

VISUELLE UNTERSCHIEDE IN SCREENSHOTS VON ARCHIVierten WEBSEITEN

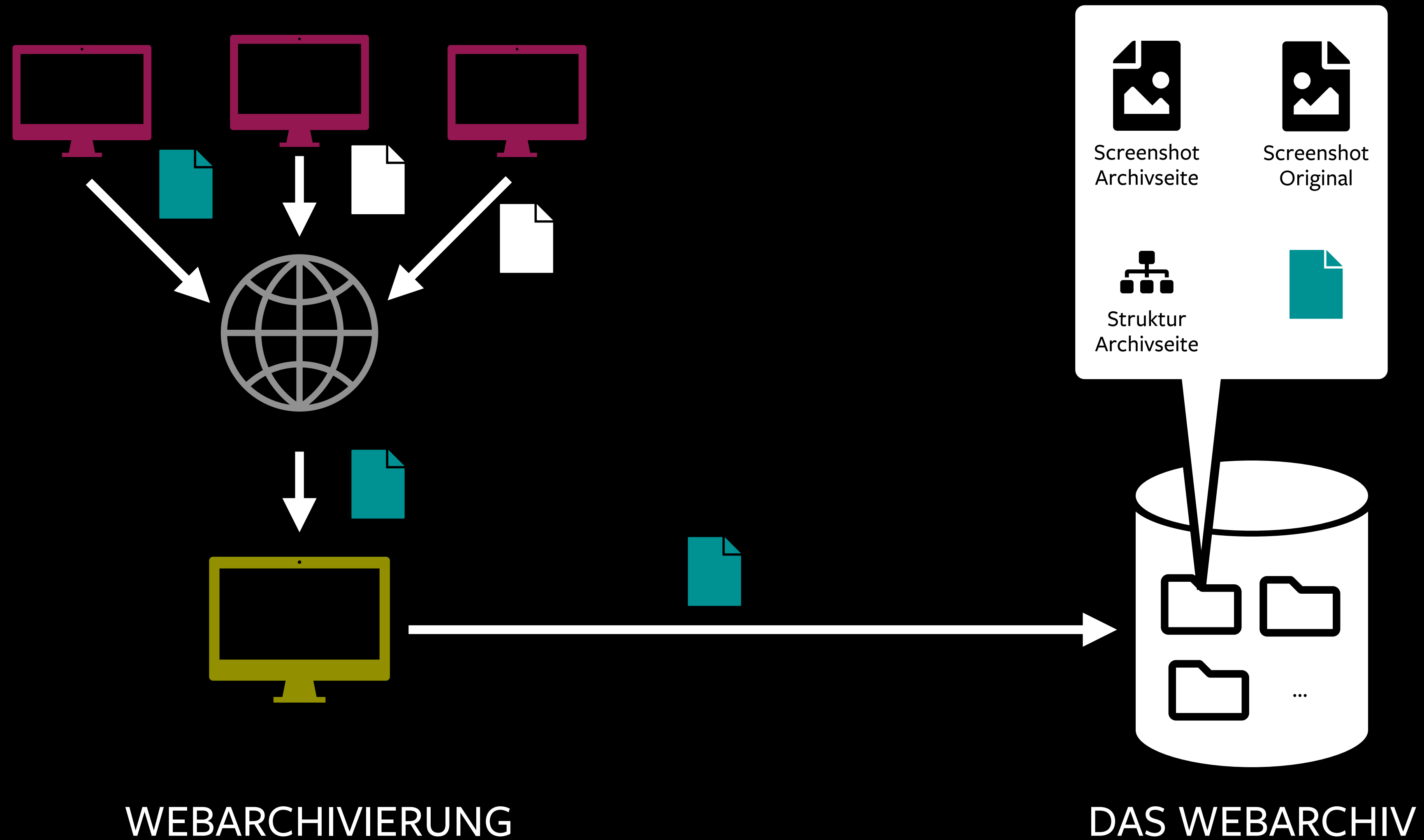
Masterarbeit von Theresa Elstner

INHALT

- ▶ MOTIVATION: QUALITÄTSBEWERTUNG VON WEBARCHIVEN
- ▶ ARCHIVIERUNGSFEHLER
 - ▶ FEHLERTYPEN
 - ▶ FEHLERMAß
- ▶ GENERELLES FRAMEWORK ZUR QUALITÄTSBEWERTUNG
- ▶ UMSETZUNG DES FRAMEWORKS
 - ▶ VERSCHIEBUNGEN
- ▶ ZUSAMMENFASSUNG

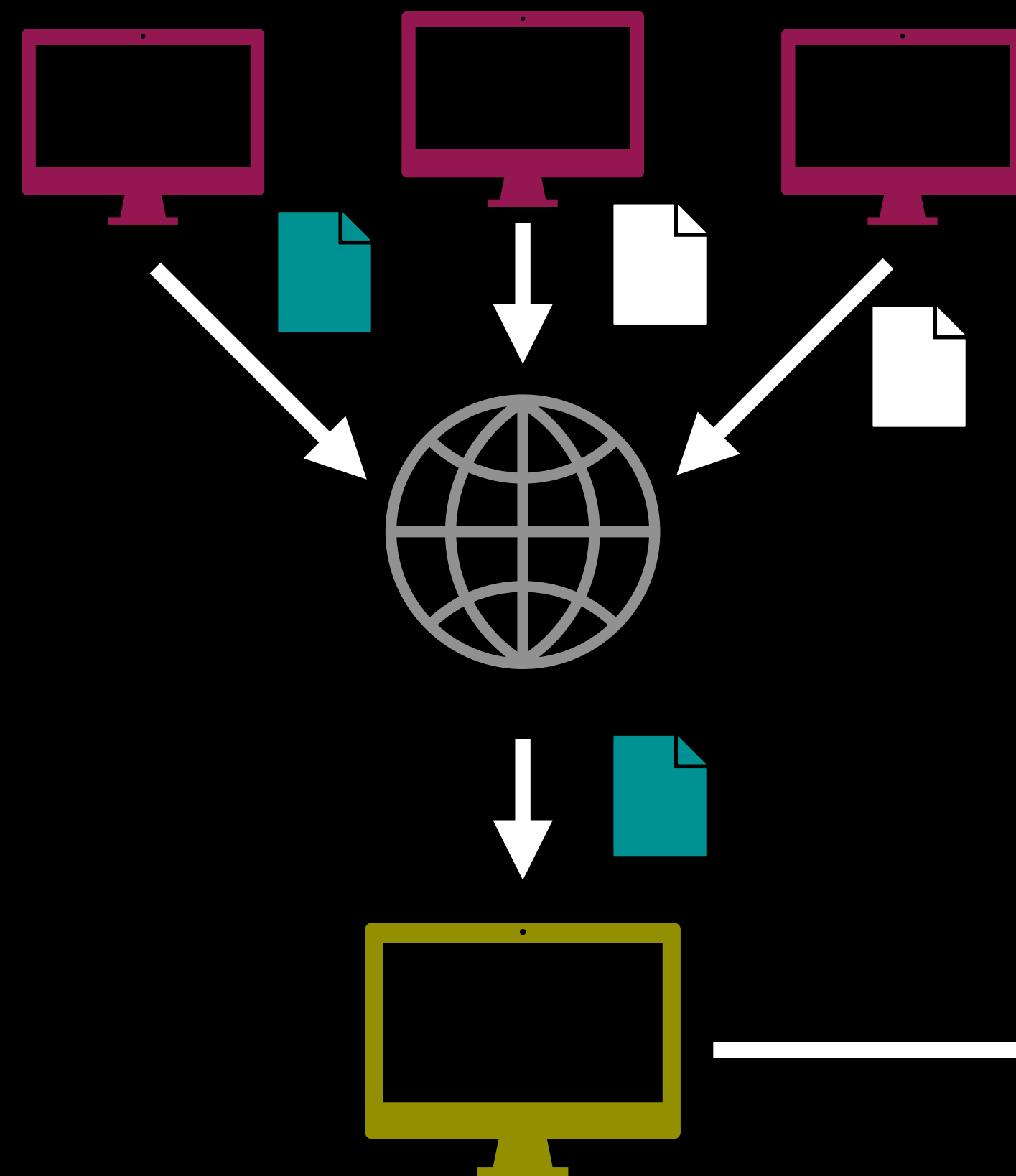
MOTIVATION

WEBARCHIVE

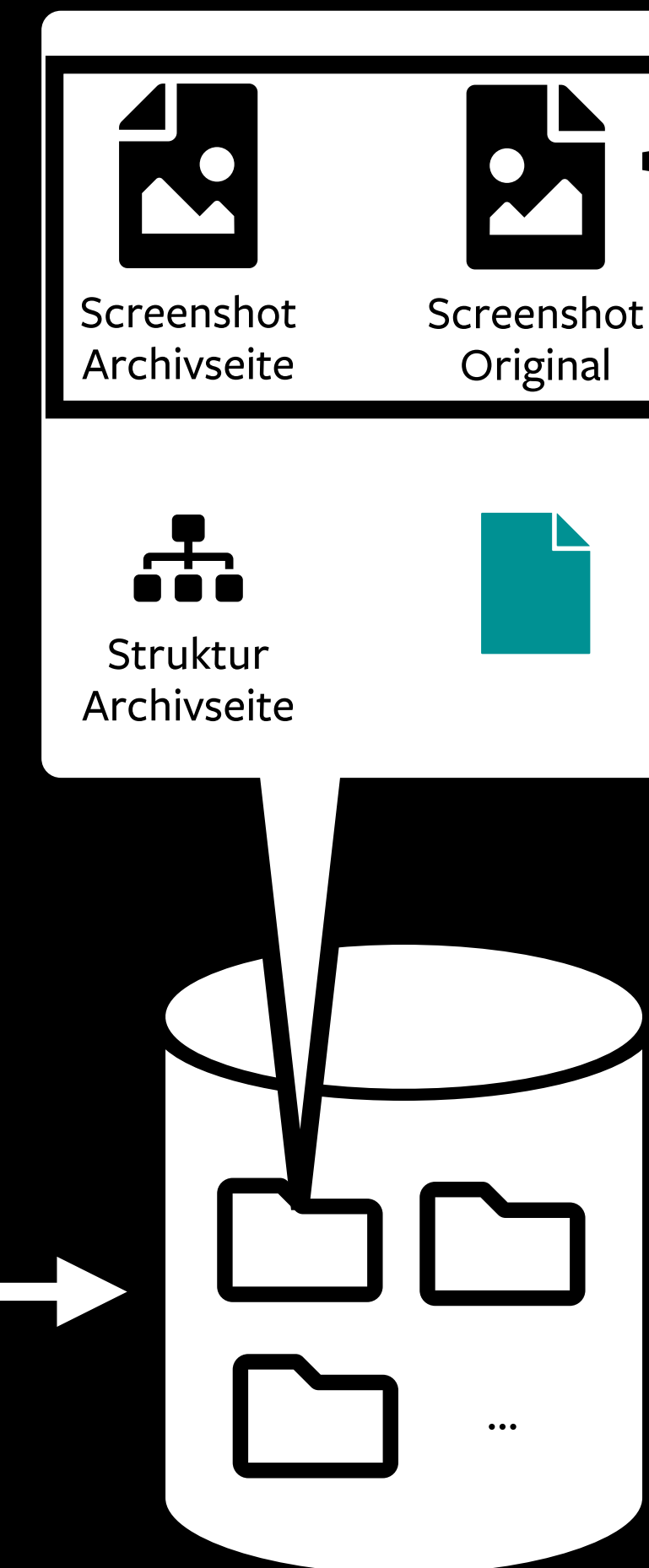


MOTIVATION

WEBARCHIVE



WEBARCHIVIERUNG



HUMAN
ANNOTATIONS

DAS WEBARCHIV

MOTIVATION QUALITÄT VON WEBARCHIVEN

► PROBLEM

- archivierte Seite ist nicht immer identisch zur originalen Webseite
- einige archivierte Webseiten unbrauchbar
- Qualitätsbewertung der archivierten Webseiten benötigt
 - manuell?
 - automatisiert!

Creators Remix Roundup: Top 10 Remixes Of The Year



<p>Including Florence + the Machine, DJ Wordy, Glasser, Major Lazer, Mark Ronson, G-Dragon, Gang Gang Dance and James Lavelle.</p>

SHARE TWEET

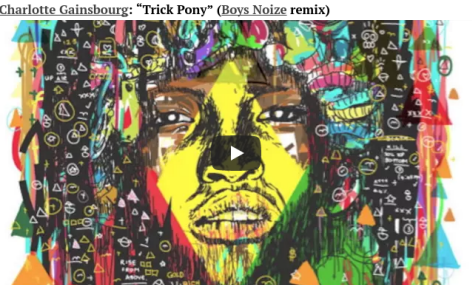
Our Creators are a talented and prolific bunch, and our inbox is always overflowing with alerts of new remixes and mashups from the incredible DJs and producers in our line-up. We just couldn't keep these fresh new tunes to ourselves because, after all, flesharing is caring. To commemorate the one year anniversary of the Creators Remix Roundup (which happens to fall on the magical day of 11-11-11, no less) we decided to revisit our favorite tunes from the past year... which ones were your favorites?



There seems to be a few different adaptations of Jamie xx's rework floating around the internet, but this 2-step version has to be our favorite. Last winter this track was in heavy, heavy rotation on our iPods and since just this morning we debuted our documentary on the enchanting Florence + the Machine, we thought it was the perfect time to bring it back. The original feels like a comforting, glamorous lullaby, but we like the added male vocals (from The xx's Oliver Sim) and Jamie's glitchy loops.

Missy Elliott: "Work It" (DJ Wordy remix)

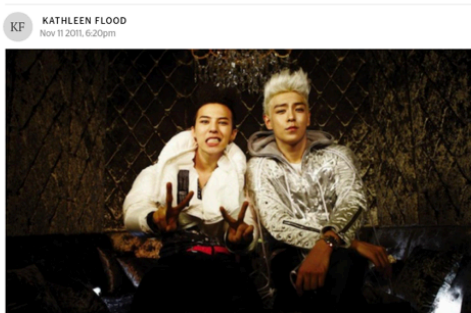
Chinese hit-man DJ Wordy been "Work[ing] It" hard on a new album under the moniker WordySoulSpeak, and we love this groovy remix of Missy Elliott's oft-remixed classic. Listen, download and see for yourself how well Wordy's soulful horns match with Missy's musical stylings.



We thought the original recording of French chanteuse Charlotte Gainsbourg's "Trick Pony" off her 2009 Beck-produced album *IRM*, was pretty good to begin with. Sometimes messing with something already pretty perfect can be overkill, but Alex Ridha of *Boys Noize* has proved that this certainly isn't the case. About 20 seconds in, he starts laying down the dirty bass, and giving some electro polish to the instrumentation—also a nice complement to Gainsbourg's bewitching vocal arrangements. This is Charlotte like you've never heard her before.

Beastie Boys feat. Santigold: "Don't Play No Game That I Can't Win" (Major Lazer remix)

Creators Remix Roundup: Top 10 Remixes Of The Year



<p>Including Florence + the Machine, DJ Wordy, Glasser, Major Lazer, Mark Ronson, G-Dragon, Gang Gang Dance and James Lavelle.</p>

SHARE TWEET

Our Creators are a talented and prolific bunch, and our inbox is always overflowing with alerts of new remixes and mashups from the incredible DJs and producers in our line-up. We just couldn't keep these fresh new tunes to ourselves because, after all, flesharing is caring. To commemorate the one year anniversary of the Creators Remix Roundup (which happens to fall on the magical day of 11-11-11, no less) we decided to revisit our favorite tunes from the past year... which ones were your favorites?



There seems to be a few different adaptations of Jamie xx's rework floating around the internet, but this 2-step version has to be our favorite. Last winter this track was in heavy, heavy rotation on our iPods and since just this morning we debuted our documentary on the enchanting Florence + the Machine, we thought it was the perfect time to bring it back. The original feels like a comforting, glamorous lullaby, but we like the added male vocals (from The xx's Oliver Sim) and Jamie's glitchy loops.

Missy Elliott: "Work It" (DJ Wordy remix)

Chinese hit-man DJ Wordy been "Work[ing] It" hard on a new album under the moniker WordySoulSpeak, and we love this groovy remix of Missy Elliott's oft-remixed classic. Listen, download and see for yourself how well Wordy's soulful horns match with Missy's musical stylings.



We thought the original recording of French chanteuse Charlotte Gainsbourg's "Trick Pony" off her 2009 Beck-produced album *IRM*, was pretty good to begin with. Sometimes messing with something already pretty perfect can be overkill, but Alex Ridha of *Boys Noize* has proved that this certainly isn't the case. About 20 seconds in, he starts laying down the dirty bass, and giving some electro polish to the instrumentation—also a nice complement to Gainsbourg's bewitching vocal arrangements. This is Charlotte like you've never heard her before.

Beastie Boys feat. Santigold: "Don't Play No Game That I Can't Win" (Major Lazer remix)



New York hip-hop legends, the Beastie Boys, finally released their much anticipated studio album *Hip Suma Committee Part Two* on May 3rd, which was delayed in part because band member Adam "MCA" Yauch was diagnosed with cancer (he's doing better now). Digga and Switch, aka Major Lazer, give their second single the reggae-fab moonbalmton "treatment" with extra spice courtesy of Santigold.



archivierte Seite

originale Seite

MOTIVATION

QUALITÄTSBEWERTUNG

- ▶ 1. Idee: Pixelvergleich
- ▶ vergleiche jeden Pixel aus Original- mit korrespondierendem im Archivbild
- ▶ je größer die Anzahl der unterschiedlichen Pixel, umso schlechter die Qualität
- ▶ ABER...

MOTIVATION

QUALITÄTSBEWERTUNG

- ... Anzahl unterschiedlicher Pixel sagt wenig über Qualität:
- relativer Fehler: ca. 32 %

archivierte Seite

Pixelunterschiede

■ : gleich ■ : verschieden

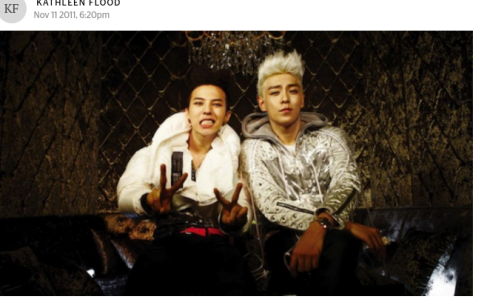
originale Seite

MOTIVATION

QUALITÄTSBEWERTUNG

- ▶ **Intuition:**
 - ▶ rekonstruiere die originale Webseite so *gut es geht* aus der archivierten Webseite
 - ▶ z.B. verschiebe falsch positionierte Elemente
 - ▶ vergleiche die Rekonstruktion mit dem Original
 - ▶ erhalte einen Fehler
 - ▶ wenn Fehler zu hoch, markiere die archivierte Seite als unbrauchbar

Creators Remix Roundup: Top 10 Remixes Of The Year



<p>Including Florence + the Machine, DJ Wordy, Glasser, Major Lazer, Mark Ronson, G-Dragon, Gang Gang Dance and James Lavelle.</p>

SHARE

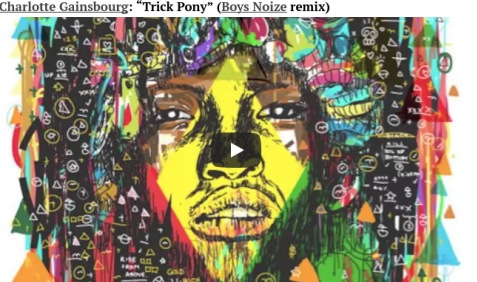
Our Creators are a talented and prolific bunch, and our inbox is always overflowing with alerts of new remixes and mashups from the incredible DJs and producers in our line-up. We just couldn't keep these fresh new tunes to ourselves because, after all, flesharing is caring. To commemorate the one year anniversary of the Creators Remix Roundup (which happens to fall on the magical day of 11-11-11, no less) we decided to revisit our favorite tunes from the past year... which ones were your favorites?



There seems to be a few different adaptations of Jamie xx's rework floating around the internet, but this 2-step version has to be our favorite. Last winter this track was in heavy, heavy rotation on our iPods and since just this morning we debuted our documentary on the enchanting Florence + the Machine, we thought it was the perfect time to bring it back. The original feels like a comforting, glamorous lullaby, but we like the added male vocals (from The xx's Oliver Sim) and Jamie's glitchy loops.

Missy Elliott: "Work It" (DJ Wordy remix)

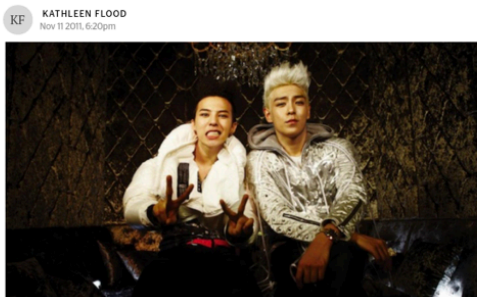
Chinese hit-man DJ Wordy been "Work[ing] It" hard on a new album under the moniker WordySoulSpeak, and we love this groovy remix of Missy Elliott's oft-remixed classic. Listen, download and see for yourself how well Wordy's soulful horns match with Missy's musical stylings.



We thought the original recording of French chanteuse Charlotte Gainsbourg's "Trick Pony" off her 2009 Beck-produced album *IRM*, was pretty good to begin with. Sometimes messing with something already pretty perfect can be overkill, but Alex Ridha of *Boys Noize* has proved that this certainly isn't the case. About 20 seconds in, he starts laying down the dirty bass, and giving some electro polish to the instrumentation—also a nice complement to Gainsbourg's bewitching vocal arrangements. This is Charlotte like you've never heard her before.

Beastie Boys feat. Santigold: "Don't Play No Game That I Can't Win" (Major Lazer remix)

Creators Remix Roundup: Top 10 Remixes Of The Year



<p>Including Florence + the Machine, DJ Wordy, Glasser, Major Lazer, Mark Ronson, G-Dragon, Gang Gang Dance and James Lavelle.</p>

SHARE

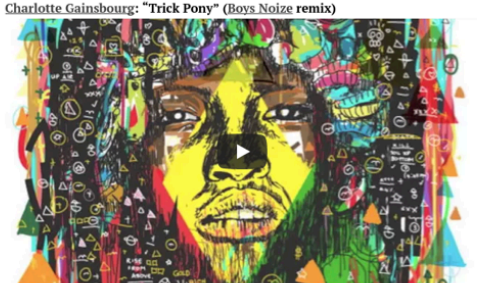
Our Creators are a talented and prolific bunch, and our inbox is always overflowing with alerts of new remixes and mashups from the incredible DJs and producers in our line-up. We just couldn't keep these fresh new tunes to ourselves because, after all, flesharing is caring. To commemorate the one year anniversary of the Creators Remix Roundup (which happens to fall on the magical day of 11-11-11, no less) we decided to revisit our favorite tunes from the past year... which ones were your favorites?



There seems to be a few different adaptations of Jamie xx's rework floating around the internet, but this 2-step version has to be our favorite. Last winter this track was in heavy, heavy rotation on our iPods and since just this morning we debuted our documentary on the enchanting Florence + the Machine, we thought it was the perfect time to bring it back. The original feels like a comforting, glamorous lullaby, but we like the added male vocals (from The xx's Oliver Sim) and Jamie's glitchy loops.

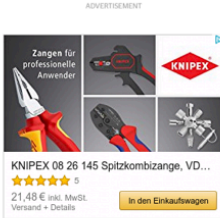
Missy Elliott: "Work It" (DJ Wordy remix)

Chinese hit-man DJ Wordy been "Work[ing] It" hard on a new album under the moniker WordySoulSpeak, and we love this groovy remix of Missy Elliott's oft-remixed classic. Listen, download and see for yourself how well Wordy's soulful horns match with Missy's musical stylings.



We thought the original recording of French chanteuse Charlotte Gainsbourg's "Trick Pony" off her 2009 Beck-produced album *IRM*, was pretty good to begin with. Sometimes messing with something already pretty perfect can be overkill, but Alex Ridha of *Boys Noize* has proved that this certainly isn't the case. About 20 seconds in, he starts laying down the dirty bass, and giving some electro polish to the instrumentation—also a nice complement to Gainsbourg's bewitching vocal arrangements. This is Charlotte like you've never heard her before.

Beastie Boys feat. Santigold: "Don't Play No Game That I Can't Win" (Major Lazer remix)

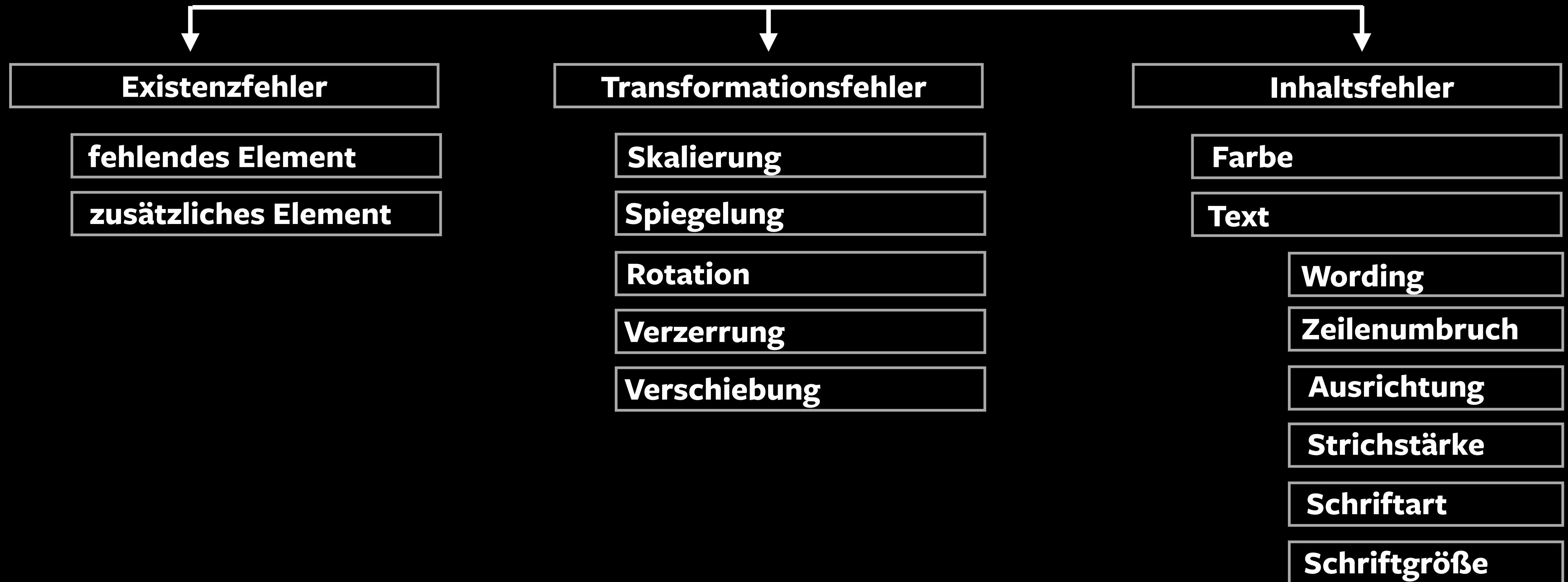


New York hip-hop legends, the Beastie Boys, finally released their much anticipated studio album *Hill Sinner Committee Part Two* on May 3rd, which was delayed in part because band member Adam "MCA" Yauch was diagnosed with cancer (he's doing better now). Diplo and Switch, aka Major Lazer, give their second single the reggae-fab moonbalmton "treatment" with extra spice courtesy of Santigold.



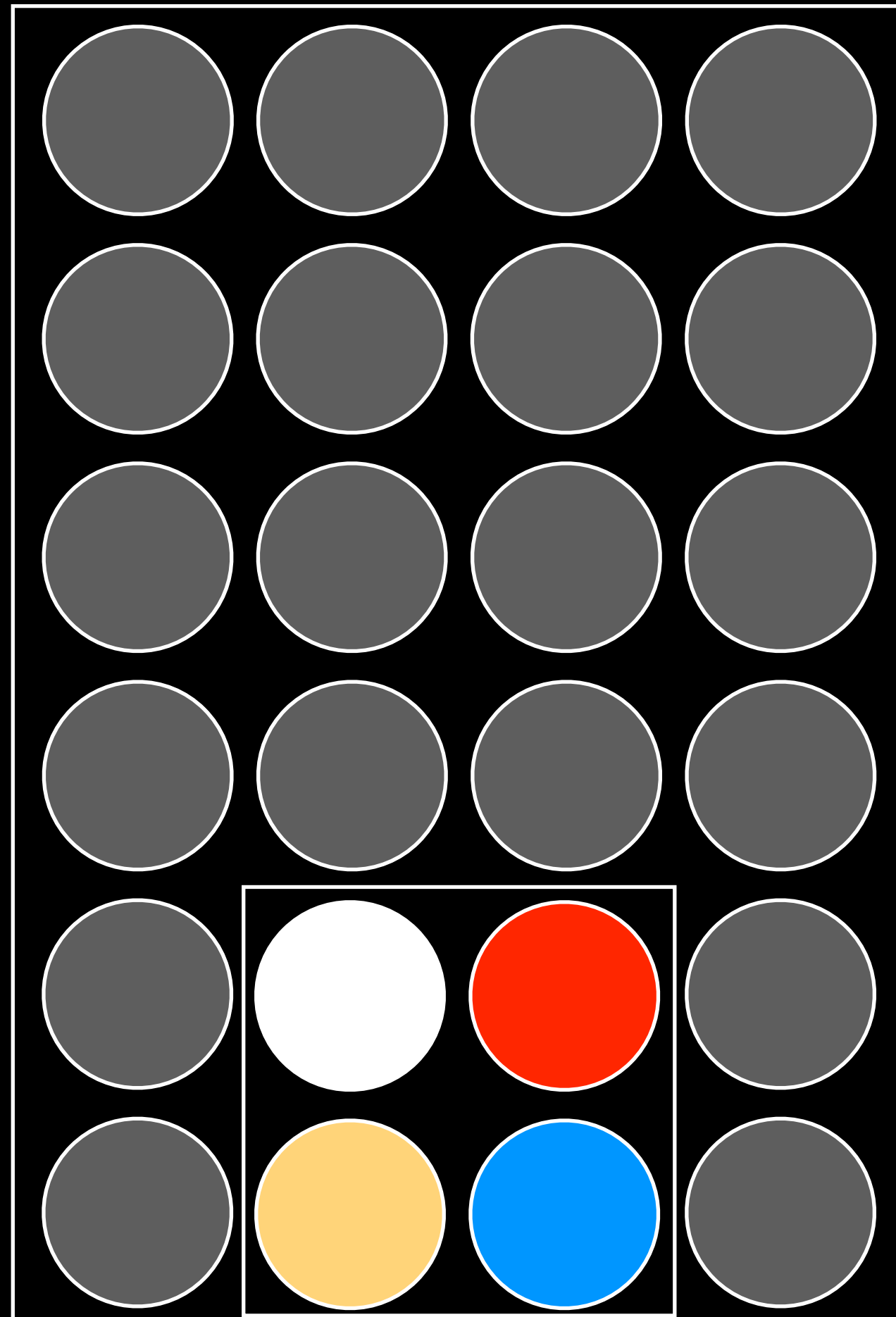
ARCHIVIERUNGSFEHLER

FEHLERTYPEN

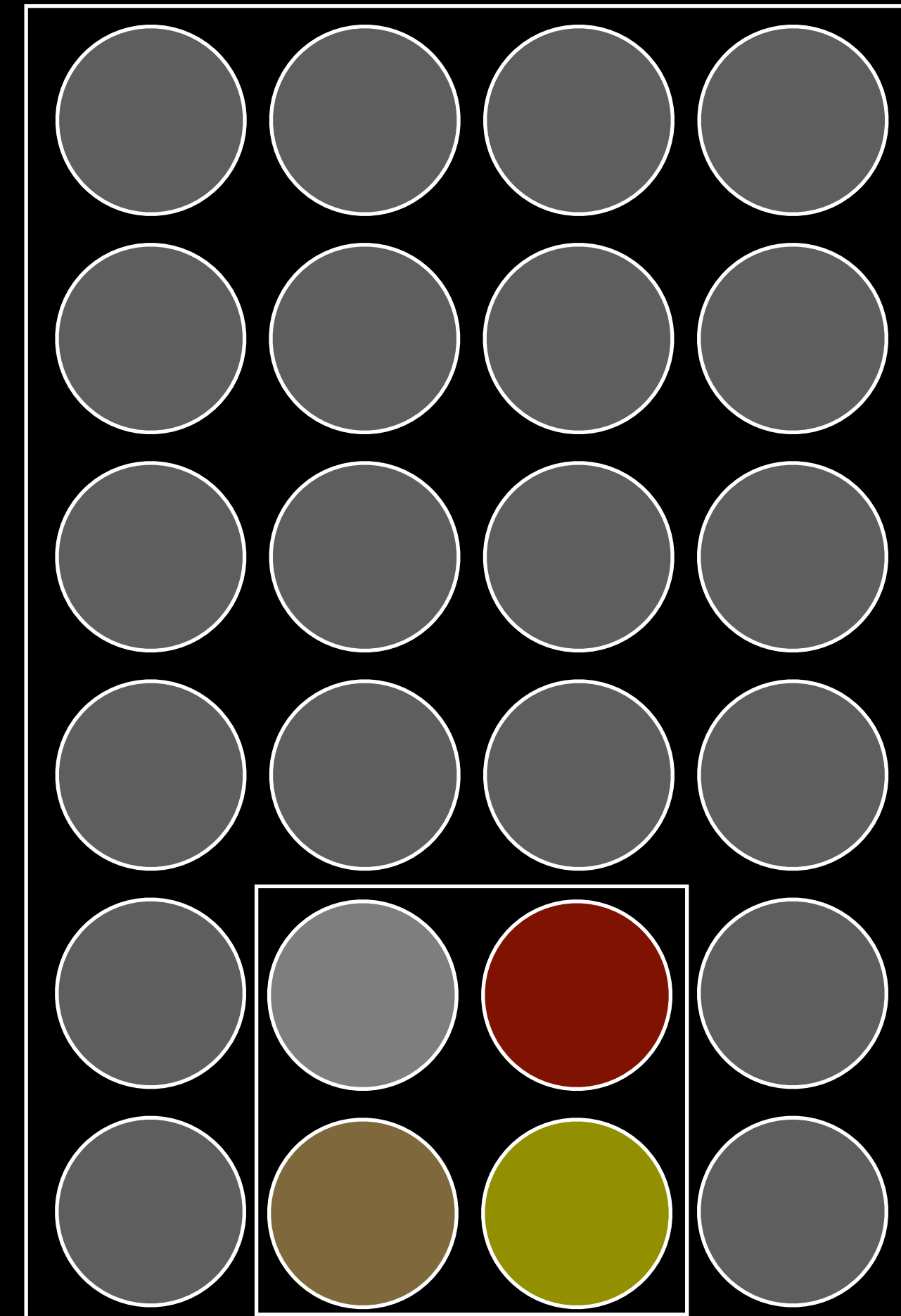


ARCHIVIERUNGSFEHLER

PROBLEM: FEHLERKATEGORISIERUNG



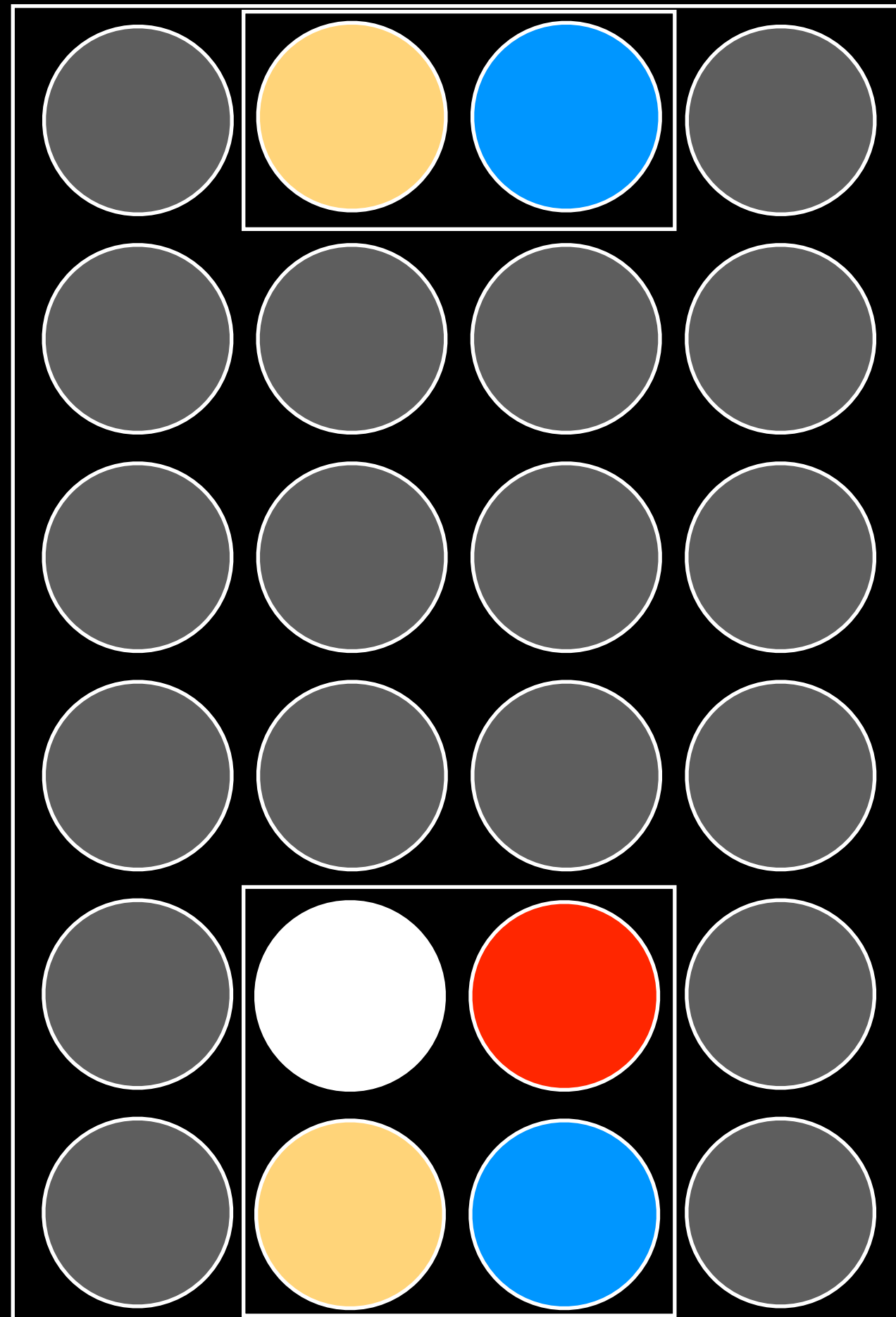
Screenshot Original



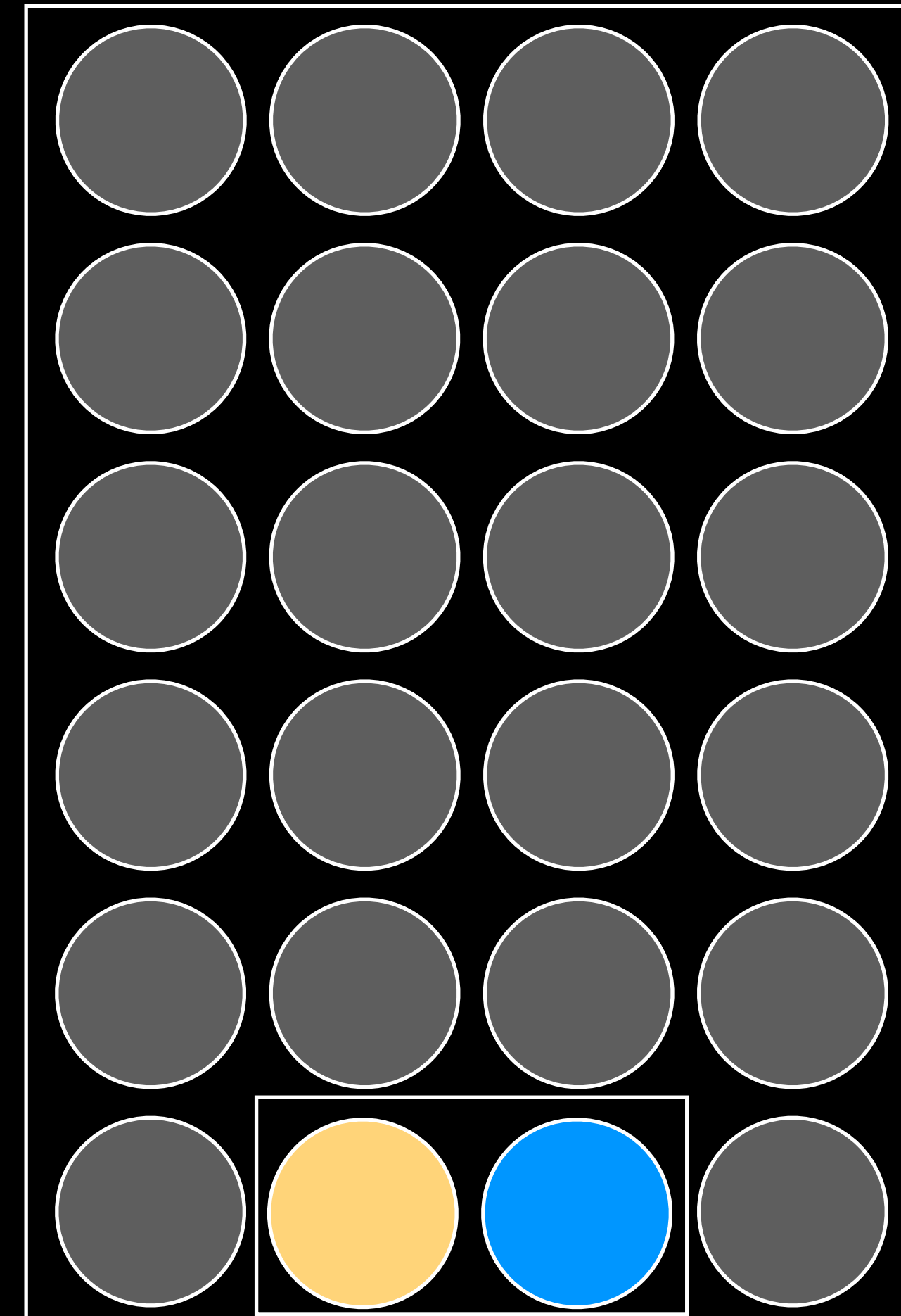
Screenshot Archiv

ARCHIVIERUNGSFEHLER

PROBLEM: ZUORDNUNG KORRESPONDIERENDER ELEMENTE



Screenshot Original



Screenshot Archiv

ARCHIVIERUNGSFEHLER

IMPLIKATIONEN

- ▶ Fehlerkategorisierung:
 - ▶ nicht alle Fehler sind eindeutig kategorisierbar
 - ▶ perfekte Zuordnung korrespondierender Elemente ist darum nicht möglich
- ▶ d.h. korrekte Rekonstruktion der originalen Webseite aus der archivierten Webseite sind nicht mit Sicherheit für alle Webseiten möglich
- ▶ es heißt aber nicht, dass nicht für einige Webseiten eine perfekte Rekonstruktion existieren kann
 - ▶ automatisierte Qualitätsbewertung möglich

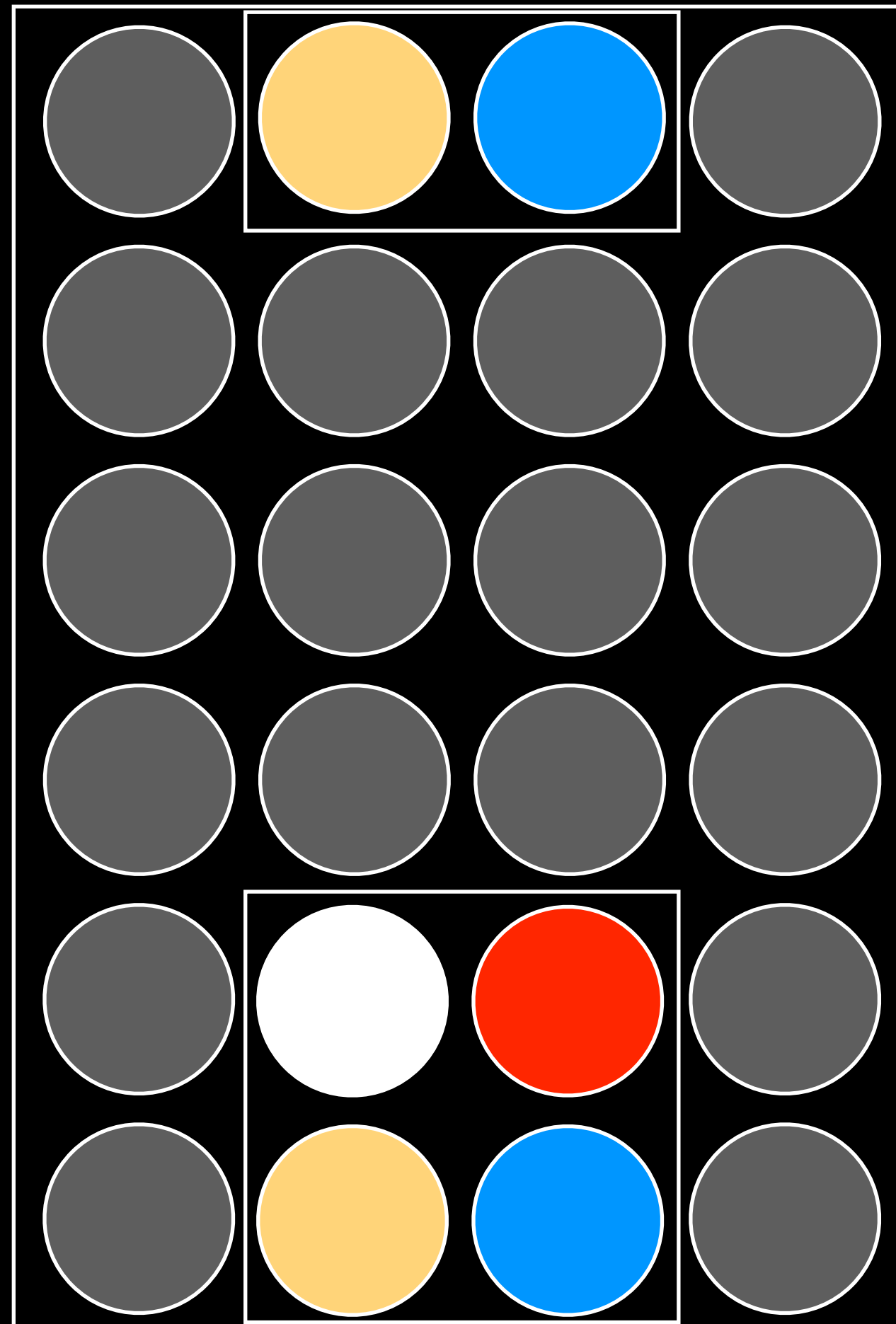
ARCHIVIERUNGSFEHLER

FEHLERMAß

- ▶ Pixel-zu-Pixel-Vergleich zwischen archiviertem und originalem Screenshot
 - ▶ angepasste Edit-Distanz nach Levenshtein:
 - ▶ **Insertion** = Pixel einfügen
 - ▶ **Deletion** = Pixel löschen
 - ▶ **Substitution** = Pixel durch einen Pixel in einer anderen Farbe ersetzen
- ▶ Fehler = minimale Anzahl der Pixel auf denen Operationen nötig sind um das Original aus dem Archivbild zu erzeugen

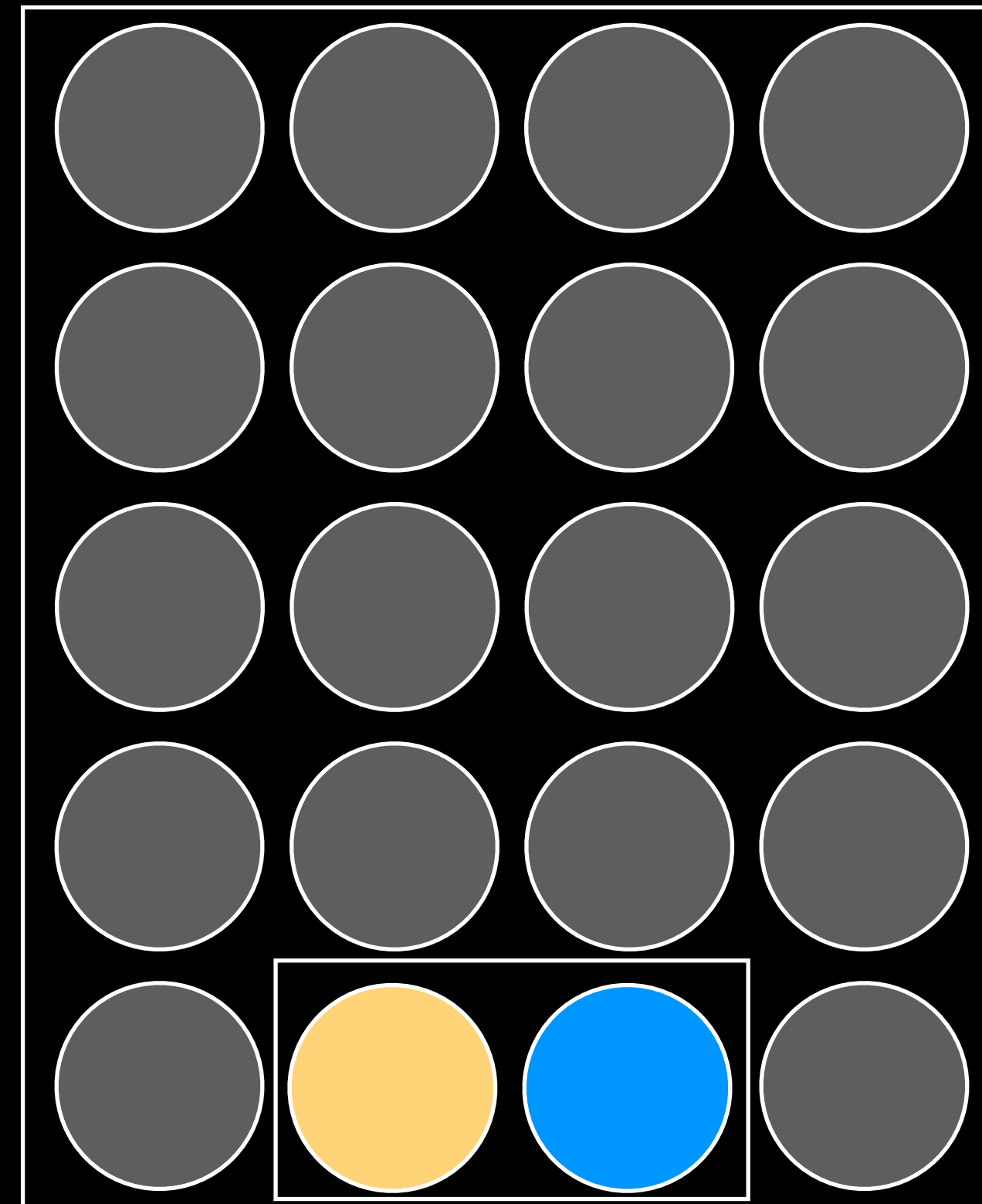
ARCHIVIERUNGSFEHLER

FEHLERMAß BEISPIEL



Screenshot Original

vs.

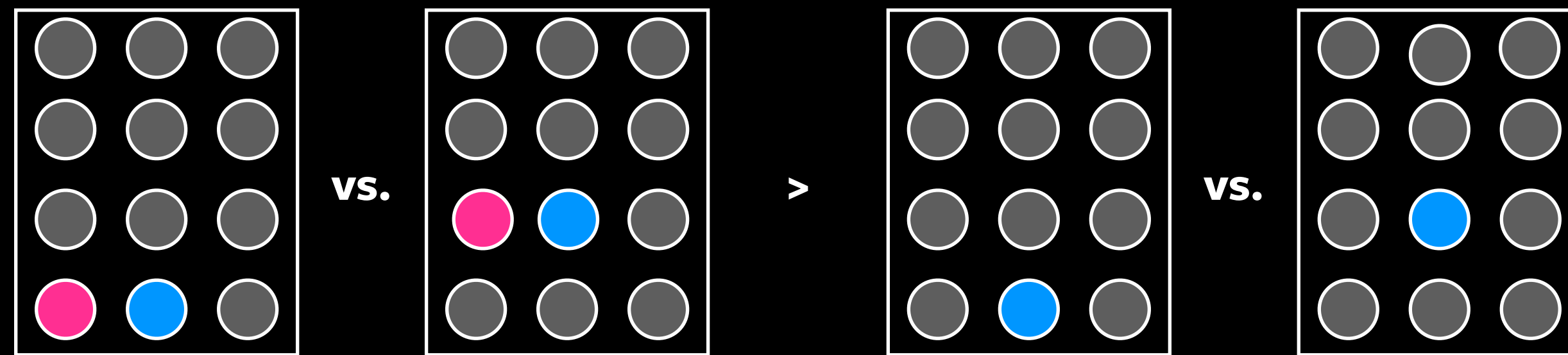


Screenshot Archiv

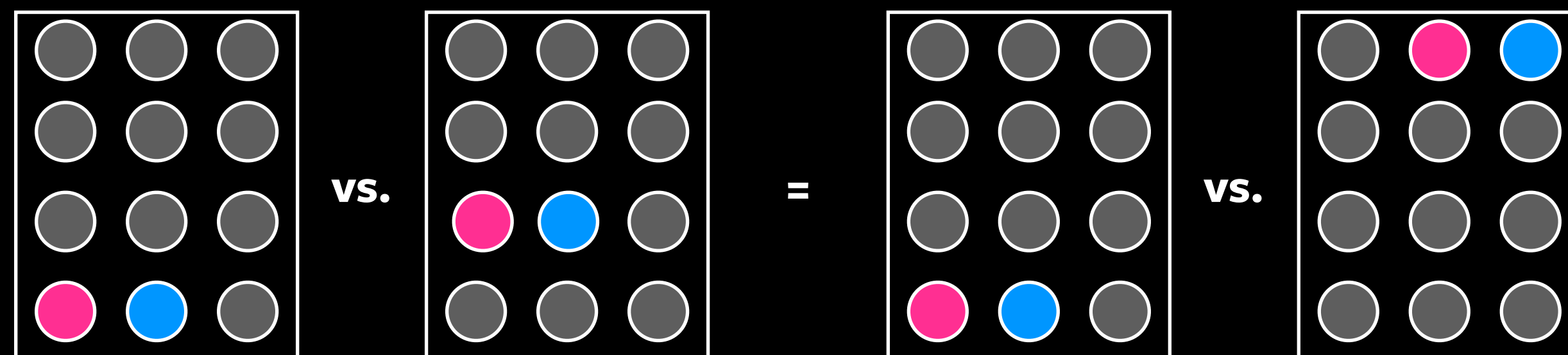
= 4 Insertion + 2 Substitution = 6 Px

ARCHIVIERUNGSFEHLER

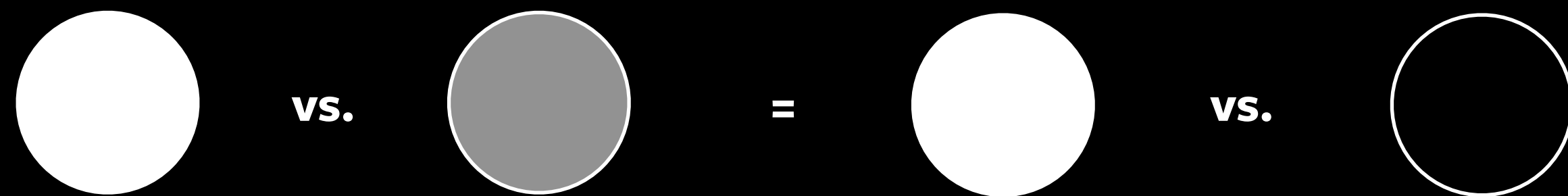
FEHLERMAß KRITIK



► Fehler abhängig von Elementgröße



► unterschiedliche Verschiebungslängen resultieren in gleich große Fehler

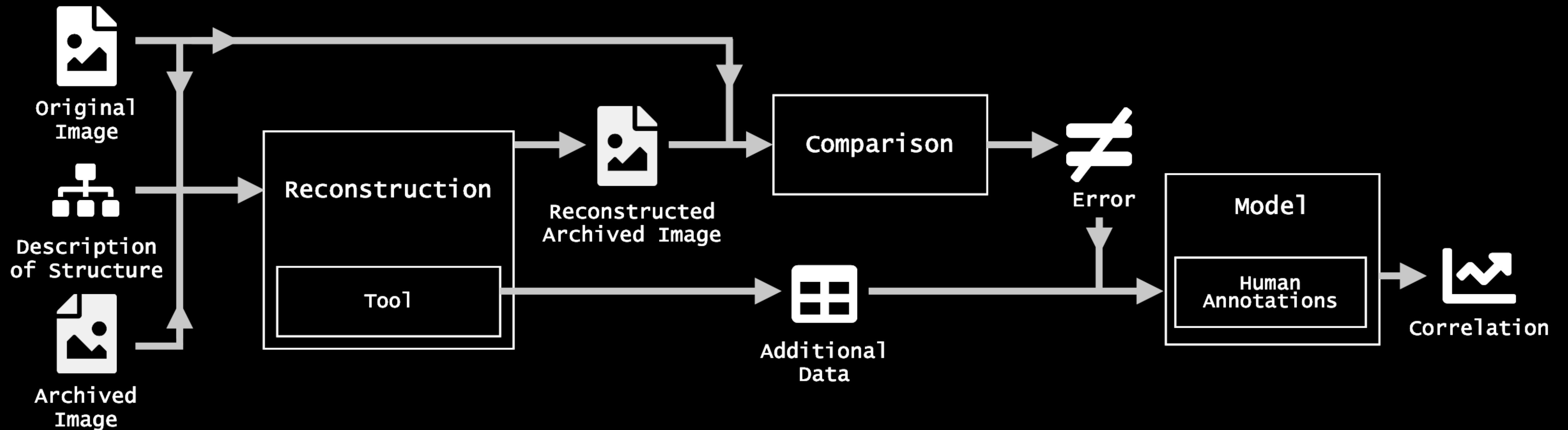


► keine Unterscheidung zwischen kleinen und großen Farbabweichungen

GENERELLES FRAMEWORK

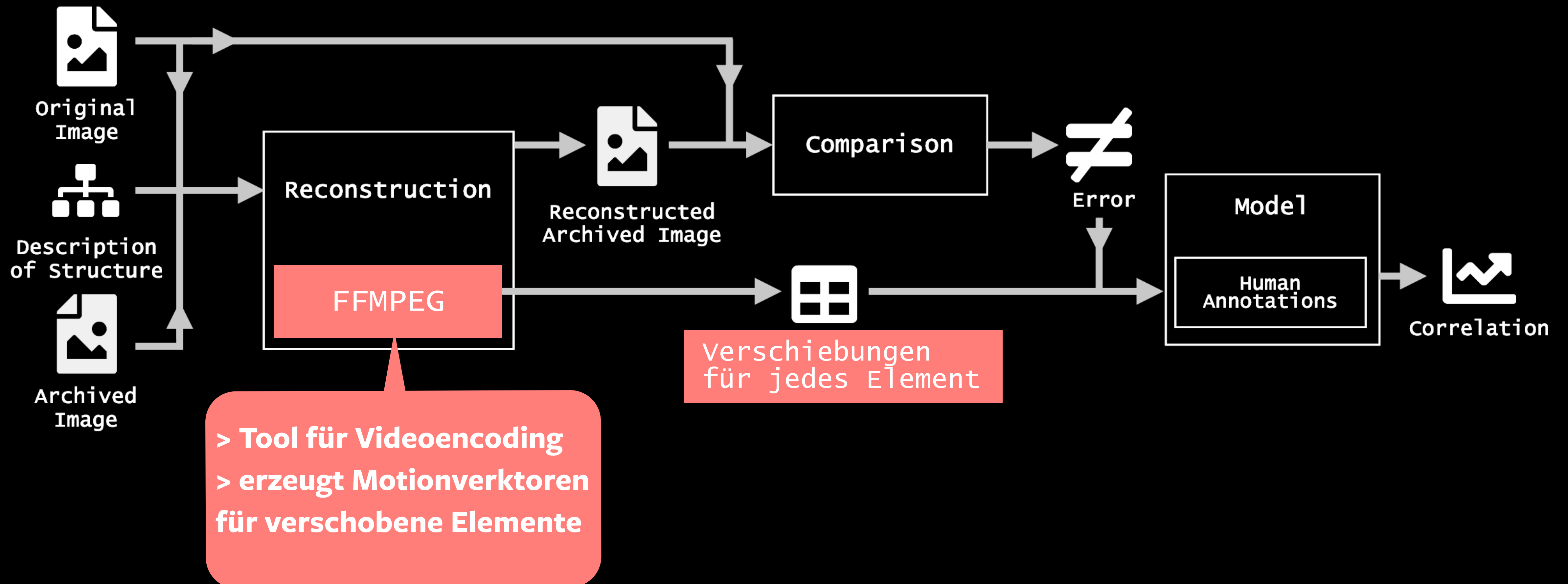
ENTWICKLUNG EINER AUTOMATISIERTEN QUALITÄTSBEWERTUNG

► Ablauf



UMSETZUNG DES FRAMEWORKS

VERSCHIEBUNGEN REDUZIEREN



UMSETZUNG DES FRAMEWORKS

VERSCHIEBUNGEN REDUZIEREN



originale Seite



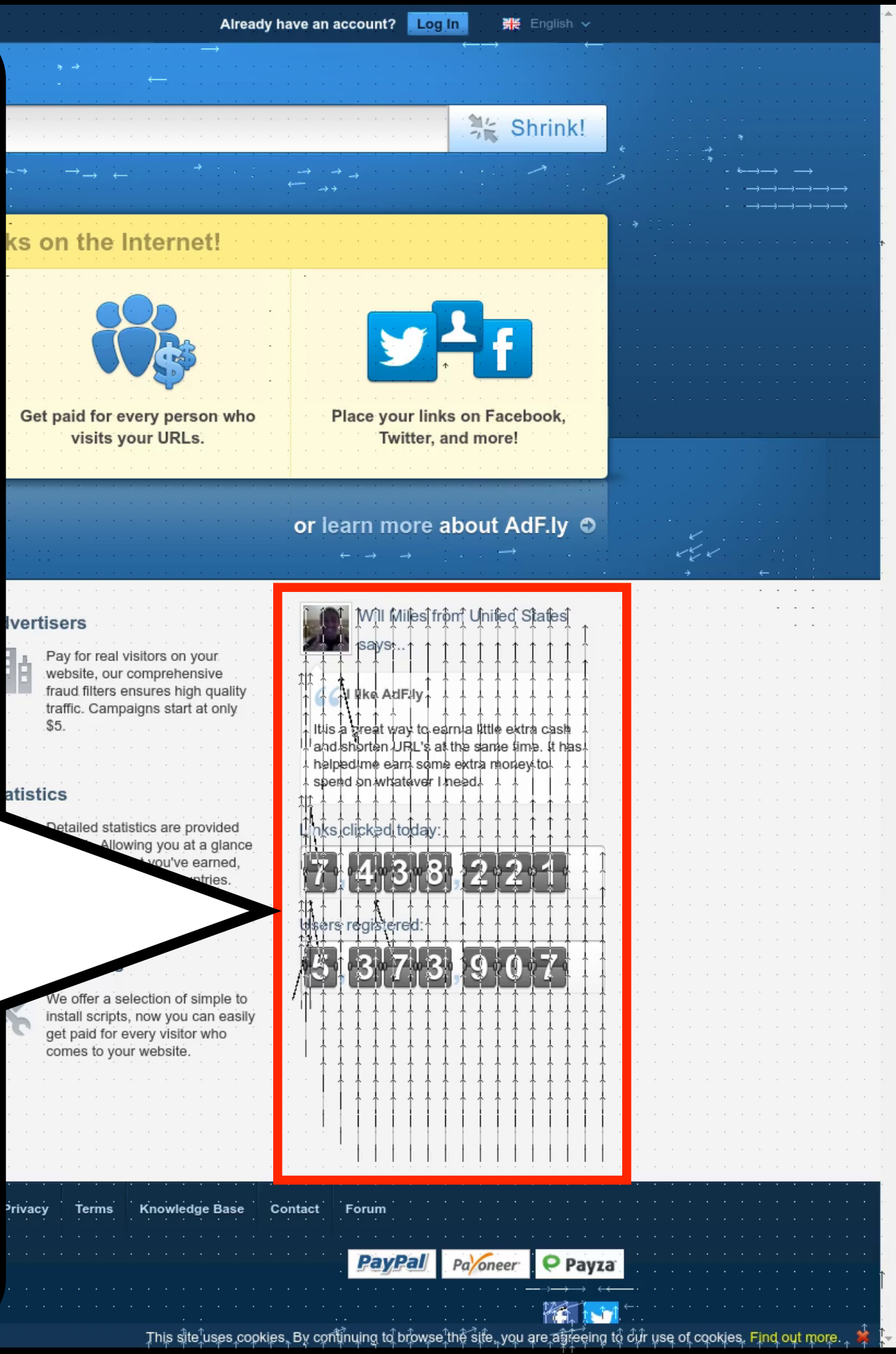
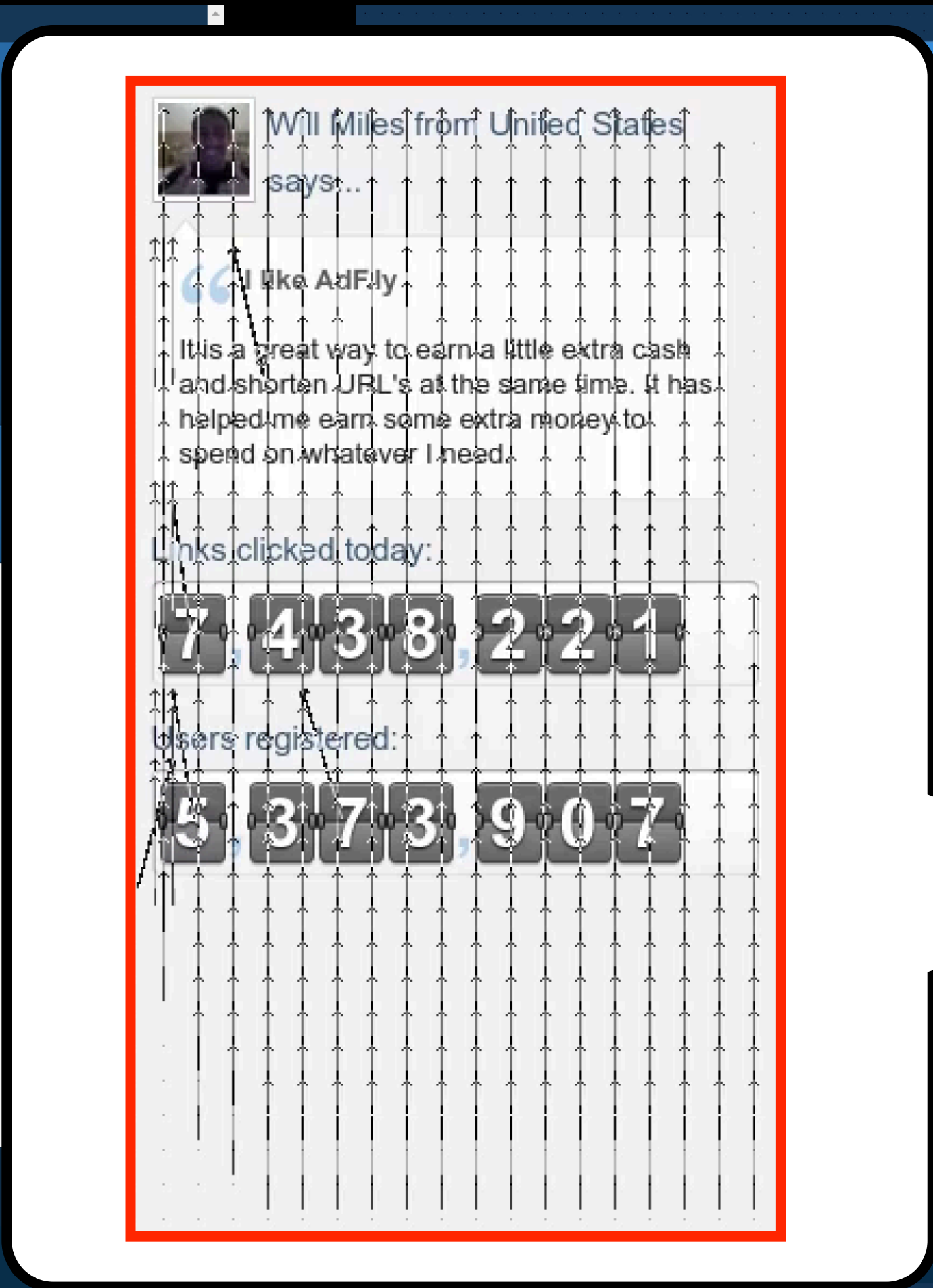
archivierte Seite

UMSETZUNG DES FRAMEWORKS

VERSCHIEBUNGEN REDUZIEREN: MOTIONVEKTOREN



originale Seite



archivierte Seite mit Motionvektoren

UMSETZUNG DES FRAMEWORKS

VERSCHIEBUNGEN REDUZIEREN



originale Seite



archivierte Seite

UMSETZUNG DES FRAMEWORKS

VERSCHIEBUNGEN REDUZIEREN + FEHLENDE ELEMENTE FINDEN?



originale Seite



rekonstruierte archivierte Seite

UMSETZUNG DES FRAMEWORKS

VERSCHIEBUNGEN REDUZIEREN + FEHLENDE ELEMENTE FINDEN?



originale Seite



rekonstruierte archivierte Seite

UMSETZUNG DES FRAMEWORKS

VERSCHIEBUNGEN REDUZIEREN + FEHLENDE ELEMENTE FINDEN?



originale Seite



rekonstruierte archivierte Seite

UMSETZUNG DES FRAMEWORKS

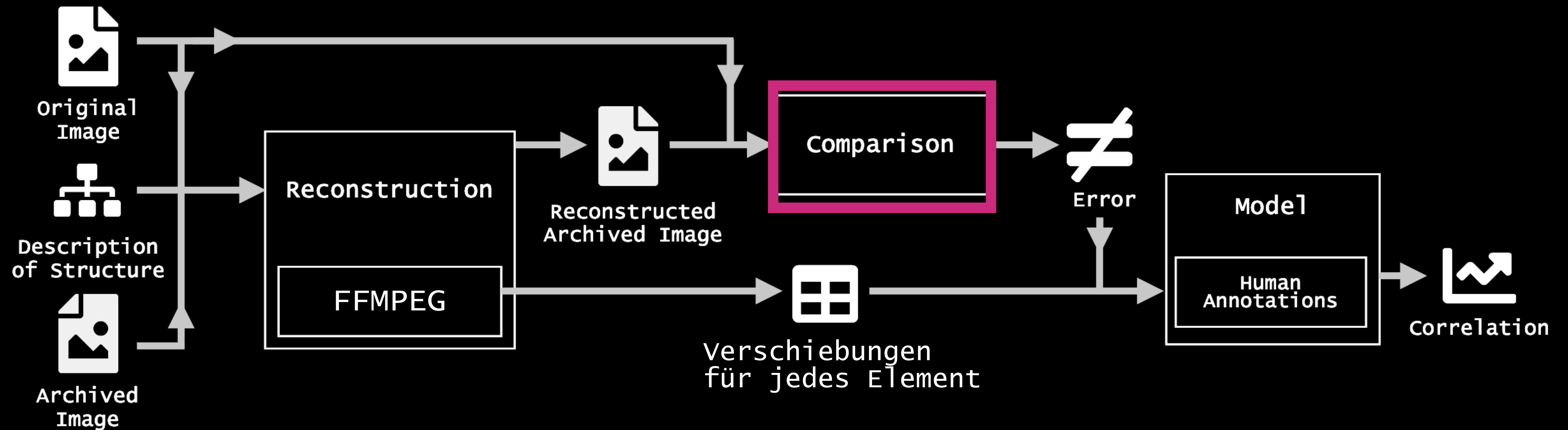
VERSCHIEBUNGEN REDUZIEREN + FEHLENDE ELEMENTE FINDEN?

► **PROBLEM:**

- Lücken der verschobenen Elemente korrespondieren nicht immer mit den fehlenden Elementen im Original

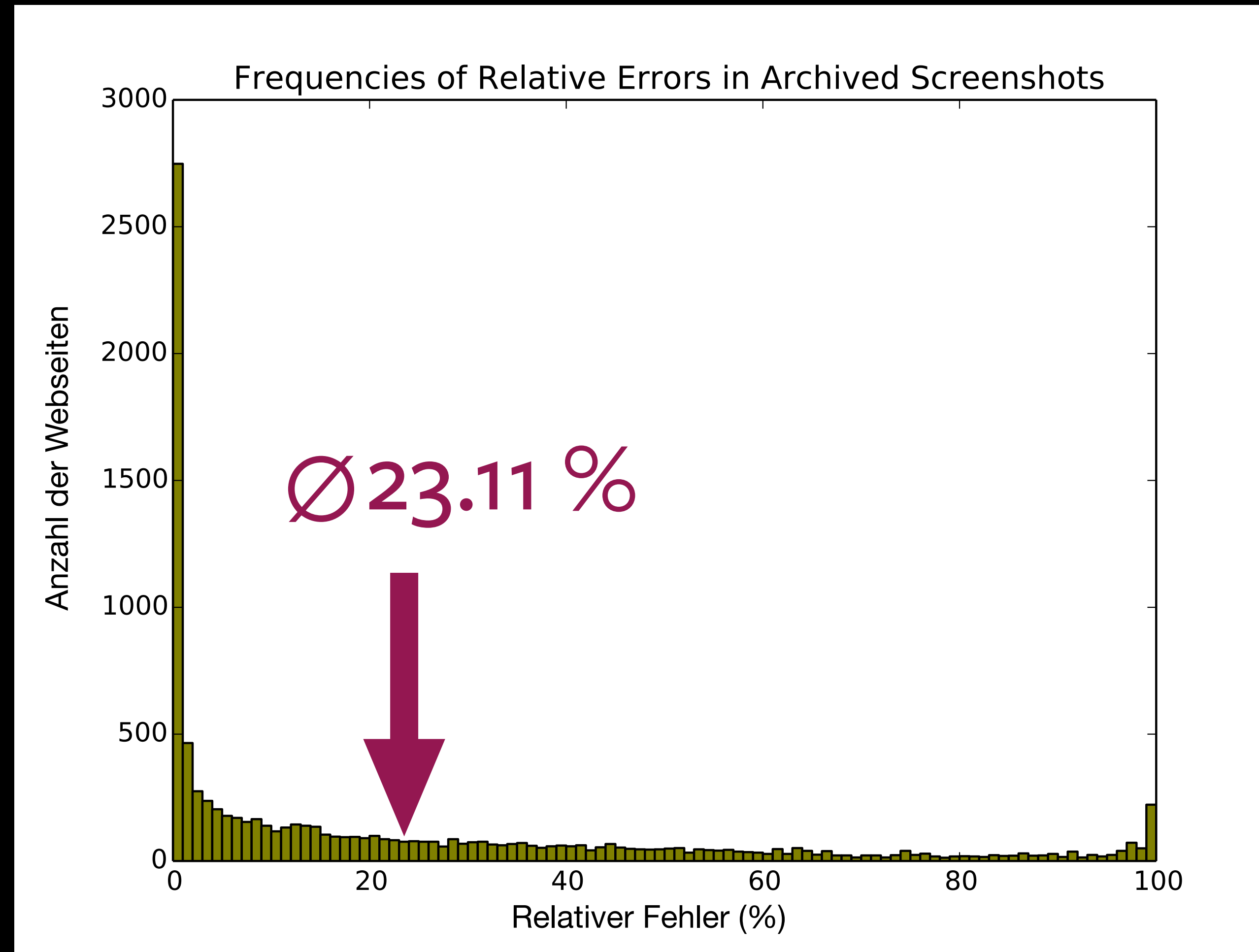
UMSETZUNG DES FRAMEWORKS

VERSCHIEBUNGEN REDUZIEREN



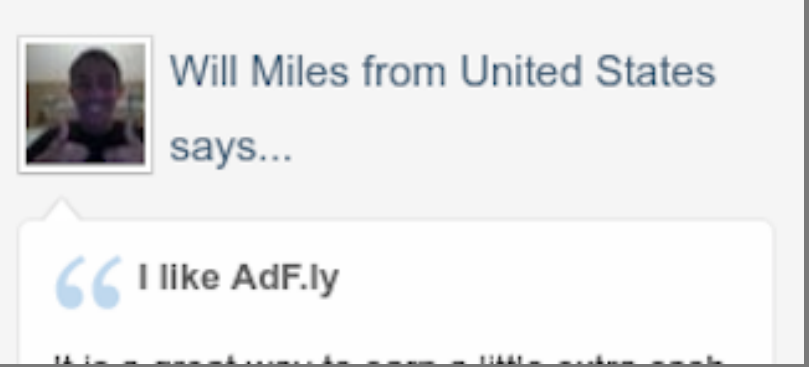
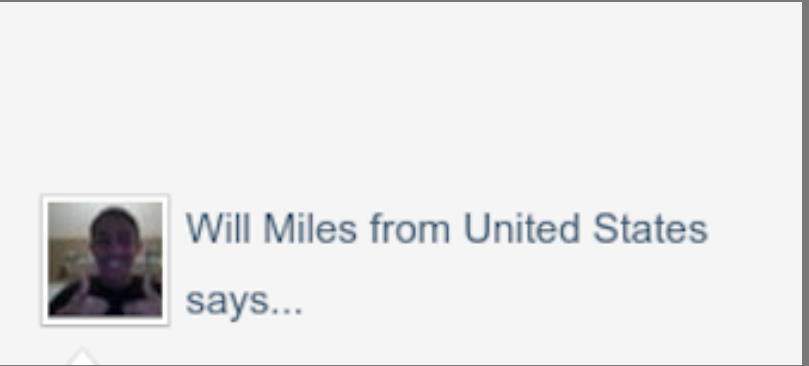
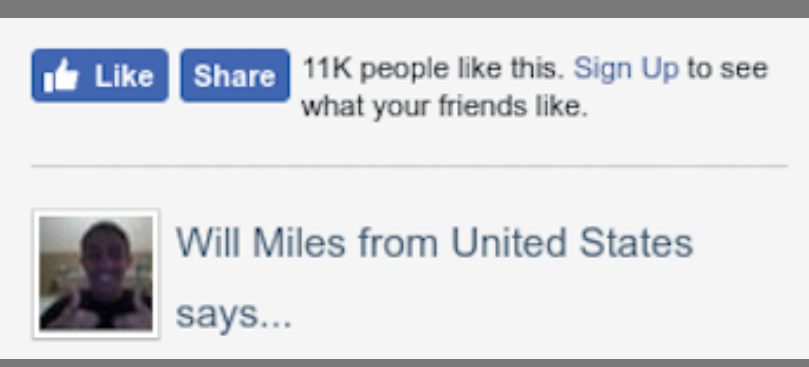
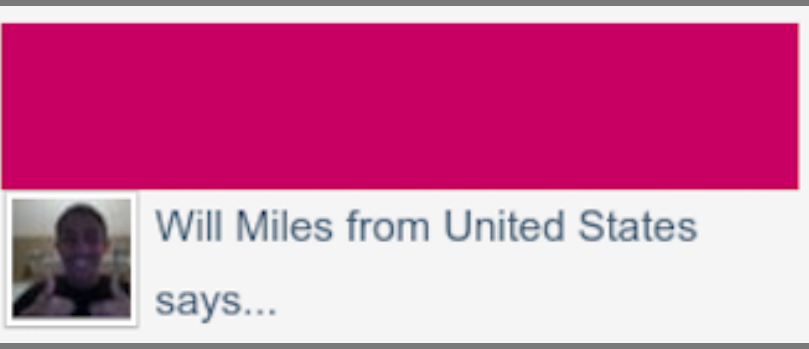
UMSETZUNG DES FRAMEWORKS

VERSCHIEBUNGEN REDUZIEREN: FEHLER DER ARCHIVSEITEN



UMSETZUNG DES FRAMEWORKS

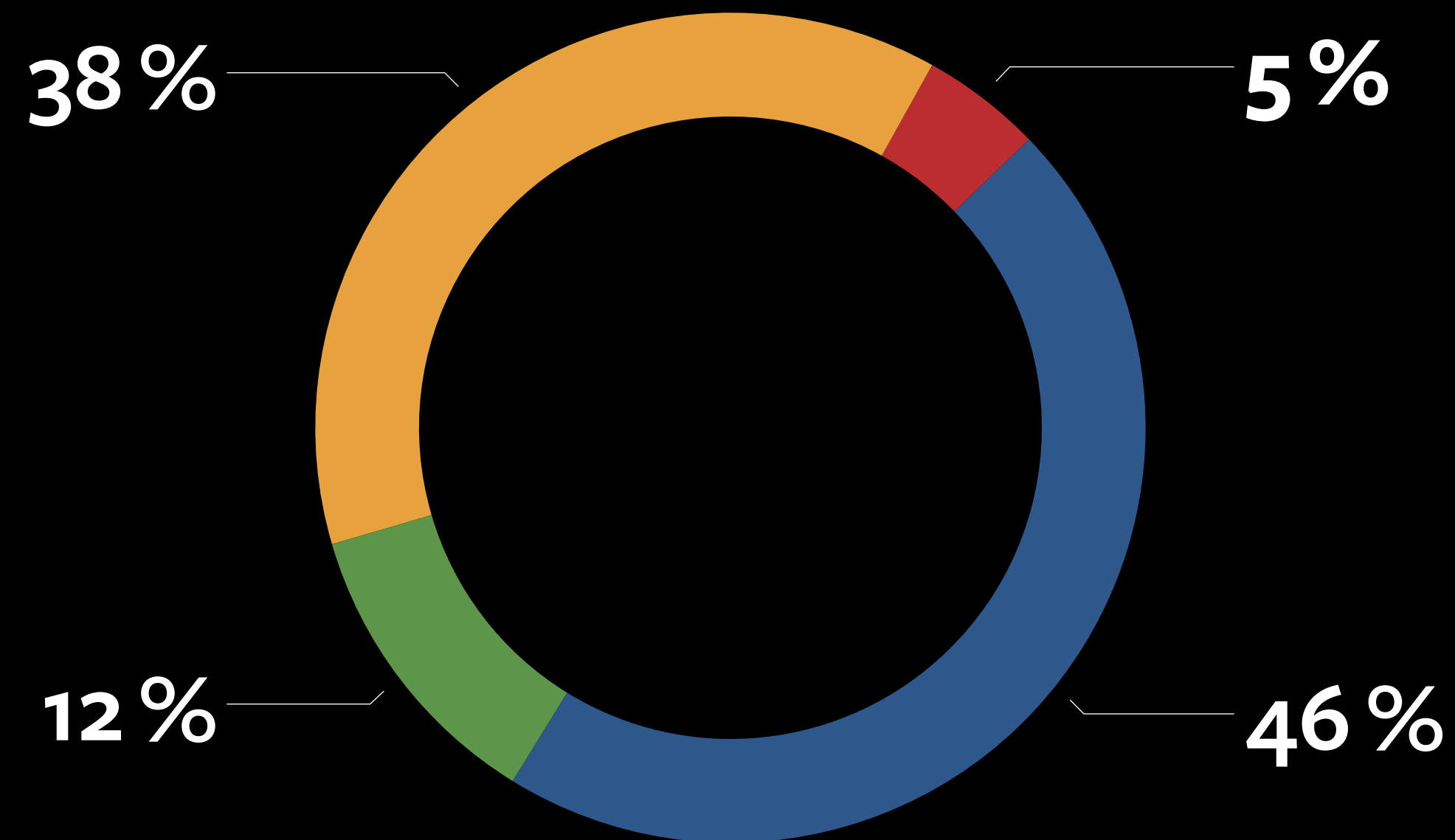
VERSCHIEBUNGEN REDUZIEREN: FEHLER DER REKONSTRUKTION

METHODE		AVG RELATIVER FEHLER	VERBESSERUNG
KEINE		23.11 %	-
HÄUFIGSTE FARBE IM ARCHIVBILD		21.07 %	2,04 %
KOPIE AUS ORIGINAL (UPPER BOUND)		20.85 %	2,26 %
VERSCHIEDEN VON ORIGINAL (LOWER BOUND)		21.34 %	1,77 %

UMSETZUNG DES FRAMEWORKS

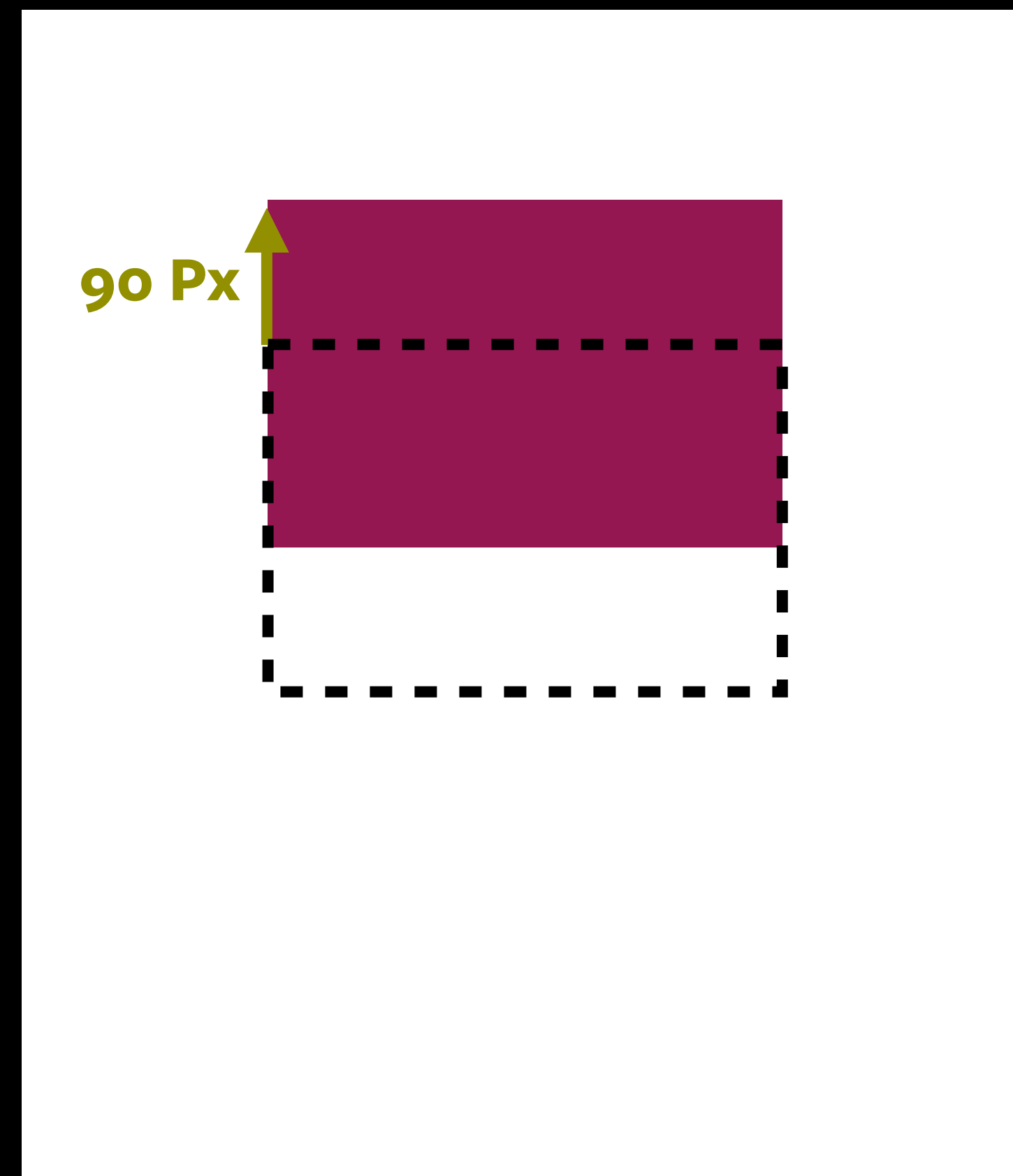
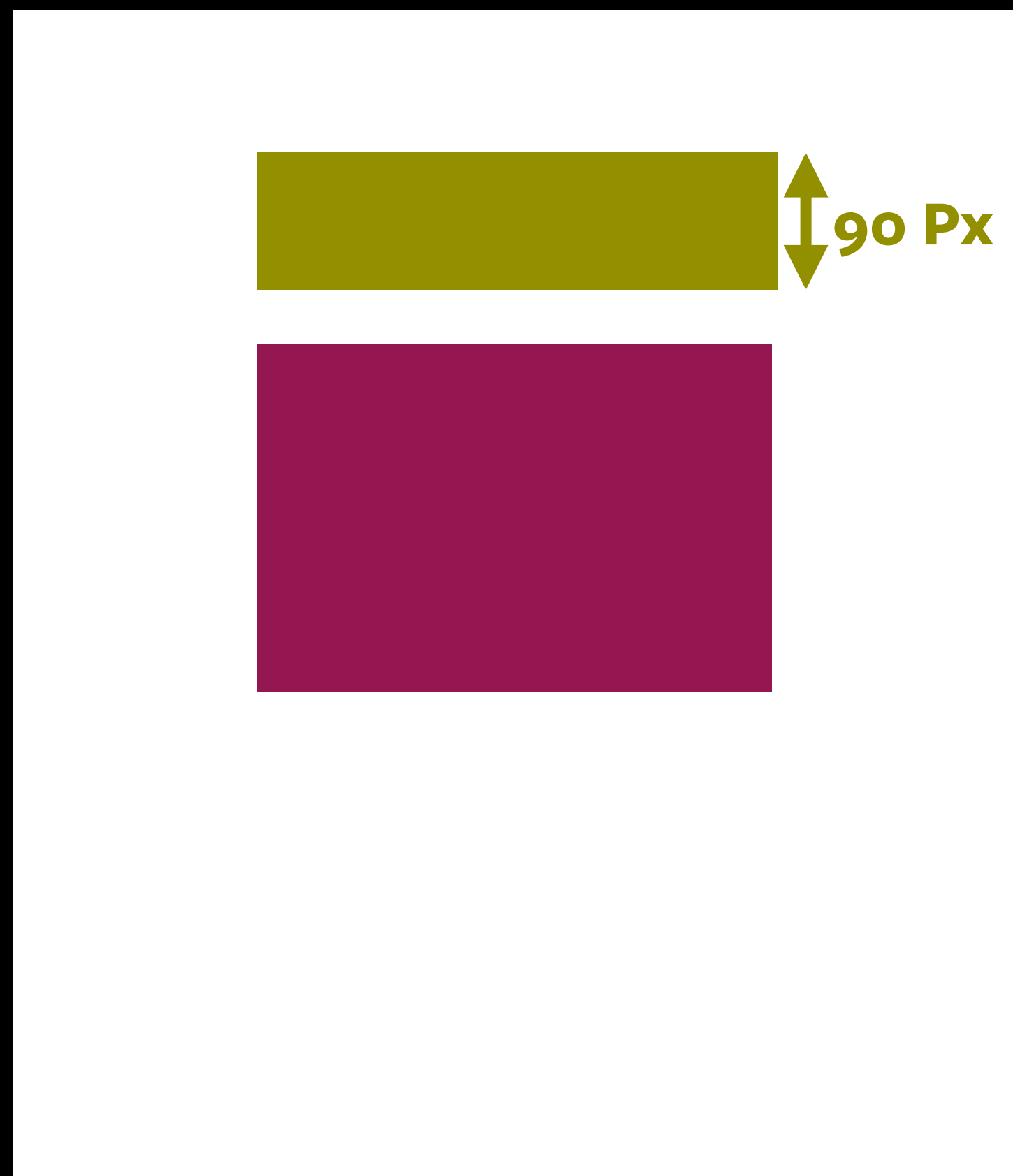
ANALYSE DER VERSCHIEBUNGEN NACH RICHTUNG

● to top ● to bottom ● to left ● to right



UMSETZUNG DES FRAMEWORKS

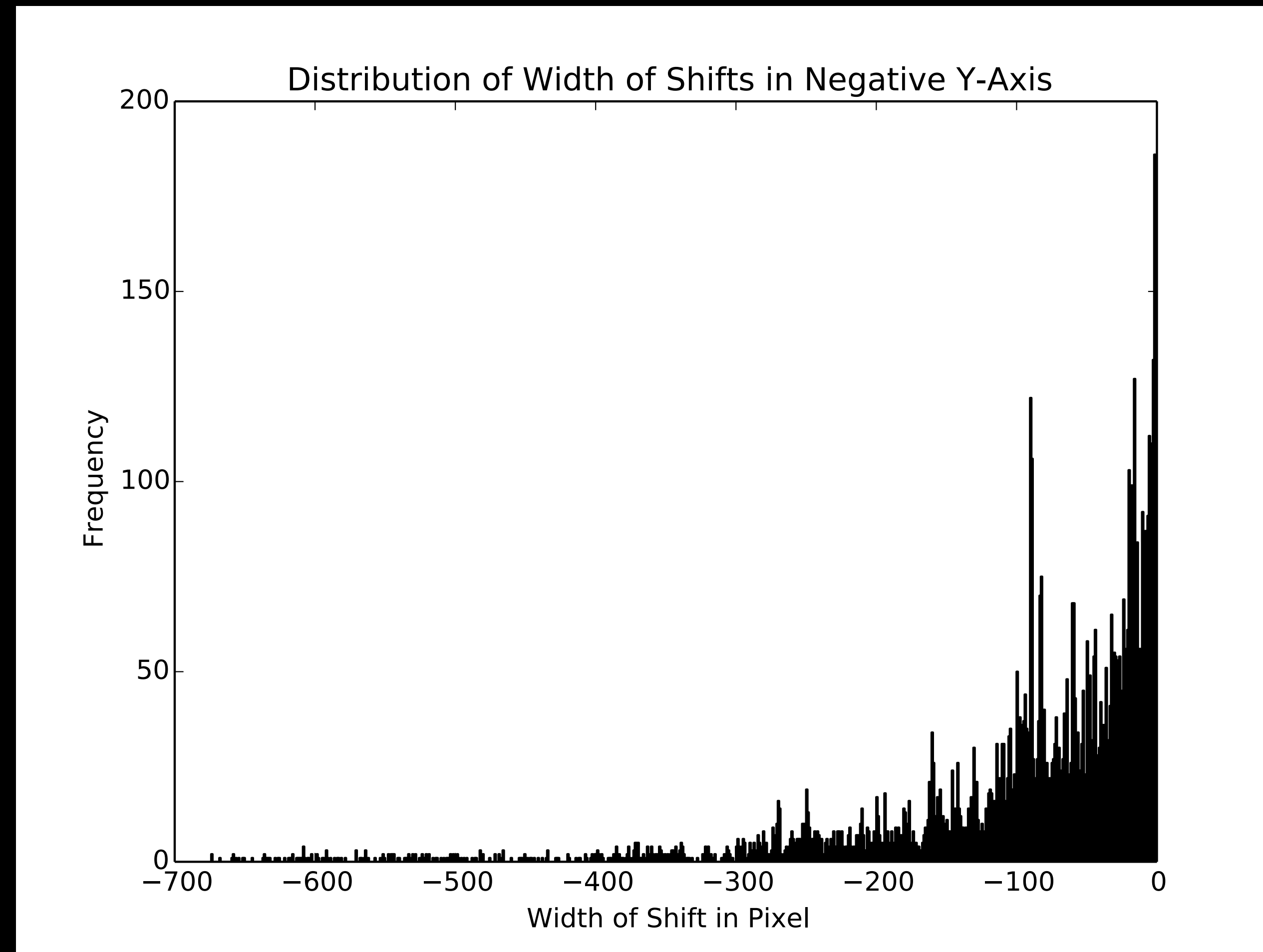
ANALYSE DER VERSCHIEBUNGEN NACH OBEN



- Zusammenhang zwischen Verschiebung und fehlendem Element ?

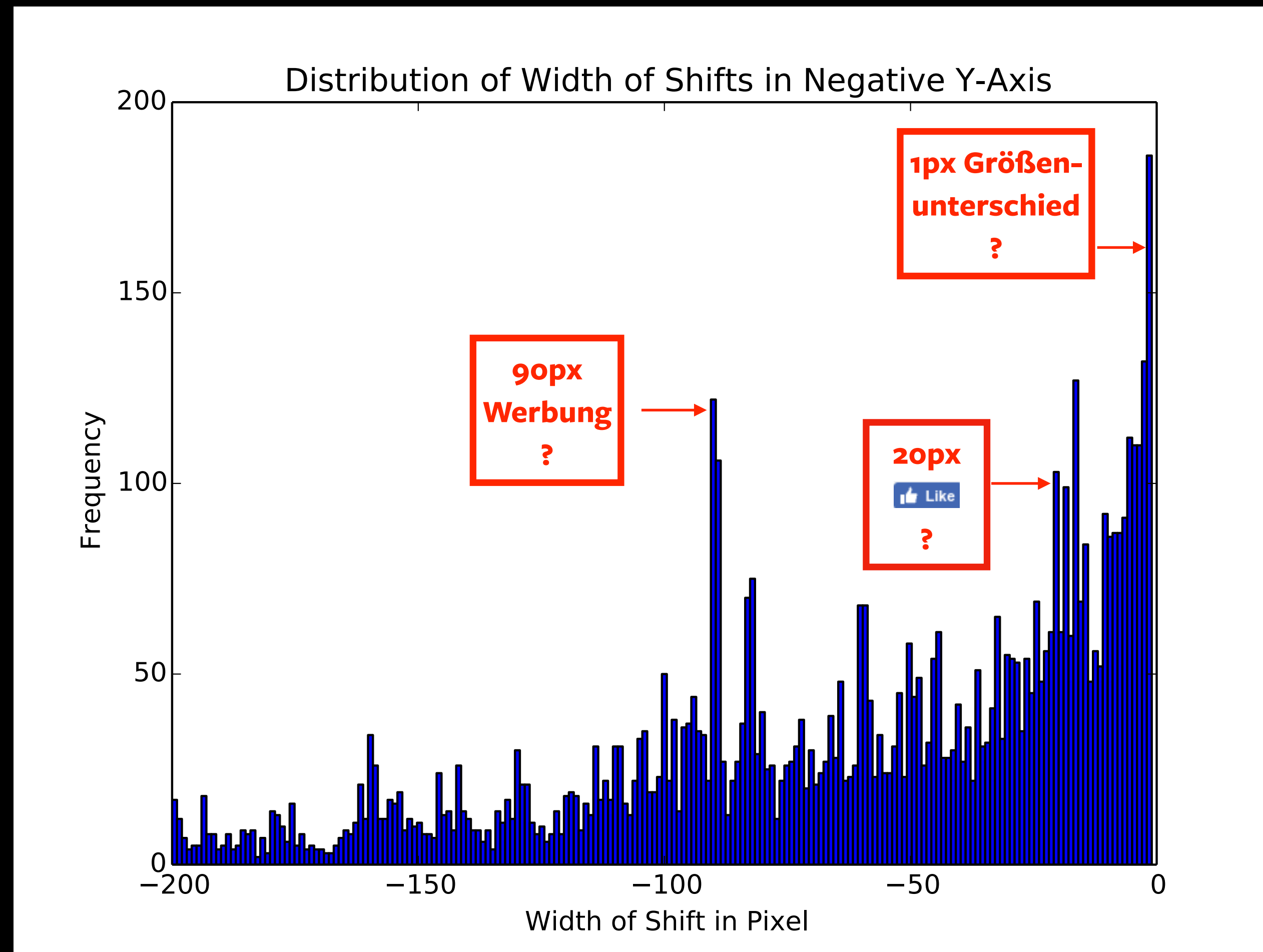
UMSETZUNG DES FRAMEWORKS

ANALYSE DER VERSCHIEBUNGEN NACH OBEN: VERSCHIEBUNGSWEITEN



UMSETZUNG DES FRAMEWORKS

ANALYSE DER VERSCHIEBUNGEN NACH OBEN: VERSCHIEBUNGSWEITEN



OFFENE FRAGEN

- ▶ weiter mit:
 - ▶ Suchen fehlender Elemente mit bestimmter Größe
 - ▶ Lineare Regression mit Human Annotations
 - ▶ neuer Threshold für Motionvektoren
 - ▶ anderer Codec, z.B. H.265 statt H.264
 - ▶ ...

ZUSAMMENFASSUNG

- ▶ verschiedene **Fehlertypen**
 - ▶ Grenzen automatischer Fehlerbehebung:
1.Fehler korrekt kategorisieren => 2. Elemente korrekt zuordnen
- ▶ angepasste Edit-Distanz als **Fehlermaß**
- ▶ **Framework** zur Entwicklung automatischer Qualitätsbewertung von Webarchiven
 - ▶ Rekonstruktion des Originalbildes, Vergleich, Korrelation
- ▶ Versuch: **Verschiebungsfehler** durch Videokompression beheben
 - ▶ nebenbei: **Fehlende Elemente?**
 - ▶ Problem: Lücken korrespondieren nicht mit fehlendem Element
 - ▶ anderer Ansatz: Verschiebungsweite als Indiz für Größe eines fehlenden Elements

QUELLEN

- ▶ The Internet Archive. Wayback machine. <https://webarchive.org>
- ▶ Johannes Kiesel, Florian Kneist, Milad Alshomary, Benno Stein, Matthias Hagen, and Martin Potthast. Reproducible Web Corpora: Interactive Archiving with Automatic Quality Assessment. Journal of Data and Information Quality (JDIQ), 10(4):17:1–17:25, October 2018
- ▶ Vladimir Iosifovich Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady, 10(8):707–710, 1966. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965