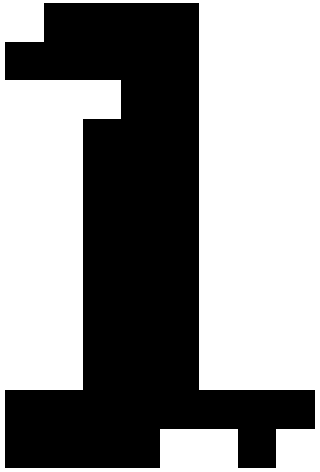
- Wiederholte Suchen mit angepassten Kandidaten führten zum gewünschten Ergebnis

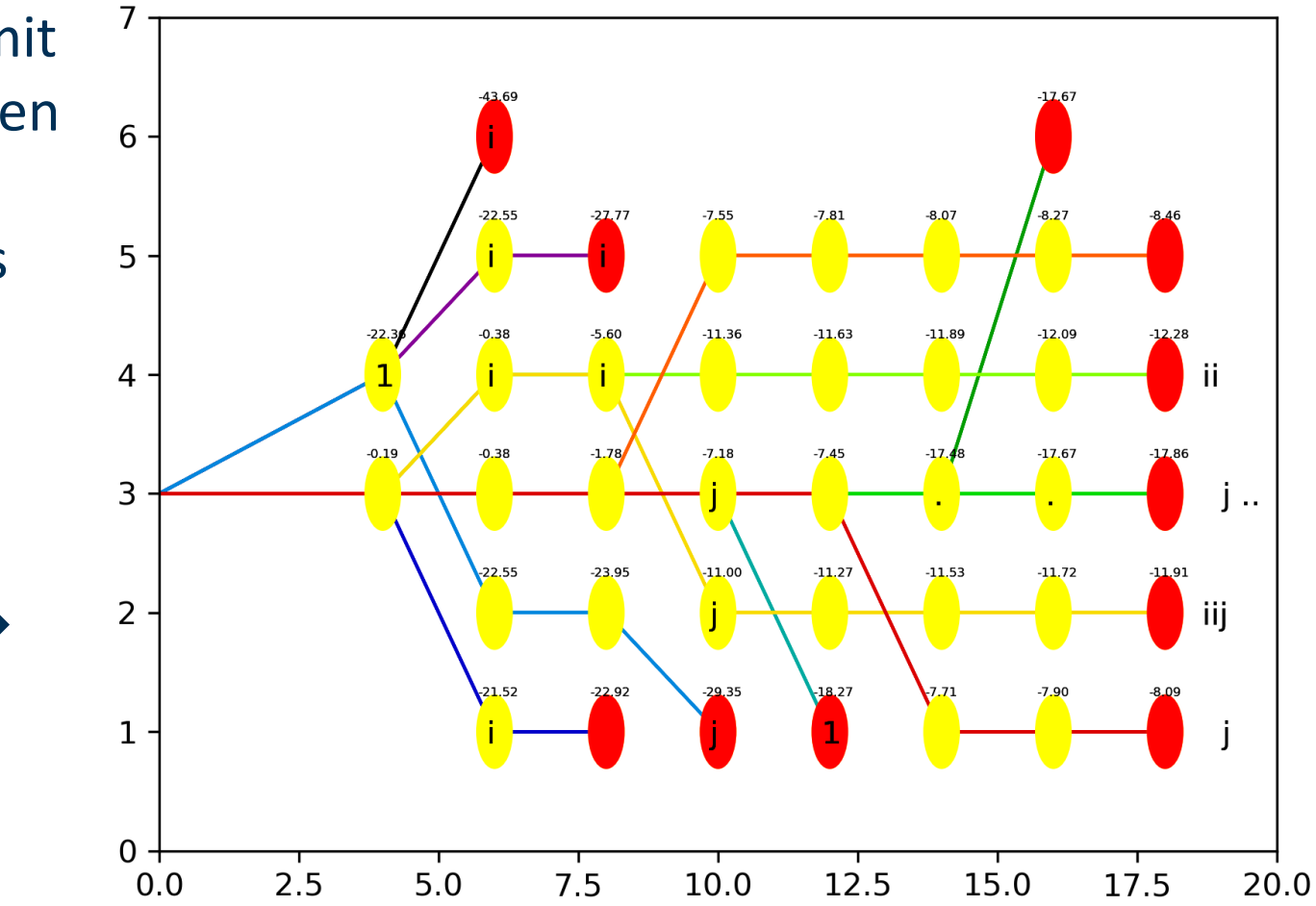




DFG-Projekt „OCR-D“
Auslesen der Strahlensuche – Kaskadierende Suche

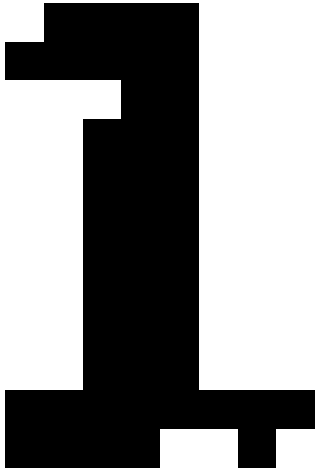
- Wiederholte Suchen mit angepassten Kandidaten führten zum gewünschten Ergebnis

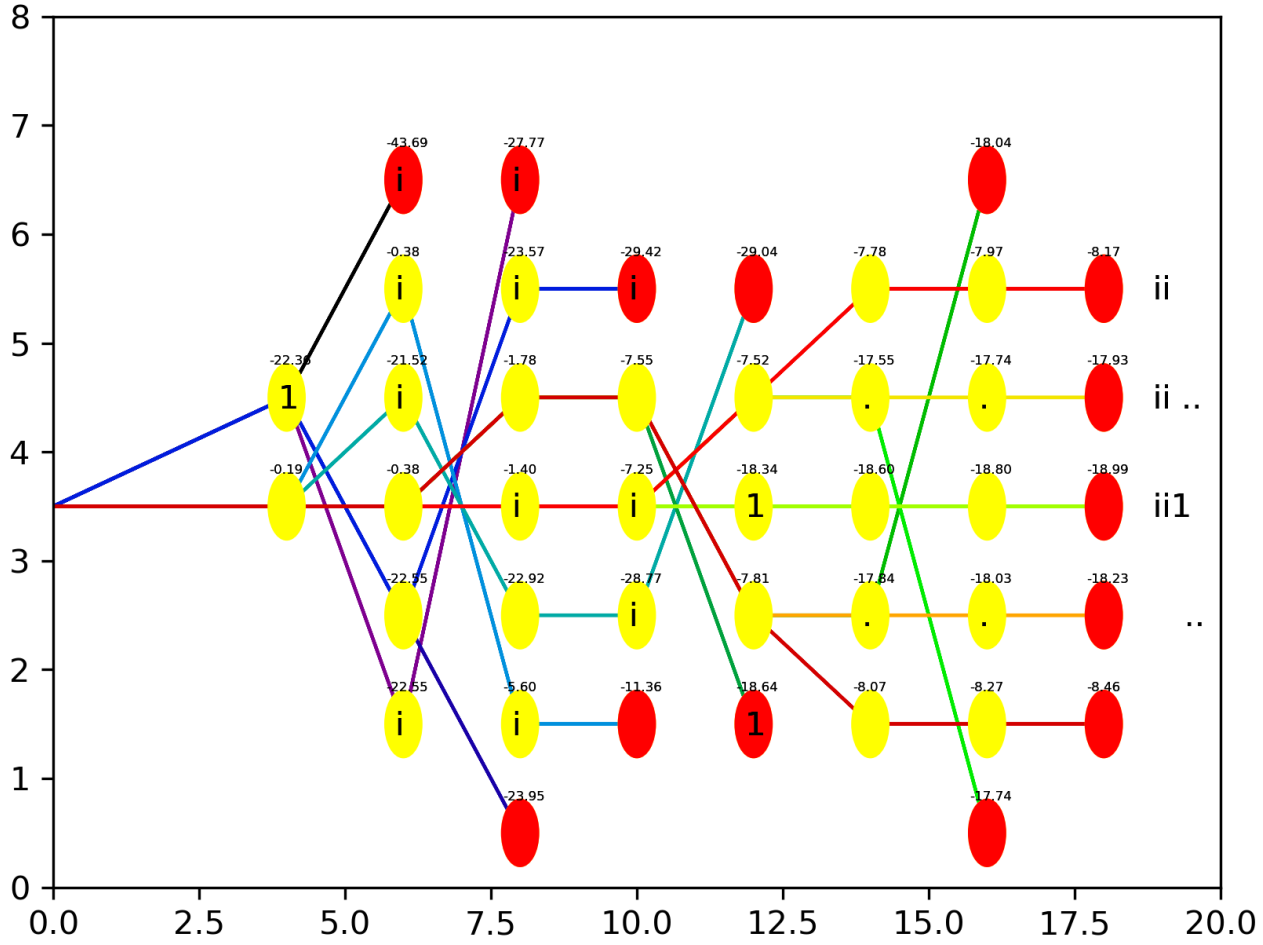




DFG-Projekt „OCR-D“
Auslesen der Strahlensuche – Kaskadierende Suche

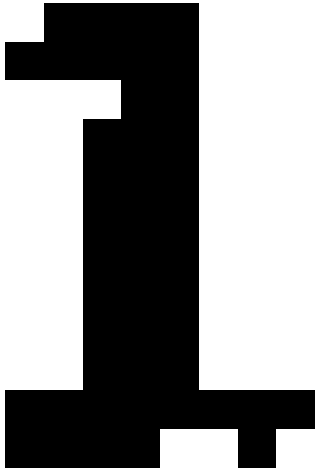
- Wiederholte Suchen mit angepassten Kandidaten führten zum gewünschten Ergebnis

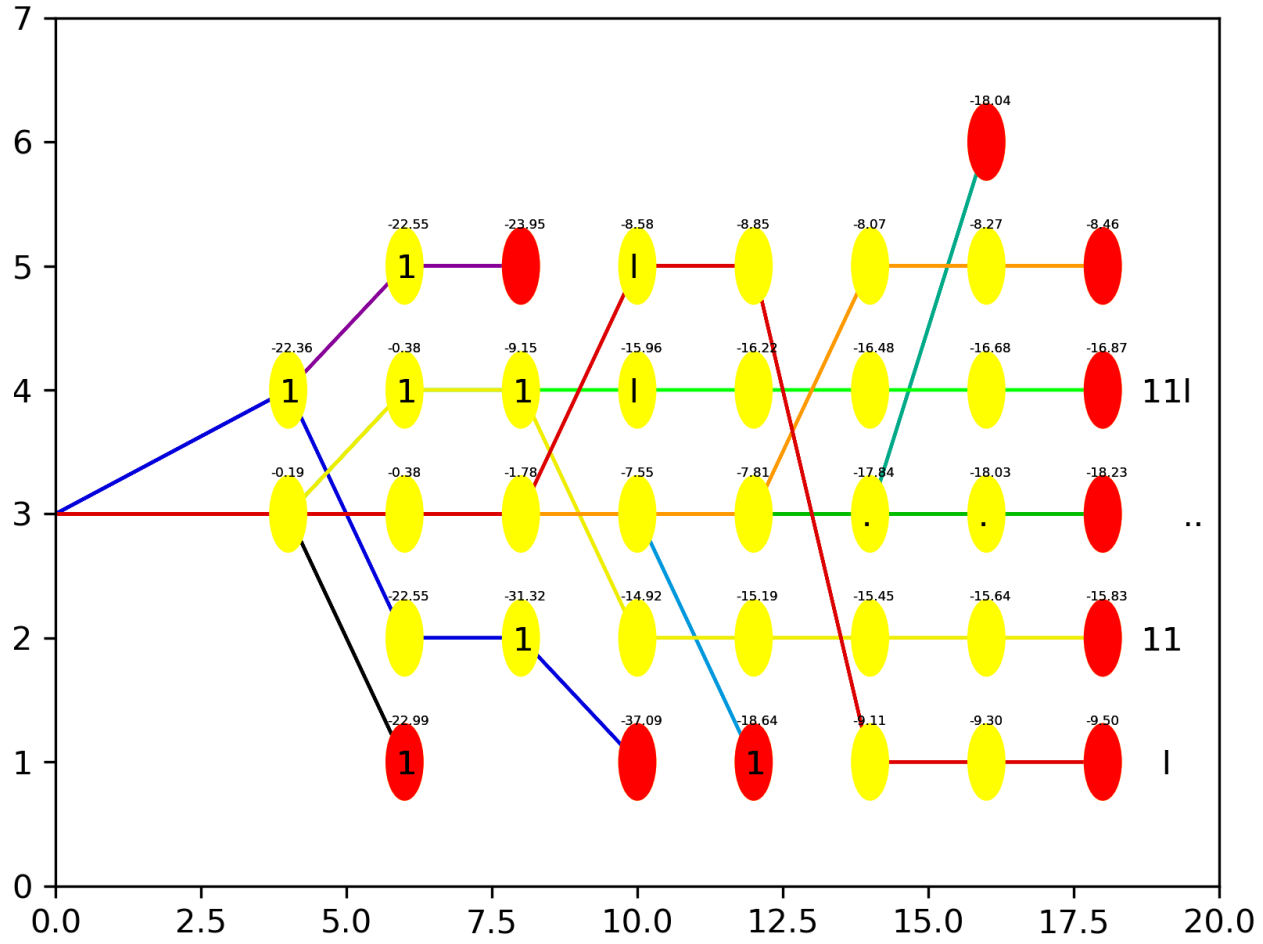




DFG-Projekt „OCR-D“
Auslesen der Strahlensuche – Kaskadierende Suche

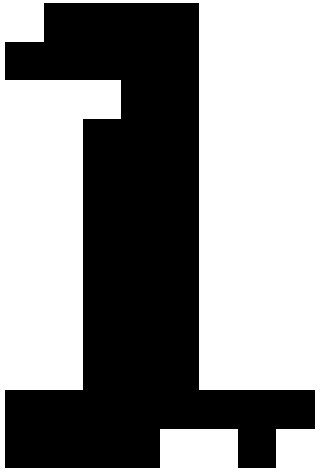
- Wiederholte Suchen mit angepassten Kandidaten führten zum gewünschten Ergebnis

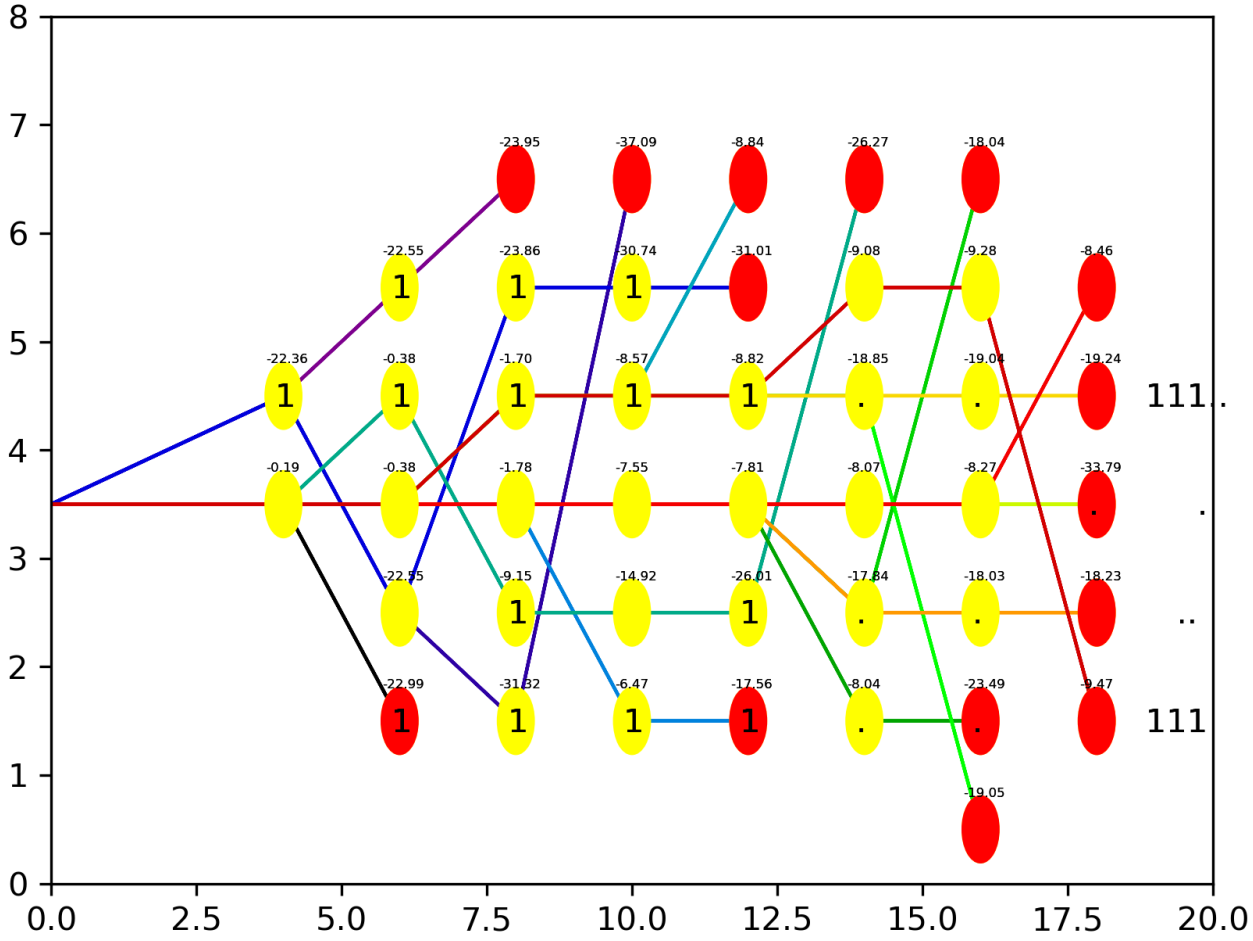




DFG-Projekt „OCR-D“
Auslesen der Strahlensuche – Kaskadierende Suche

- Wiederholte Suchen mit angepassten Kandidaten führten zum gewünschten Ergebnis

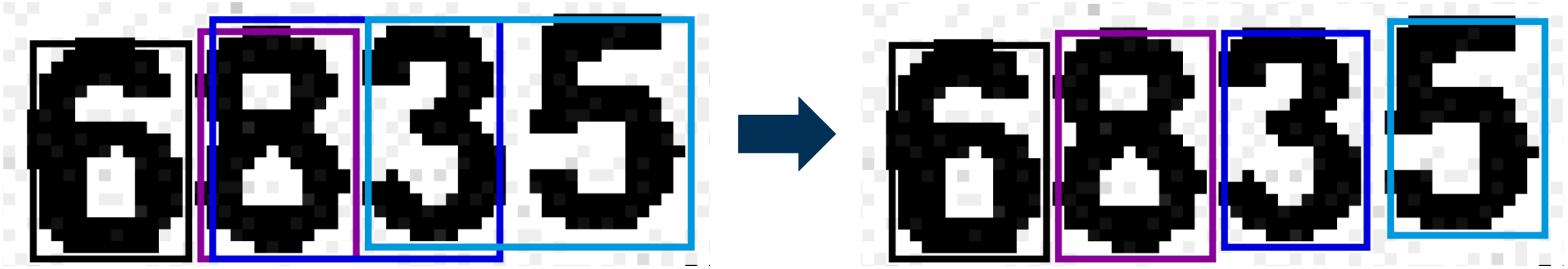




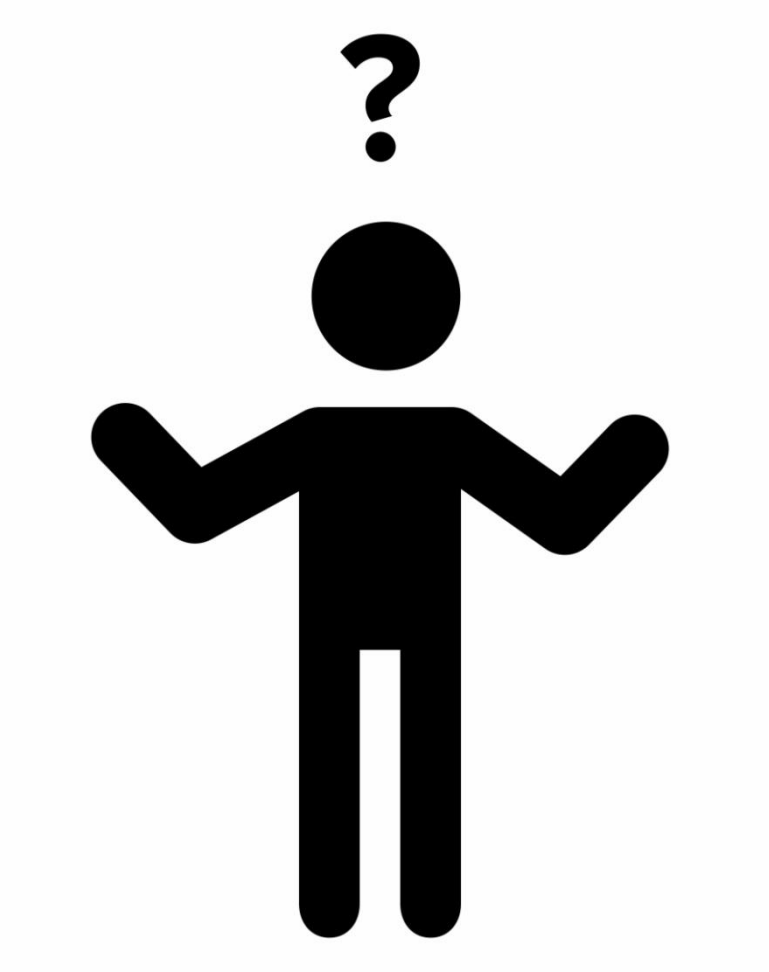
DFG-Projekt „OCR-D“
Positive Nebeneffekte

- Allgemeine Anwendbarkeit der Schnittstelle

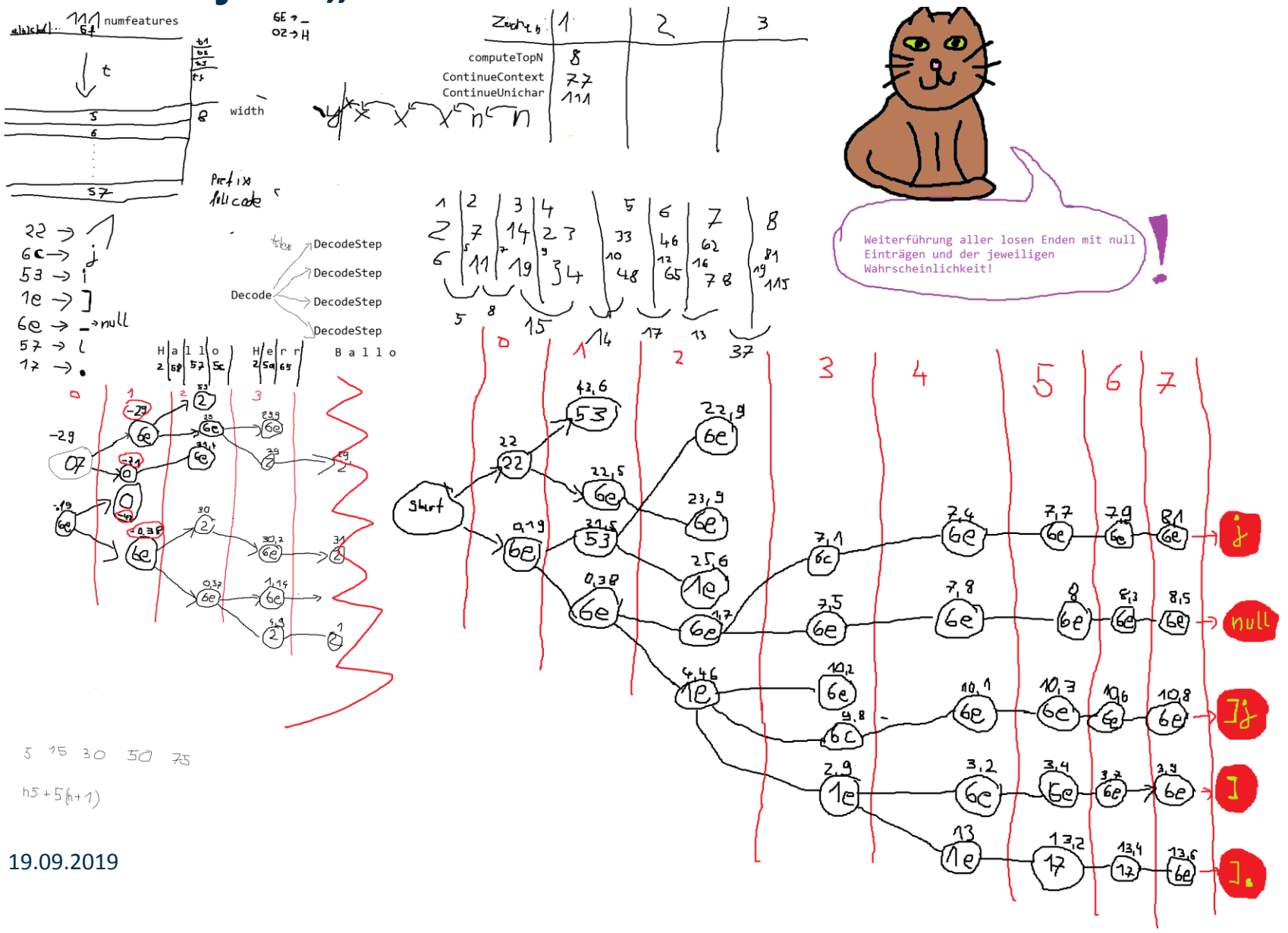
- Wiederverwendung von Teilalgorithmen zur Erstellung besserer Begrenzungsrahmen im Standardprozess von Tesseract



DFG-Projekt „OCR-D“



DFG-Projekt „OCR-D“



DFG-Projekt „OCR-D“