(h[s])®

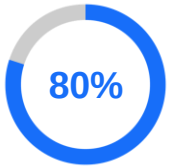
Hyperscience Platform

- 지능형 문서 처리(IDP-Intelligent Document Processing) -

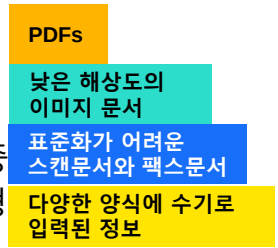
현황

Document processing is step zero

We live in a human-readable world.

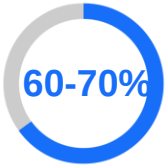


기업 데이터중 80%가 비정형



$8 Billion

기업은 고객 및 기타 조직에서 생성된 데이터를 추출하고 사용하기 위해 매년 800 억 달러 이상을 지출



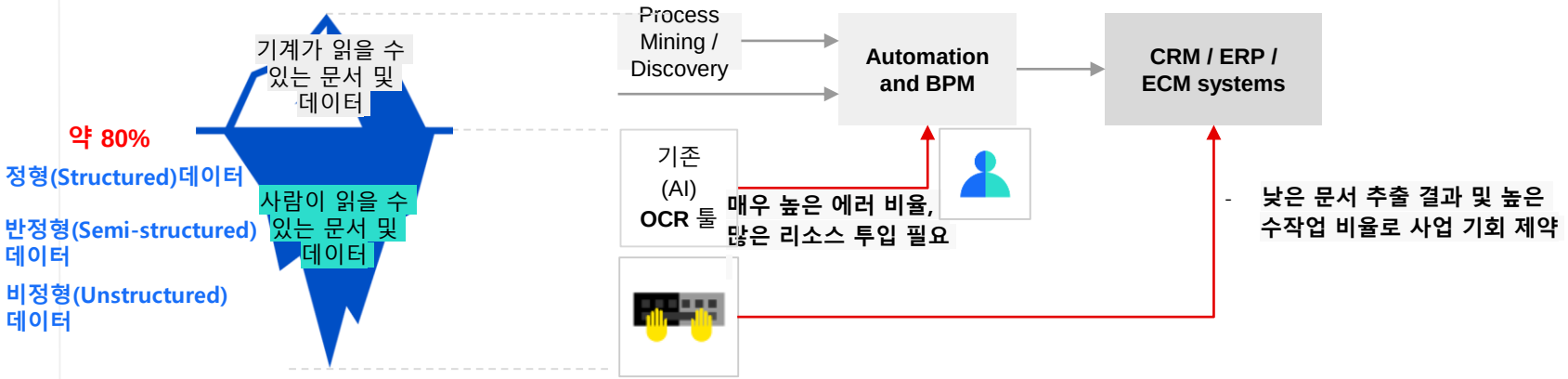
기업 내 모든 데이터의 70%가 분석에 사용되지 않음

7.5 Billion

매년 75억개의 고유한 문서와 15조 개의 사본이 생성됨

*Deloitte executive survey, 2020.

직면한 문제



문제 #1 수동 입력 한계

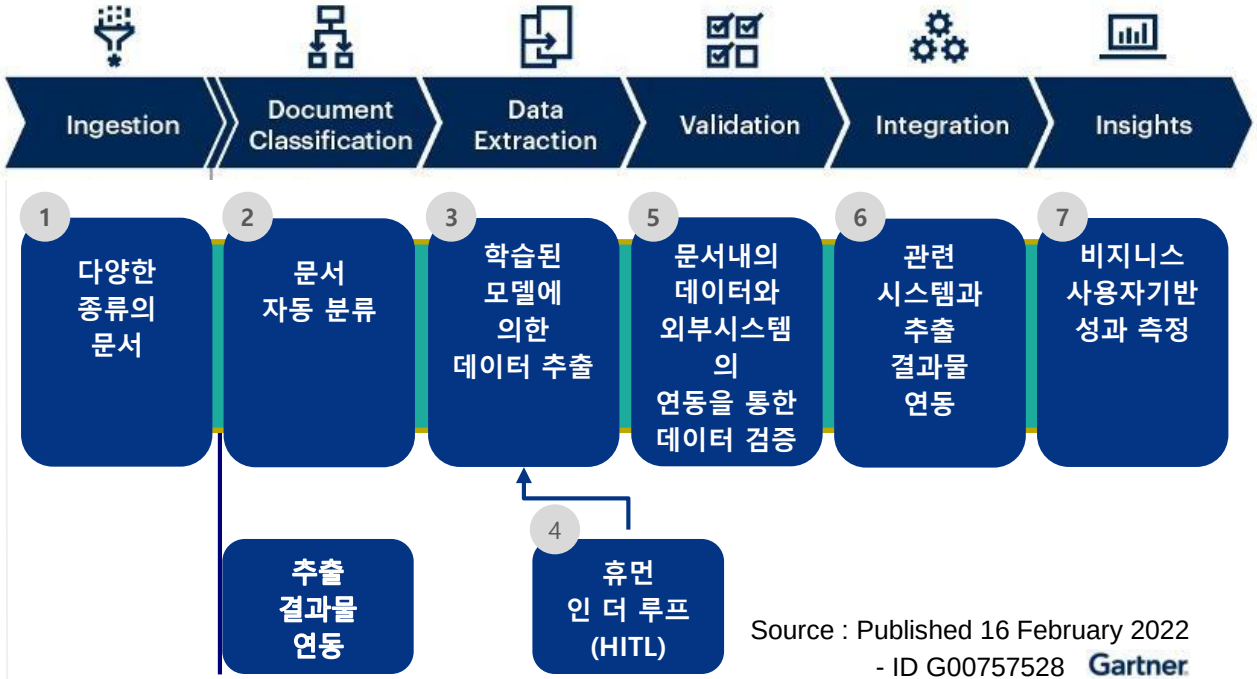
- 높은 운영 비용
- 리스크 증가
- 비효율적인 인력 활용
- 낮은 고객 만족도
- 기회 비용 손실

문제 #2 (AI) OCR 한계

- 수기 인식 불가
- 고해상도 이미지 필요
- 모호한 문서 내용
- 낮은 신뢰성, 수동 검토 및 검증 필수
- 성능 개선 한계
- 자동화 한계

지능형 문서 처리(IDP)의 정의 by 가트너

기본적인 IDP 워크플로우의 단계



Source : Published 16 February 2022
- ID G00757528 Gartner

목적

문서의 분석 및 워크플로우 자동화를 위해 주로 출력물과 수기 작성 문서에서 텍스트 데이터 정보를 추출하고 이전에 처리한 적이 없는 문서 형식에도 쉽게 적응하고 다양한 모델을 통해 데이터를 추출

핵심

-모든 유형의 정형 및 반정형 문서 처리
-다국어 지원
-엔드 투 엔드 프로세스 워크플로우 자동화
-사용 용이성(로우 코드/노코드 플랫폼)

차별화

-수기 작성 문서 및 서명을 포함한 반정형 문서 처리
-산업군 또는 도메인별 사전 훈련된 솔루션
-트레이닝 데이터 요구 사항을 최소화하는 소규모 데이터 접근 방식
-Human-in-Loop를 통한 데이터 검증
-개인 식별 정보(PII) 및 민감한 데이터 수정

하이퍼사이언스

지능형 문서인식(IDP: Intelligent Document Processing) 시장의 리더

$300M

3억불 투자 유치

300+

글로벌 임직원

뉴욕, 런던, 소피아, 베를린, 뮌헨, 토론토, 호주, 인도, 한국

주요 고객군

공공기관, 금융(은행, 생명보험, 화재보험외), BPO, 의료, 제조(공정 디지털라이제이션등)등

지원 언어

한국어, 영어, 스페인어, 독일어, 프랑스어, 이태리어, 포르투칼어, 아랍어, 일본어

투자사

기술 파트너

주요 고객

Everest Group

" 하이퍼사이언스는 빠른 성장을 계속해 왔으며 사용자가 특정 사용 사례에 맞는 맞춤형 워크플로우를 만들 수 있는 모듈화된 아키텍처를 포함하여 IDP 플랫폼을 크게 개선했습니다. 직관적인 사용자 인터페이스, 손으로 쓴 텍스트를 처리할 수 있는 기능, 분석 및 보고 기능, 독특한 로드맵도 성공에 긍정적으로 기여했습니다"

.Ashwin Gopakumar, Practice Director, Everest Group

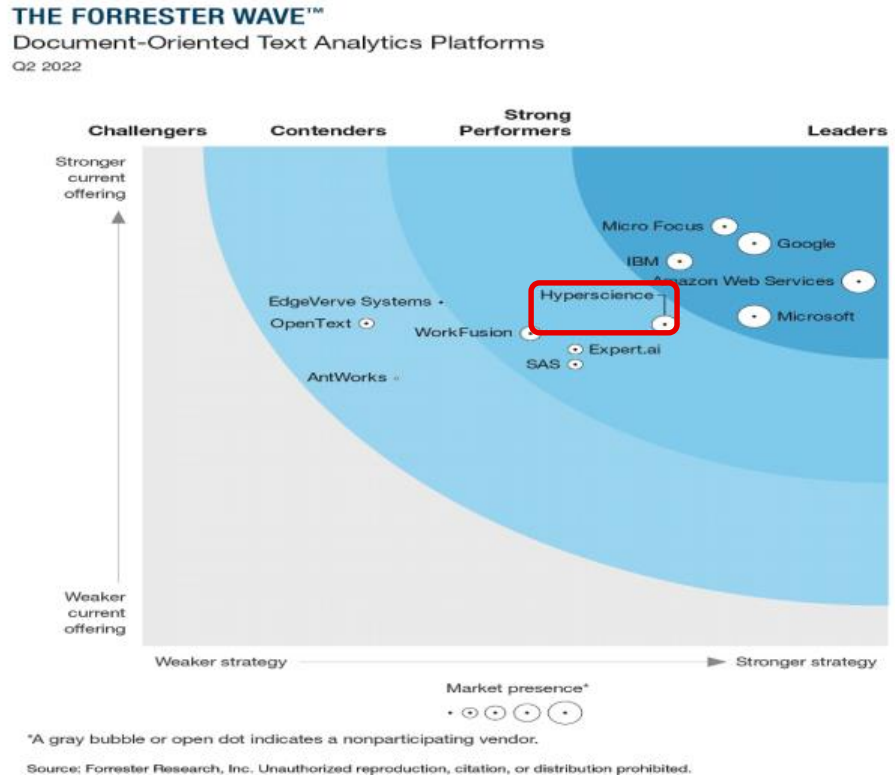
Download full report here.

THE FORRESTER WAVE

"하이퍼사이언스는 이번 Wave에서 전체 고객 기준 최고 점수를 받았습니다. 고객은 문서 캡처와 같은 기능을 제공하고 다양한 문서 유형 및 양식에 대해 사전 훈련된 다양한 ML 모델을 포함하여 모든 텍스트 분석 사용 사례를 처리하는 하이퍼사이언스의 능력을 높이 평가했습니다. 또한 플랫폼 보안, 확장성, DevOps 및 공급업체 전문 서비스에 대해서도 언급했습니다.."

Download full report here.

Also read the Forrester Consulting's Total Economic Impact™ (TEI) study commissioned by Hyperscience



주요 선택 이유

검증된 정확성

고객이 목표 정확성을 결정하고 입력. 고객의 복잡한 문서를 학습하고 목표 정확도 수준을 달성하기 위해 인간과 기계에 작업요청을 최적화함.
하이퍼사이언스는 실수를 하지 않기 위해 인간에게 더 빨리 도움을 요청하며 QA(Quality Assurance)시스템 내장

포춘 500 금융 서비스 회사에서 99.5% 필드레벨 정확성 확보

빠른 적용

고객 워크플로우 요구 사항에 맞게 사용자 정의된 사용하기 쉽고 매우 유연한
플랫폼

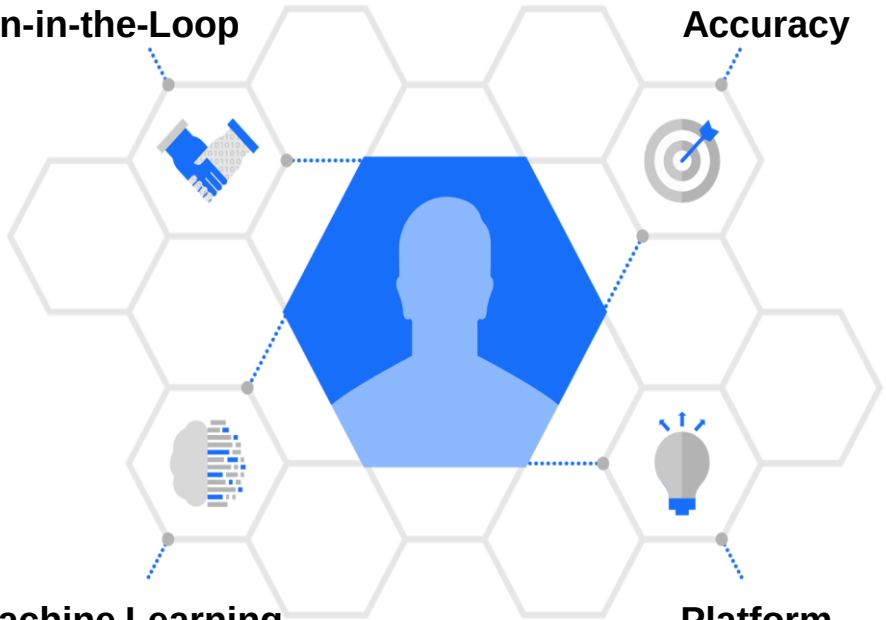
미국 연방 정부 기관에 2개월내 적용

비용 감소

동적으로 트리거되는 휴먼인더루프(Human-in-the-Loop) 및 정확성을 희생하지 않는 자동화

포춘 500 보험회사에서 수작업 검사 91% 감소

하이퍼사이언스 플랫폼 개요



Hyperscience is transforming the future of work with

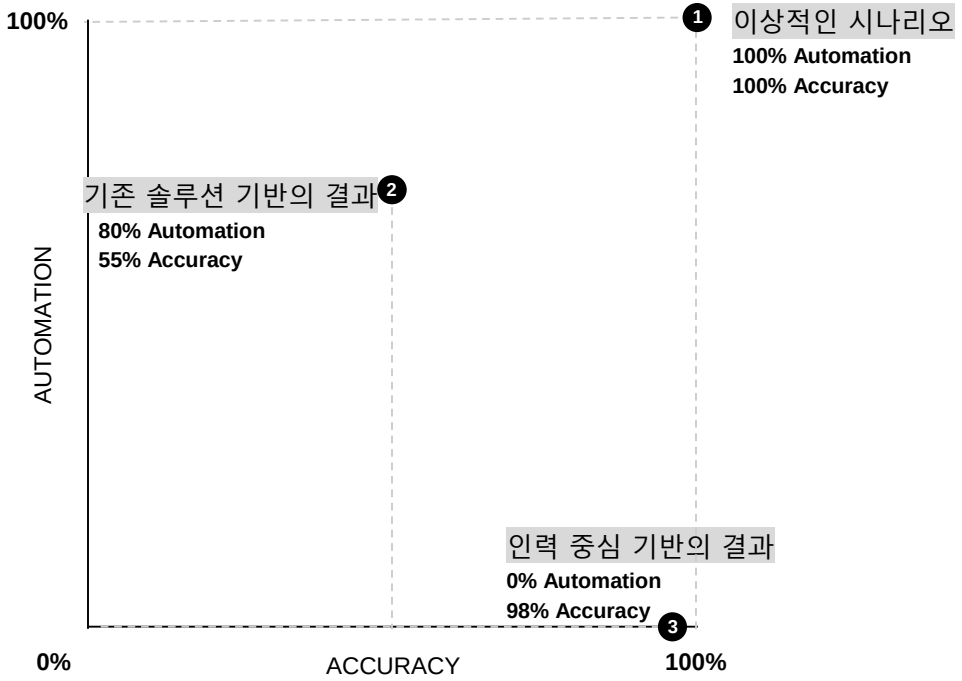
Human Centered Automation.

대부분의 솔루션들은 사람의 개입을 고려하지 않고 데이터 품질의 중요성을 간과하는 "완전 자동화" 접근 방식을 사용합니다.
하이퍼사이언스는 잘못된 숫자 하나가 청구 또는 대출 신청이 승인 또는 거부되고 누군가의 삶에 큰 영향을 미치는 차이일 수 있다는 점을 이해합니다. 그렇기 때문에 높은 정확성이 우리에게 가장 중요하며 고객이 설정한 필수 정확성 수준에 대해 자동화를 실현합니다.

기대 효과 – 도입 전

THE AUTOMATION VS. ACCURACY TRADEOFF

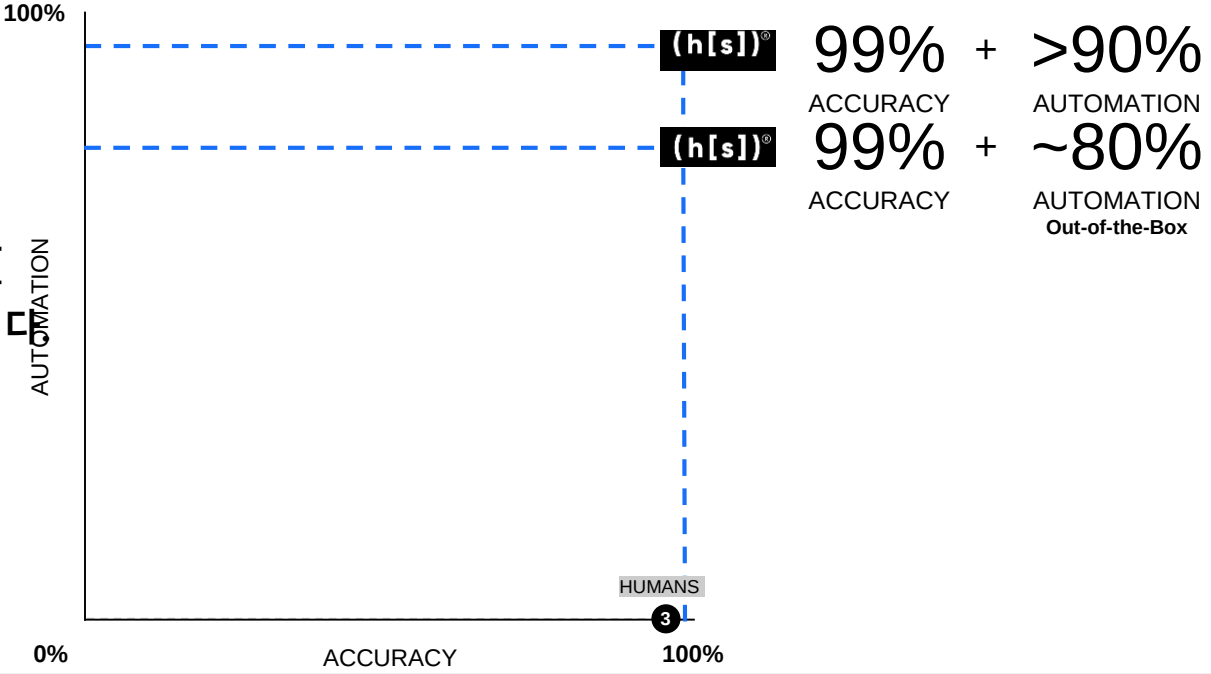
3가지 시나리오



기대 효과 – 도입 후

WITH AUTOMATION OVERTIME.

하이퍼사이언스의 AI엔진은
고객의 데이터를 지속적으로
학습하여 데이터의 정확도를
향상시켜 자동화률을 높입니다.

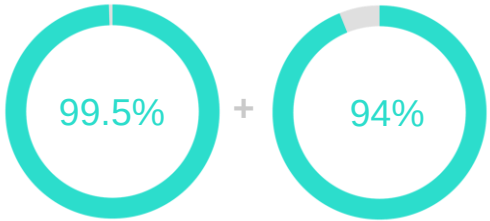


도입 사례

2022 Fortune 500 금융 서비스 고객

천만장+문서

문서 데이터 추출
속도를 3배 이상 개선

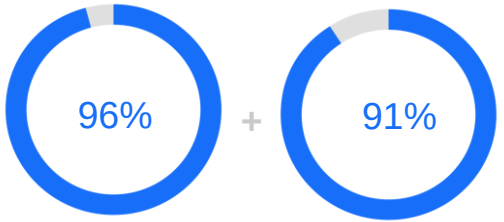


정확도 자동화

2022 Fortune 100 생명 보험 서비스 고객

100% → 9%

수동 검사 비율을
9%로 줄임



정확도 자동화

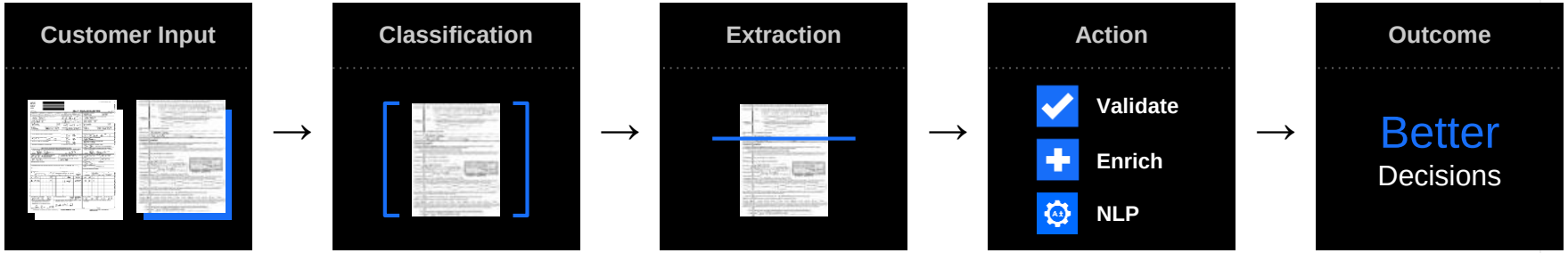
생명보험 & 퇴직연금 가입 자동화

내부 검사 자동화

(h[s])®

Platform Overview

하이퍼사이언스 플랫폼 구조



다양한 포맷의 디지털 문서
(폴더리스너, 메시지큐, 이메일등등)

문서 자동 분류
(자동 그룹화/분할/순서조정)

자동 데이터 추출,
필요시 약 1% 미만의
필드 교정,
로컬 학습으로
정확도 지속 향상

DB연결,
데이터검증,
비지니스 로직 적용

JSON등
정형 포맷으로
데이터 제공

주요 기능

Flows Studio | Supervision | Case Collation | Reporting | Software Integrations

처리 문서 종류

정형 문서(Structured)

보험금 청구 서류, 대출 서류 등 동일 신청 양식

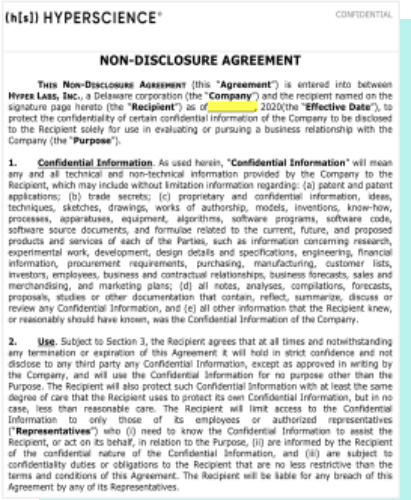
-필드 위치 고정-

반정형 문서 (Semi-structured)

수출입 문서, 각종 은행 서류, 처방전등

-필드 위치 변화 + 이름-

비정형 문서(Unstructured)

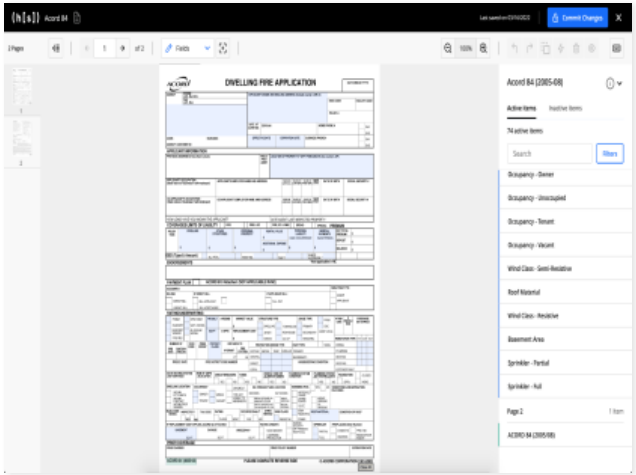


계약서, 처방전 등 완전 비정형 문서

-필드 위치 변화 + 컨텐츠-

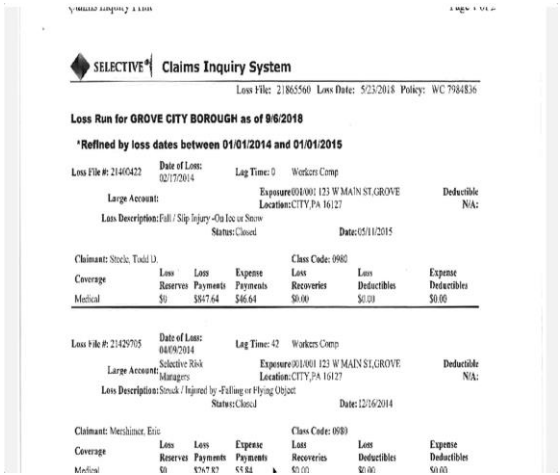
처리 문서 종류 – 계속

정형 문서



레이아웃별 1개의 빈 문서만 등록 필요, 추출하고자 하는 필드, 체크박스, 사인등을 ML으로 학습된 다양한 데이터 타입으로 1회 작업만 필요,
레이아웃이 일부 변경되어도 해당 문서의 변형버전관리(Variant) 기능으로 개발자 또는 제조사 지원없이 쉽게 대응 쉽게 대응

반정형/비정형 문서



제약 없는 다양한 폼 지원, 적은 수량으로 모델 학습 수행, 빠른 트레이닝 작업을 지원하는 포인트&클릭 인터페이스,
(i.e, 인보이스 또는 진료내역서등과 같은 추출 대상 필드가 달라지는 유사 문서들에 대해 각각 추출 모델을 만들 필요가 없어 1개의 인보이스 모델과 진료내역서 모델이 데이터를 추출하므로 개발자 또는 제조사 지원 불필요)

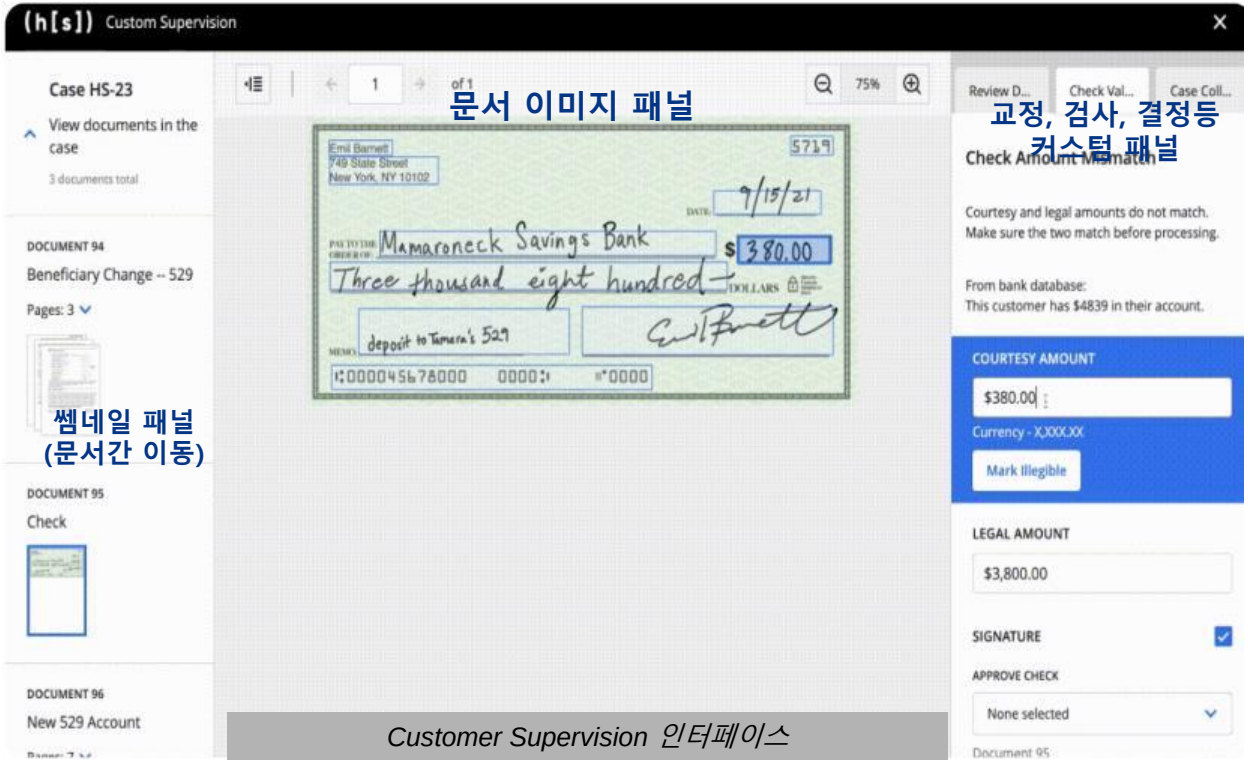
직관적인 UI/UX
비즈니스 사용자들이 직접 정형/반정형 문서의 레이아웃을 만들어 문서 처리

레이아웃 관리(Layout Versioning)
레이아웃은 계속 추가되거나 변경되는 경우 새로 업데이트하면서 벤더나 개발자 및 IT도움 없이 쉽게 적용 및 관리

정확도(Accuracy)
비즈니스 사용자들이 Ui에서 목표 정확도를 설정

휴먼인더루프(Human-in-the-loop)
비즈니스 유저들이 직관적인 Ui를 통해 부정확한 추출 정보를 교정해 나가면서 문서 형태를 학습시켜 정확도를 지속적으로 향상

차별화 기술 – 비니지스 요건 기반 Custom UI

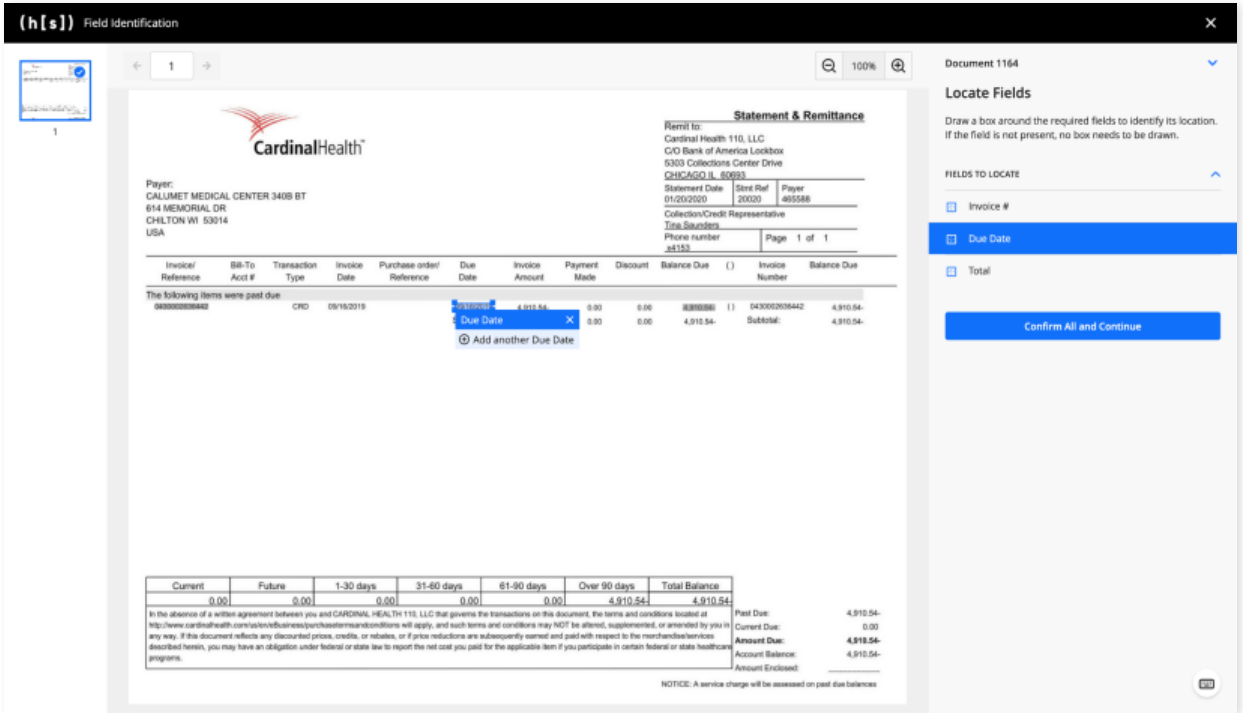


Customer Supervision 인터페이스

커스텀 수퍼비젼(Custom Supervision)
비즈니스 규칙에 따라 사람이 개입할 수 있는 맞춤형 작업 인터페이스 구축
(예: 문서의 검토 및 승인 결정등)

의사결정 지원(Facilitate Decisions)
Custom Supervision은 추출된 데이터에 대해 하이퍼사이언스가 결정을 내리는데 도움이 되도록 활용

차별화 기술 – Human in the loop + ML



Human-in-the-loop + machine learning = better performance

빠른 처리 속도

빠른 휴먼인더루프(Human-in-the-loop) 작업은 기계가 자신있게 원하는 것을 찾고 기록하는데 도움 제공. 머신 러닝은 기존 OCR 또는 룰기반 엔진들이 처리 할 수 없는 문제 해결

빠른 문서 모델링

포인트 앤 클릭 사용자 인터페이스는 반구조화된 문서에 대한 빠른 ML 학습 모델 교육을 지원

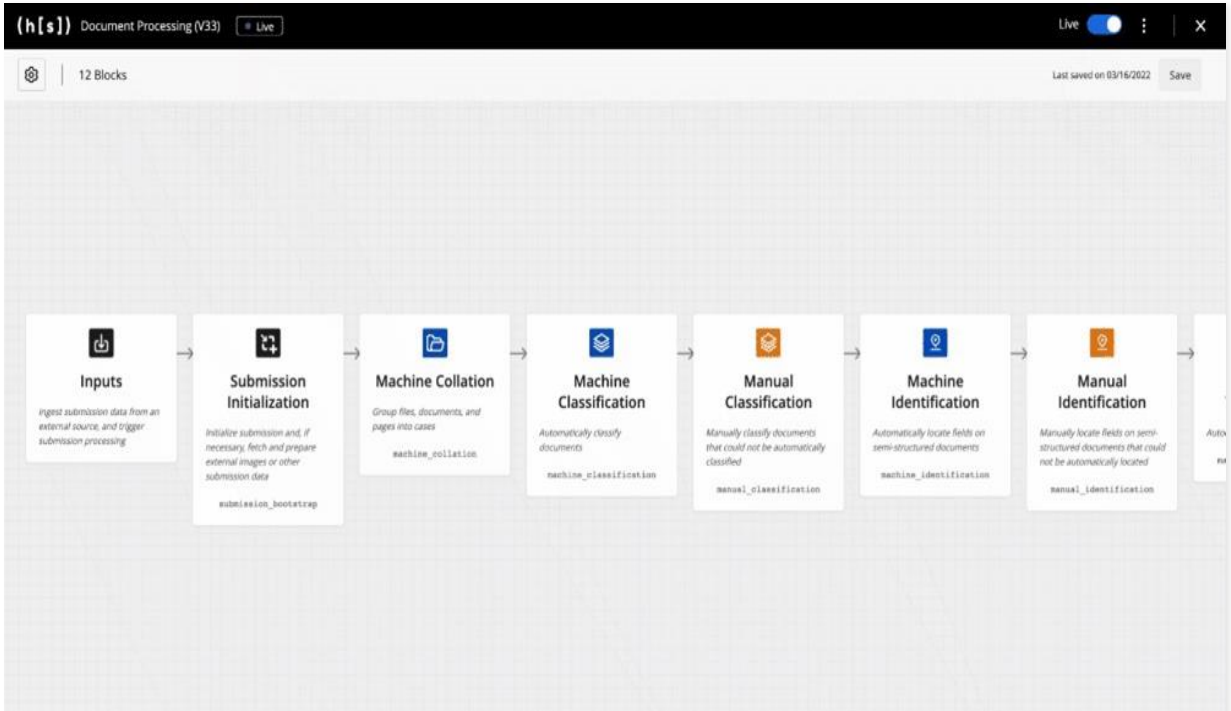
지속적인 정확도 상승

내장된 QA 프로세스로 시간 경과에 따른 정확도 향상

호환성과 연결성

사람이 개입할 수 있는 모든 작업은 자체 개발된 플랫폼내에서 제공되는 위젯을 통해 타 외부 시스템에 호출 가능

차별화 기술 – 맞춤형 블럭(Custom Code Block)



Flow Studio 인터페이스

블럭(Blocks)
문서와 관련된 관련 비즈니스
프로세스를 기반으로 하는 플로우로
즉시 사용할 수 있는 모듈화된 블록

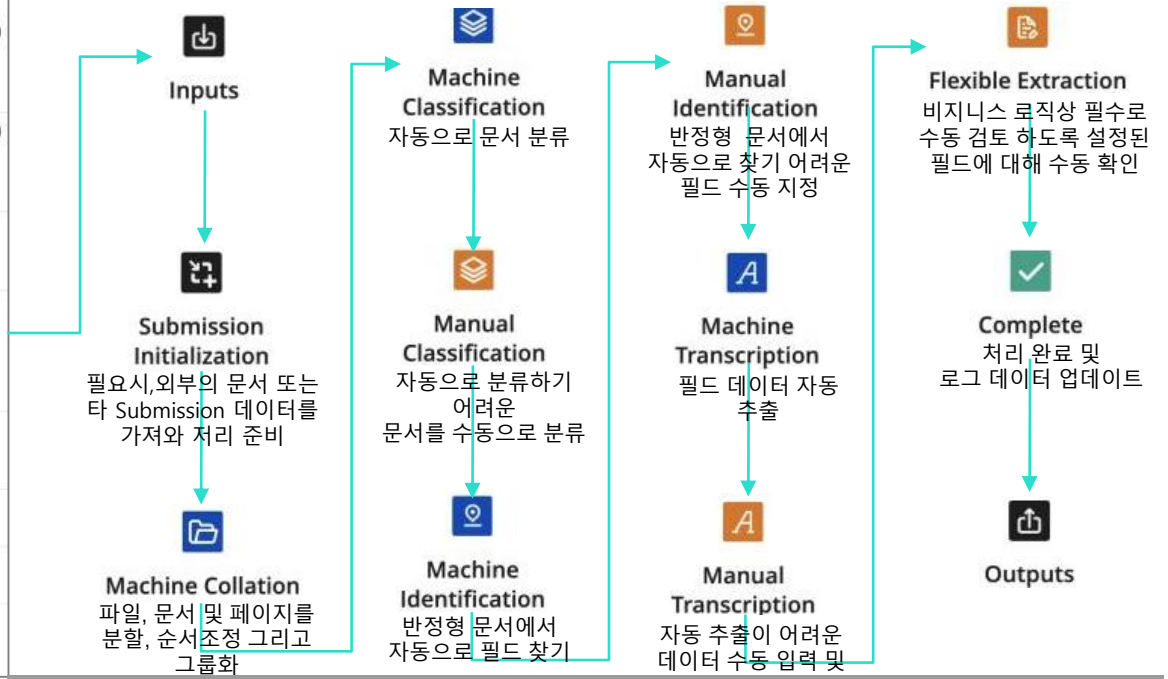
플로우 스튜디오(Flow Studio)
플로우 관리를 위한 사용자 친화적인
인터페이스

SDK(Developer Tools)
플로우 SDK는 블록 및 플로우가
비지니스를 쉽게 지원하도록
사용하기 쉬운 Python 라이브러리 제공

활용의 예
- AI와 비즈니스 규칙을 결합하여 데이터 추출 수행
- 특정 필드 서식 지정
- 교차 필드 유효성 검사
- 외부 시스템과 연결하여 필드 유효성 검사
- 외부 시스템과 연결하여 관련 데이터 가져오기
- 출력 파일 유형 사용자 정의
- 제출 문서들 필드들의 연관 유효성 검증
- 문서 내용을 기반으로 특정 팀으로 라우팅

차별화 기술 – 맞춤형 블럭(Custom Code Block)

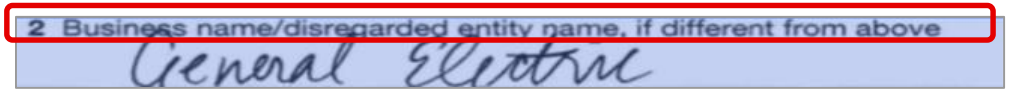
공통	• 재무제표 • 소득세징수액집계표(전년도 연말정산분) • 사업자등록증 • 인감증명서 • 본인서명사실확인서
법인, 조합	• 등기사항전부증명서(구:법인등기부등본) • 정관 또는 조합규약 • 법인 인감증명서 • 이사회의사록 사본
신용평가 작성대상	• 신용조사의뢰서 • 정 밀 • 약 식
담보물 감정(부동산)	• 토지 및 건물등기부등본 • 건축물관리대장 • 토지대장 • 도시계획확인서 • 기계기구목록
임대차조사	• 임대차계약서 • 소유자 주민등록등본 • 임주자 주민등록등본
저당권 설정	• 담보물 소유자 인감증명서 • 등기권리증
연대보증인	• 보증인 인감증명서 • 보증인 재산세 과세증명서 또는 영수증
할인어음 및 외상채권 대출 취급 시	• 어음관련업체 사업자등록증 사본 • 세금계산서 • 합계잔액시산표(전년도)



Use case : 기업운전자금 대출 관련 문서 일괄 처리

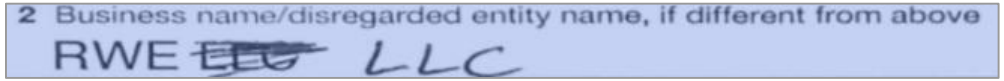
요건에 따라 SDK로 필요한 비지니스 로직을 만들어 필요한 곳에 추가,
일괄 제출된 문서들의 관련 필드의 정합성 검증 및 주요 항목 검증 로직 반영등등

차별화 기술 – 필드 속성 학습 AI

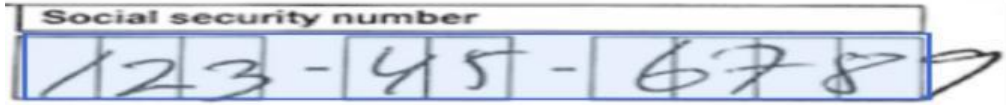


GENERAL ELECTRIC

123 ELECTRON STILL HILL ST



RWE LLC



123-45-6789

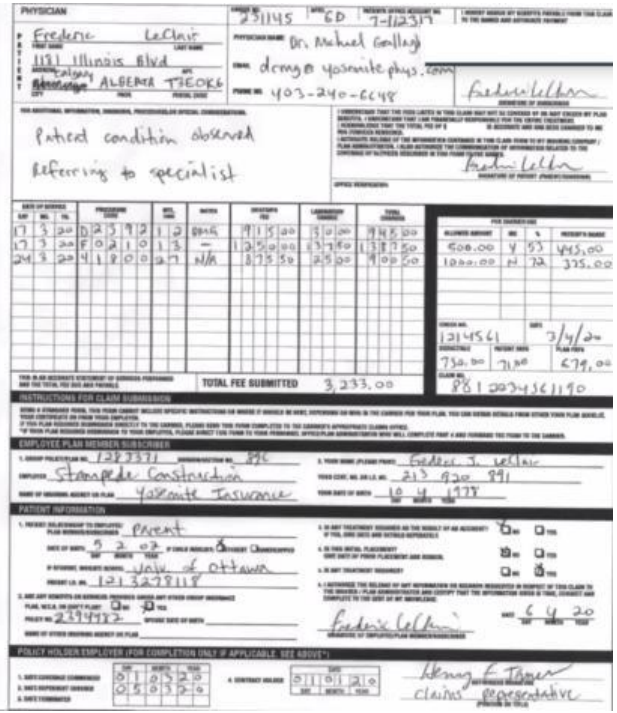
배경 가이드 글씨 자동 제거
불필요한 데이터는 제외하고
필요한 데이터만 정확하게 추출

필드 속성 학습
주소 필드의 경우 구조를 학습해서
앞쪽에는 숫자가 나오고 중간에는
숫자가 나오지 않으므로 숫자 1과 매우
유사한 소문자 l을 문자로 인식

사람의 의도 학습
필드의 속성 및 의도를
학습하는 검증된 학습 엔진

박스를 벗어난 글자 인식
별도의 작업 없이 정해진 박스를 벗어
나는 숫자 및 문장 추출

차별화 기술 – 다양한 문서 자동 처리



해상도 비종속
100DPI이하의 문서 추출 지원

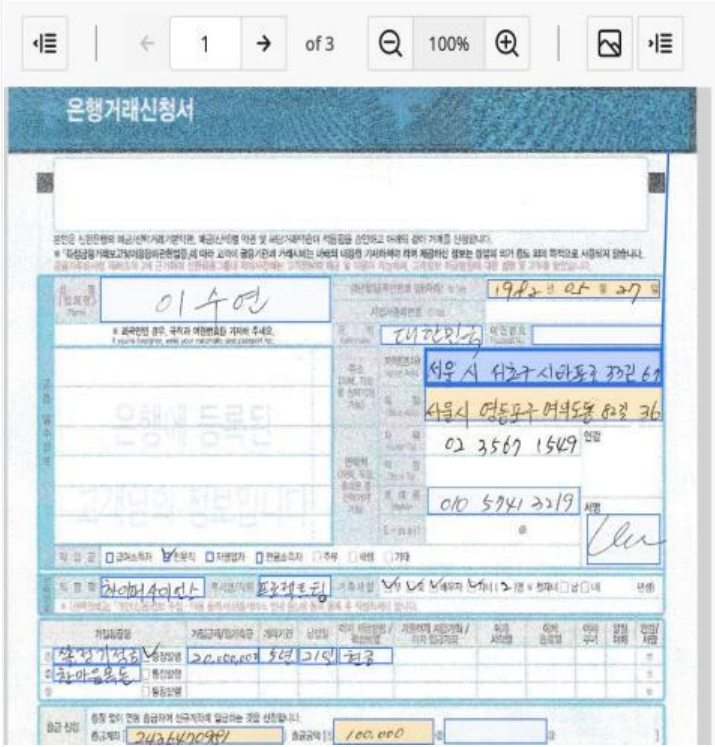
노이즈 제거
문서의 가독성을 저해하는 노이즈 제거

뒤틀림 교정
추출에 어려움을 야기할 수 있는
비틀어지거나 뒤틀린 문서 자동 교정

뒤집힌 문서 교정
위아래가 뒤집힌 문서 자동 교정

이미지 보정외

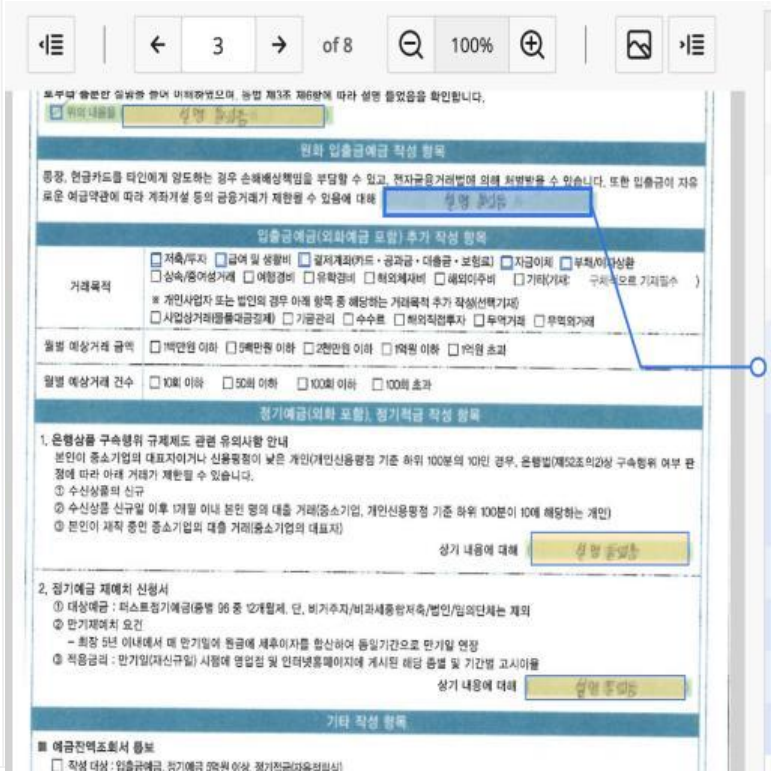
차별화 기술 – 뛰어난 수기 인식



Raw	서울시 서초구신반포로 33길 67
Identified by	Machine
Transcribed by	Machine
Show more	
직장	서울시 영등포구여의도동 82길 36
직장-전화번호	
휴대폰	010 5741 3219
서명	Signed
급여소득자	Not Checked
전문직	Checked
자영업자	Not Checked
연금소득자	Not Checked
직장명	하이퍼사이언스
부서명	프로젝트팀
1-가입상품명	쏠정기적금

한글 수기 인식
높은 수기 인식률

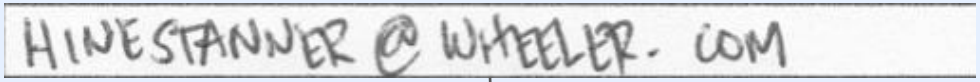
차별화 기술 – 뛰어난 수기 인식(계속)



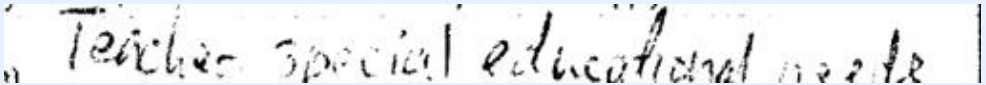
한글 수기 인식
높은 수기 인식률

PAGE 3	
예금보험관계-위의 내용	Checked
예금보험관계 설명 확인	● 설명 들었음
금융실명거래 확인 작성 확인-위의 내용	Checked
금융실명거래 확인-설명들었음	● 설명 들었음
원화입출금예금 작성항목 확인-설명들었음	
● Supervision Required, But Disabled Supervision was required for this field, but supervision was disabled for the submission.	
Normalized	설명 들었음
Raw	설명 들었음
Identified by	Machine
Transcribed by	Machine

차별화 기술 – QA를 통한 정확도 지속적인 상승



HINESTANNER@WHEELER.COM



Teacher special educational needs

축적되는 교정 데이터(Fine tuning)
기계의 요청에 의한 사람의 도움으로
학습데이터가 내부 HS에 축적되고
시간이 지날수록 기계의 인식률 상승

(h[s])®

Demo

KYC Income Validation Use case

Verifies reported income as part of mortgage package submissions

원천징수(Pays As You Go) 문서 검사

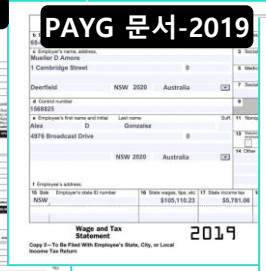
주택 융자 신청폼에 고객이 입력한 수입금액과 PAYG 문서의 수입금액 비교

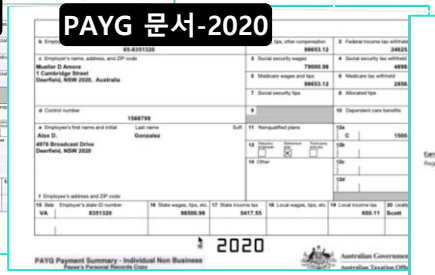
제출된 3개월이내 은행 거래 상태 문서 검사

제출된 최근 2개월 급여 지급 명세표 검사

모든 비지니스 요건 로직 블럭

비지니스 요건에 최적화된 UI -결과에 따른 결정등





KYC과정에서 필요한 4가지 정형 및 반정형 문서를
비지니스 로직을 적용하여 하나의 작업으로 자동검사 처리 및 데이터 추출

(h[s])®

Appendix

도입시 고려 사항

타 포인트(OCR or AI OCR)솔루션을 운영중이거나 새로 도입을 고려중인 조직의 고민 사항들

- 수기 한글 문서 미지원 또는 매우 낮은 정확도 – 수기 문서는 자동화 대상 제외로 인한 미처리된 문서 자동화
- 레이아웃이나 템플릿 설정의 어려움 – 도입후 고객사의 자체 운영의 한계
- 일부 변경된 정형문서의 처리 및 문서 해상도 문제 – 변경된 문서에 대한 알고리즘 재작성 및 100 DPI 이하 문서 처리 어려움
- 기본 추출에 긴 시간 필요 – 기본 모델링에 많은 시간 소요
- 위치나 포맷이 달라지는 반정형 문서 처리의 어려움 – 각 문서별로 개발자나 제조사가 각각 모델을 모두 개발해야 함
- 데이터 추출과정에서 필요한 외부 시스템과의 연동의 제약 – API등을 통한 연동 미지원
- End to End 자동화 미지원 – 단순히 문서에서 데이터만 추출하여 전처리 그리고 비지니스 로직기반 검증등의 별도 후처리를 위한 개발 필요
- 잘못된 데이터 추출의 가능성을 줄일 수 있는 기능 미비 – 최소의 리소스 투입으로 지속적인 교정을 통해 정확도와 자동화 고도화 필요
- 교정 데이터의 정보의 로컬 학습의 부재로 반복적인 데이터 수정 – 교정 데이터가 시스템에 반영되어 반복 교정을 최소화 및 정확도 개선
- 다양한 문서의 노이즈 처리 – 기울어지거나, 뒤집어진, 노이즈를 제거에 별도의 전처리작업 필요
- 여러 다른 종류의 문서 일괄 자동 처리 – 문서들의 자동 분할/순서조정/그룹화등 요건 미지원
- 문서 데이터에 비지니스 로직 반영의 한계 – 불완전 판매 , 대출 계약, 보험금 청구 업무 등등에서 함께 동반되는 다른 종류의 문서들에서 필수 데이터 검증 로직 적용의 어려움
- 도입후 관련 노하우 내재화의 어려움 – 개발자의 이직 또는 제조사의 지원 수준에 영향을 받는 기술적 구조

(h[s])®

THANK YOU!