



#GlobalAzure
#GlobalAzureMilano

Seeking limits of Document Intelligence while extracting data from complex documents

Zahhar Kirillov

Senior Project Manager, EPAM Systems (Switzerland)

Azure AI Document Intelligence (formerly Form Recognizer) is a good old Optical Character Recognition (OCR) powered by machine-learning models: cloud service to automate data extraction in applications and workflows.

Engineers benefit from Prebuilt models ready to understand Invoices, IDs and variety forms, but also can train own models (classifiers, templates, neural)



01

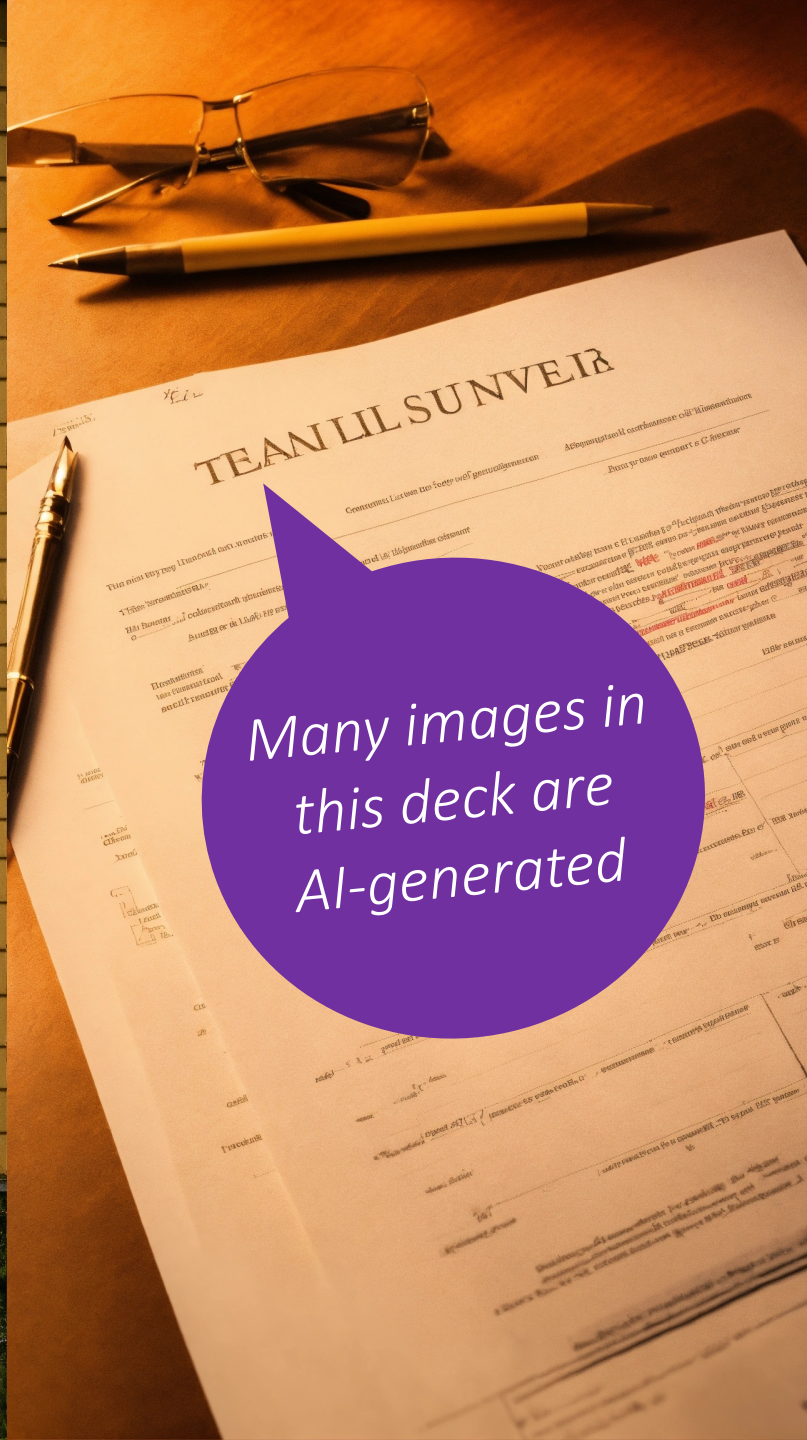
Business Case




Once upon a time...



Oops!
We
Have a
Problem
:-(



Many images in this deck are AI-generated

A person with dark hair, wearing a light-colored long-sleeved shirt and dark pants, is sitting on a white leather sofa. They are looking down at a laptop computer on their lap, with their right hand on the keyboard and their left hand holding a smartphone. The background is a blurred indoor setting with a window and some plants.

To get a loan, banks asks applicant to have **a life insurance** for the whole repay period.

To get a life insurance, applicant needs **evidence of insurability** (proof of identity and good health)

Dichiarazione di Buono Stato di Salute del Debitore Cedente

Agli effetti della validità della Polizza, il sottoscritto **DICHIARA** sotto la propria responsabilità di essere in buona salute, ed in particolare:

si prega di contrassegnare le risposte annerendo completamente le rispettive caselle SI NO

1. Di avere una **differenza tra la propria altezza** (espressa in cm) ed il **proprio peso** (espresso in Kg) compresa tra 85 e 110 (es. 175 – 70 = 105)

2. Di avere riconosciuta da una compagnia di assicurazione o dall'INPS una **invalidità permanente** totale (superiore al 66%) oppure di percepire **pensione o assegno di invalidità** per infortunio o malattia o di aver presentato domanda per ottenerla (in caso di risposta affermativa è necessario inviare alla compagnia il relativo verbale di invalidità con grado e diagnosi)

Dichiara di **aver avuto**, negli ultimi 10 anni, una diagnosi, cure continuative, trattamenti, interventi chirurgici, essere stato/a ricoverato/a o di essere in attesa di ricovero e/o intervento chirurgico o aver assunto farmaci (ad esclusione dei farmaci assunti per il ipertensione o controllo del colesterolo) in merito a:

3. **Malattie cardio-vascolari** quali: infarto del miocardio, ischemia coronarica, patologie delle valvole cardiache, aritmie cardiache, trattamento anticoagulante orale, portatore di defibrillatore impiantabile, scompenso cardiaco cronico, aneurismi arteriosi (dell'aorta o di altre arterie periferiche), tromboembolia polmonare, l'ipertensione arteriosa grave non controllata dalla terapia o che abbia causato danni ad organi e/o apparati

4. **Malattie cerebrovascolari**, inclusi ictus e attacco ischemico transitorio, aneurismi dei vasi cerebrali

5. **Malattie oncologiche**: cancro, neoplasie maligne, incluso il tumore alla pelle (melanoma), leucemie e linfomi

6. **Malattie del sistema nervoso**: morbo di Alzheimer, morbo di Parkinson, paresi/plegia, distrofia muscolare, sclerosi multipla, sclerosi laterale amiotrofica e altre patologie neurodegenerative

7. **Malattie respiratorie gravi**: enfisema, bronchite cronica ostruttiva (BPCO), asma che abbiano reso necessario un ricovero ospedaliero, fibrosi polmonare, insufficienza respiratoria cronica, tubercolosi

8. **Altre patologie**: diabete, gotta, disturbi ormonali (tiroide, ghiandola surrenale), anemia, alterazione della coagulazione, AIDS, insufficienza renale cronica, artrite reumatoide, cirrosi epatiche

9. **Ai soli fini statistici**: ha completato il ciclo vaccinale anti COVID?

DICHIARA inoltre di prosciogliere dal segreto professionale tutti i medici, gli ospedali e gli istituti di cura in genere che siano in possesso di notizie di carattere sanitario che lo riguardano; di essere consapevole di dovere dare sollecita comunicazione a, per il tramite del Contraente, di eventuali nuovi fattori inerenti il proprio stato di salute intervenuti tra la data di sottoscrizione del presente Modulo e la data di erogazione del Finanziamento, al fine di consentire all'Assicuratore la corretta valutazione del rischio assicurato.

DATI DEL MEDICO CURANTE: Nome: _____ Cognome: _____
 Indirizzo dello Studio _____ Telefono _____

AVVERTENZA Le dichiarazioni non veritiere, inesatte o reticenti rese dal soggetto Debitore Cedente possono compromettere il diritto alla prestazione ai sensi degli artt. 1892 e 1983 del Codice Civile. Prima della sottoscrizione, il Debitore Cedente deve verificare l'esattezza delle dichiarazioni riportate. (Il Debitore Cedente ha il diritto di non sottoscrivere la presente Dichiarazione di Buono Stato di Salute e di produrre il Questionario Medico o di sottoporsi a visita medica per certificare il proprio stato di salute. Il costo della visita medica è a carico del Debitore Cedente).

Luogo: _____ Data: _____ Firma del Debitore Cedente (leggibile) _____

Altezza (cm): _____
 Peso (kg): _____
 Pressione arteriosa: min _____ max _____

Si prega di compilare la sezione "Precisazioni" in modo leggibile e di contrassegnare le risposte annerendole completamente le rispettive caselle

N.	Domanda	SI	NO	Precisazioni
01	Esercita regolarmente attività sportive?	<input type="checkbox"/>	<input type="checkbox"/>	<i>Se Sì, quali: e con quale frequenza?</i>
02	Negli ultimi 2 anni ha fatto uso di prodotti da fumo, tabacco, o nicotina in un'altra forma?	<input type="checkbox"/>	<input type="checkbox"/>	<i>Se Sì, indicare prodotto: e quantità media a settimana?</i>
03	È o è stato negli ultimi 10 anni in consultazione o in trattamento in relazione al suo consumo di alcol (incluso chiarimenti speciali / visite / servizio di consulenza)?	<input type="checkbox"/>	<input type="checkbox"/>	<i>Se Sì, quando? Da parte di chi? Precisare nome ed indirizzo?</i>
04	Fa o ha fatto uso di droghe negli ultimi 10 anni?	<input type="checkbox"/>	<input type="checkbox"/>	<i>Se Sì, quali? Per quanto tempo? Quando l'ultima volta?</i>
05	Negli ultimi 5 anni ha sofferto, o attualmente soffre, di malattie per cui è stata/è necessaria una cura farmacologica continuativa?	<input type="checkbox"/>	<input type="checkbox"/>	<i>Se Sì, quali? Motivo? Inizio e fine periodo di assunzione?</i>
06	E' mai stata riconosciuta una invalidità permanente oppure percepisce pensione o assegno di invalidità per infortunio o malattia o ha presentato domanda per ottenerla? <i>In caso di risposta affermativa inviare alla compagnia verbale di invalidità con grado e diagnosi</i>	<input type="checkbox"/>	<input type="checkbox"/>	<i>Se Sì, motivo? Da quando a quando?</i>
07	Ha mai ottenuto un rifiuto o un differimento ad una richiesta di assicurazione (vita, invalidità, malattie gravi) oppure un'accettazione a condizioni speciali o con sovra premio?	<input type="checkbox"/>	<input type="checkbox"/>	<i>Se Sì, quale assicurazione? Quando? Motivazione?</i>
08	Nella sua parentela consanguinea ci sono stati casi di malattie di cuore o della circolazione, malattie del sistema nervoso, malattie mentali, tumori maligni, colpi apoplettici, malattie ereditarie, suicidi, alcolismo o diabete prima dei 55 anni?	<input type="checkbox"/>	<input type="checkbox"/>	<i>Se Sì, quali malattie? Grado di parentela? Età alla diagnosi?</i>
09	Ha o ha avuto delle malattie, disturbi o affezioni negli ultimi 10 anni: a. dell'apparato respiratorio, quali asma grave, bronchite cronica o ricorrente, polmonite ricorrente, tubercolosi polmonare o altro? <i>In caso di risposta affermativa allegare la documentazione medica specifica già in suo possesso</i>	<input type="checkbox"/>	<input type="checkbox"/>	<i>Se Sì, quali? Quando? Per quanto tempo? Adesso è guarito?</i>



Oggetto: Comunicazione di Liquidazione
 decorrenza 1 agosto 2019
 codice fiscale MRNM _____

La informo che la richiesta presentata il 18 luglio 2019 è stata accolta e che Le è stata liquidata la pensione ai superstiti, categoria SO numero 220 _____, con decorrenza dal 1 agosto 2019.
 L'importo mensile della pensione alla decorrenza è di euro 791,30.
 Il calcolo alla decorrenza è stato effettuato nella misura del 60,00% della pensione numero 11542022 categoria VO Sede 4901 di importo pari a euro 1.318,82, spettante al dante causa ALB _____, sulla base della contribuzione versata o accreditata in suo favore e dei redditi dallo stesso posseduti.

LE TRATTENUTE SULLA PENSIONE

Sulla pensione sono operate trattenute per:
 - incumulabilità con redditi IRPEF esclusi quelli derivanti da altra pensione di reversibilità come previsto dalla legge 335/95 articolo 1, comma 41.

L'IMPORTO MENSILE SPETTANTE

Dal	Importo a calcolo	IMPORTI DELLA PENSIONE		Trattenuta ONPI	Importo spettante
		Trattenuta L. 335/95 art. 1 c.41	Trattenuta L. 335/95 art. 1 c.43		
08/2019	791,30	395,64	0,00	0,01	395,65
13/2019	329,71	164,85	0,00	0,01	164,85

IMPORTO DEGLI ARRETRATI

Sono stati determinati arretrati per il periodo dal 1 agosto 2019 al 30 settembre 2019.
 Nella sottostante tabella viene riportato il credito spettante suddiviso per anno di riferimento:

Anno	Importo pensione	Importo totale
2019	791,32	791,32
Totale	791,32	791,32
Trattenuta per ritenuta IRPEF		0,00
Trattenute per ritenuta IRPEF anni precedenti		0,00
Trattenute IRPEF arretrati anni in corso		0,00
Trattenute IRPEF arretrati anni precedenti		0,00
Trattenute per quota associativa sindacale		0,00
Trattenute per contributo ex ONPI		0,02
Importo al netto delle trattenute		791,30

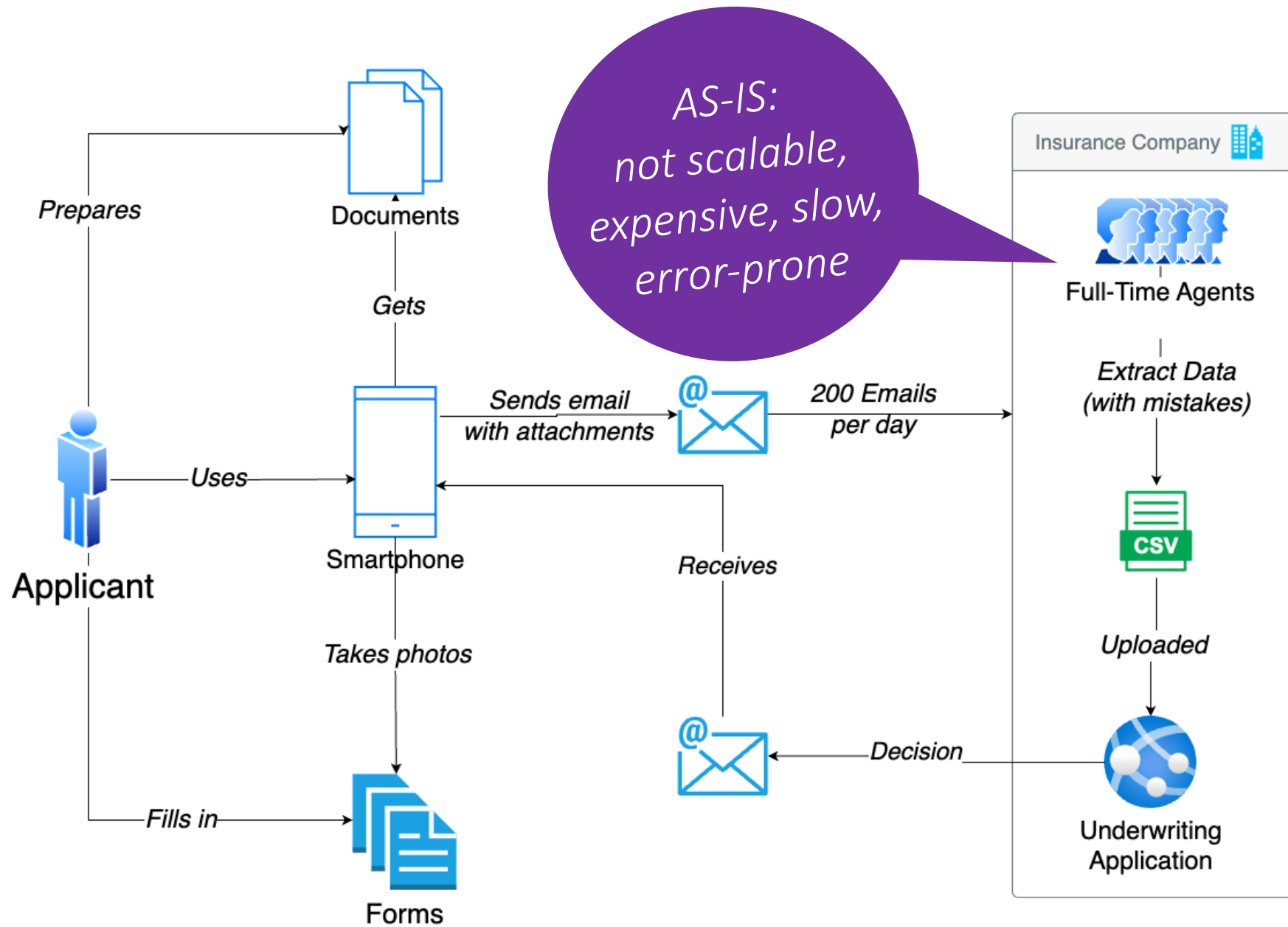
Good Health Declaration

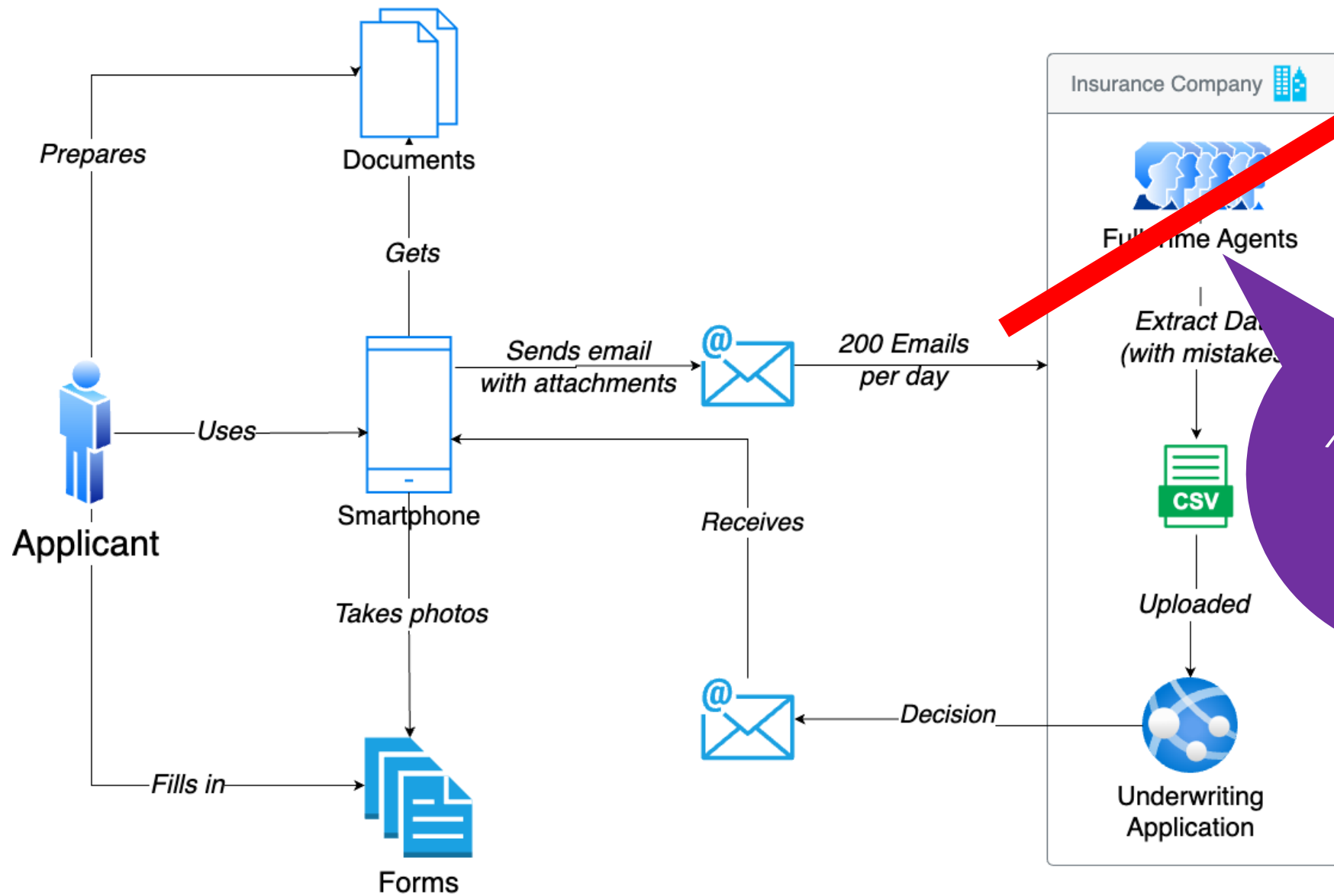
Medical Questionnaires

Social Security Documents

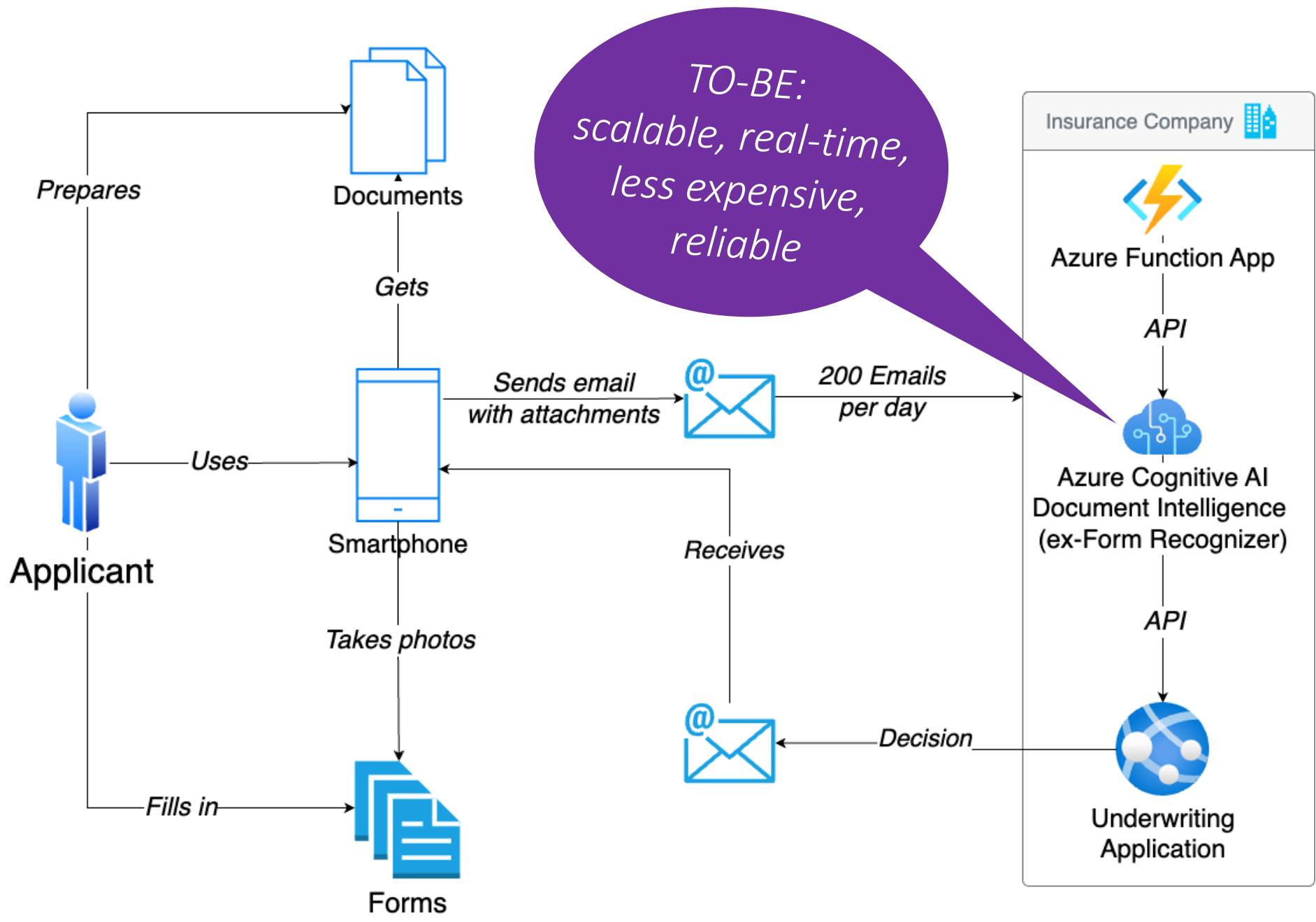
02

Architecture





OH NO!
AI is here to
replace
humans!



03

Solution

First obstacle: Availability of new features/API

*Real Magic to
impress customer
or to deal with
“Very Creative Layouts”,
also good to
build dynamic apps*

Current GA release (available worldwide)

- Form Recognizer 4.1.0 (2023-08-10)
- REST API 2023-07-31

What we need is **Document Intelligence**:

- 1.0.0-beta.2 (2024-03-05)
- REST API 2024-02-29-preview
- Only available in East US, West US2 and West Europe Azure Regions
- Totally new client with many renamed methods and attributes
- Supports 'QueryFields' – extract up to 20 custom fields from any prebuilt-model without having to train a custom model
- Supports Markdown output
- Supports Incremental training
- No stream input, only URI or Base64

Next obstacle: Free Tier is too limited for development

- Max document size 4MB (vs 500MB)
- 2 pages maximum (vs 2000)
- 5 models max can be composed (vs 200)
- 1 request per second (vs 15)
- No Premium Features
 - Query fields
 - Font face (e.g Arial, Times) detection
 - High resolution (handy for A3+ page sizes)
 - Formula extraction

What we were expected to build?

Features

- Verify if application is complete
 - Is any ID present?
 - Are all questionnaires (depend on business rules, e.g. applicant age) fully filled in?
 - Are all papers signed?
- Verify if application is consistent
 - Are all documents belong to the same person?
 - Are there any expired documents?

Challenges

- Handwritten input
- Variety of layouts
- Too small fields
- Corrections
- Multiple docs in one PDF file
- Multi-page forms with pages put in random order within PDF
- Date format DD/MM vs MM/DD
- Signature comparison
- Poor quality / bad angle photos

...relativamente al proprio stato di salute. Tali
prestazioni assicurate nonché la stessa cessazione dell'Assicurazione.

Il sottoscritto DEBITORE CEDENTE:

Nome e cognome	MARIO BIANCHI	Sesso	M	Luogo di Nascita	ROME	Data di nascita	23/12/76
Codice Fiscale	BNCMRA70A204501	Città	ROMA	Prov.	RO	CAP	10107
Indirizzo Resid.	VIA CARDINALE	Città	ROMA	Prov.	RO	CAP	10107
Tipo documento*	I	Numero	12345678	Prov.	RO	CAP	10107
Data rilascio	1/1/76	Luogo Rilascio	COMUNE DI ROMA	Prov.	RO	CAP	10107
FINANZIAMENTO - Ente erogatore:	COMUNE DI ROMA	Numero	17	Prov.	RO	CAP	10107
Durata (mesi)	18	Montante €	118000	Prov.	RO	CAP	10107
(*) 01=carta d'identità; 02=patente di guida; 03=passaporto		Finanziamento	30000				

Overlap

DICHIARA
• di essere consapevole che il Contraente intende sottoscrivere

prestazioni assicurate nonché la stessa cessazione dell'Assicurazione.

Il sottoscritto DEBITORE CEDENTE:

Nome e cognome	Markos Monselas	Data di nascita	12/12/1999				
Codice Fiscale	1522E39WF39J	Luogo di Nascita	GRICO				
Indirizzo Resid.	Weissstraße	Città	Napoli	Prov.	11	CAP	15
Tipo documento*	S	Numero	1851	Rilasciato da	TSX37		
Data rilascio	15/6/1971	Luogo Rilascio		Prov.	11	CAP	15
FINANZIAMENTO - Ente erogatore:	1466	Numero Finanziamento	00037	Prov.	11	CAP	15
Durata (mesi)	12	Montante €	1.000€	Quota mensile del Finanziamento			
(*) 01=carta d'identità; 02=patente di guida; 03=passaporto							

Date format?

Correction

DICHIARA

Initial design

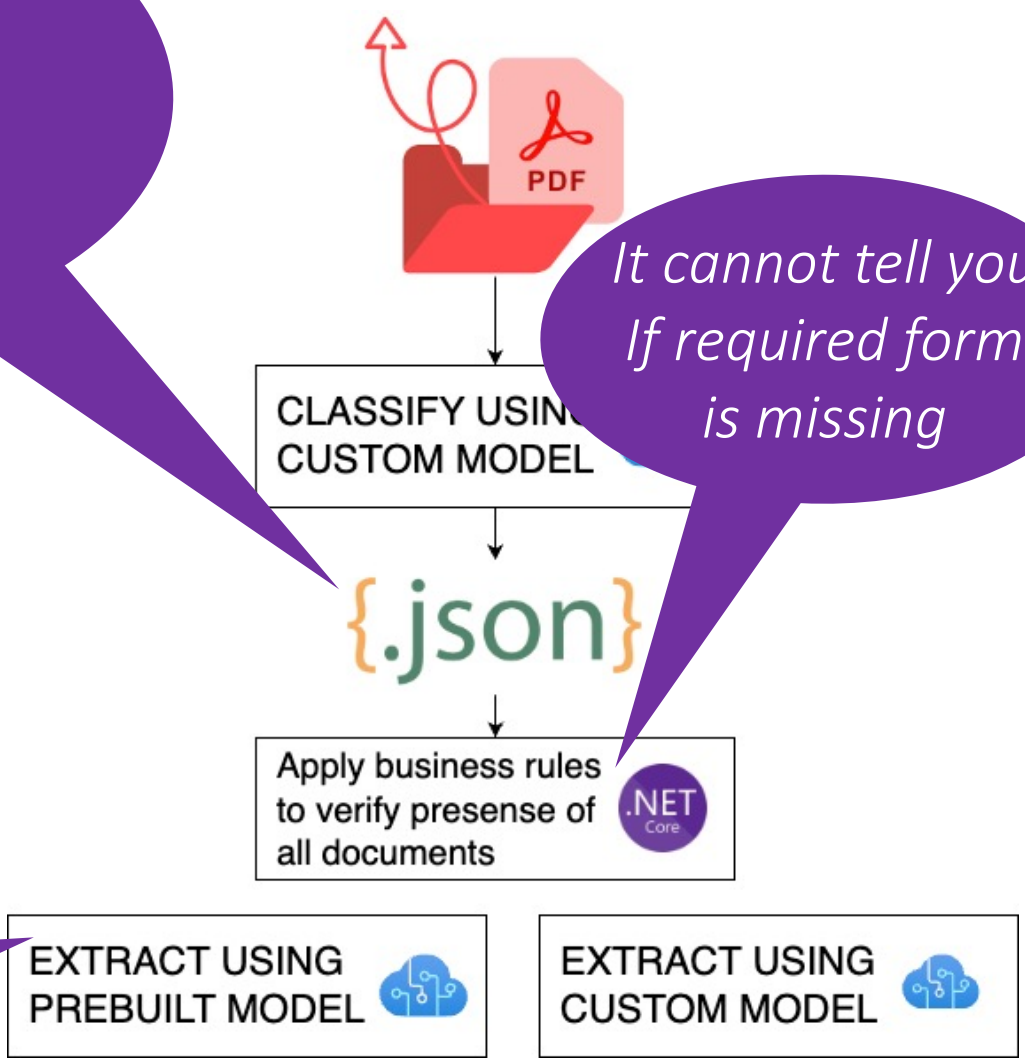
200 applicants per business
 up to 10 pages per applicant.
 40'000 pages/month

Poor result if input has "junk" pages or multi-paged forms in random page order

It cannot tell you if required form is missing

Model	Costs, \$
Custom to Classify	1600 (commitment + overage)
Prebuilt to extract IDs	40 (pay as you go)
Custom to extract data from all other documents	1440 (commitment + overage)
Total	3080

Prebuilt models are aimed to US-market

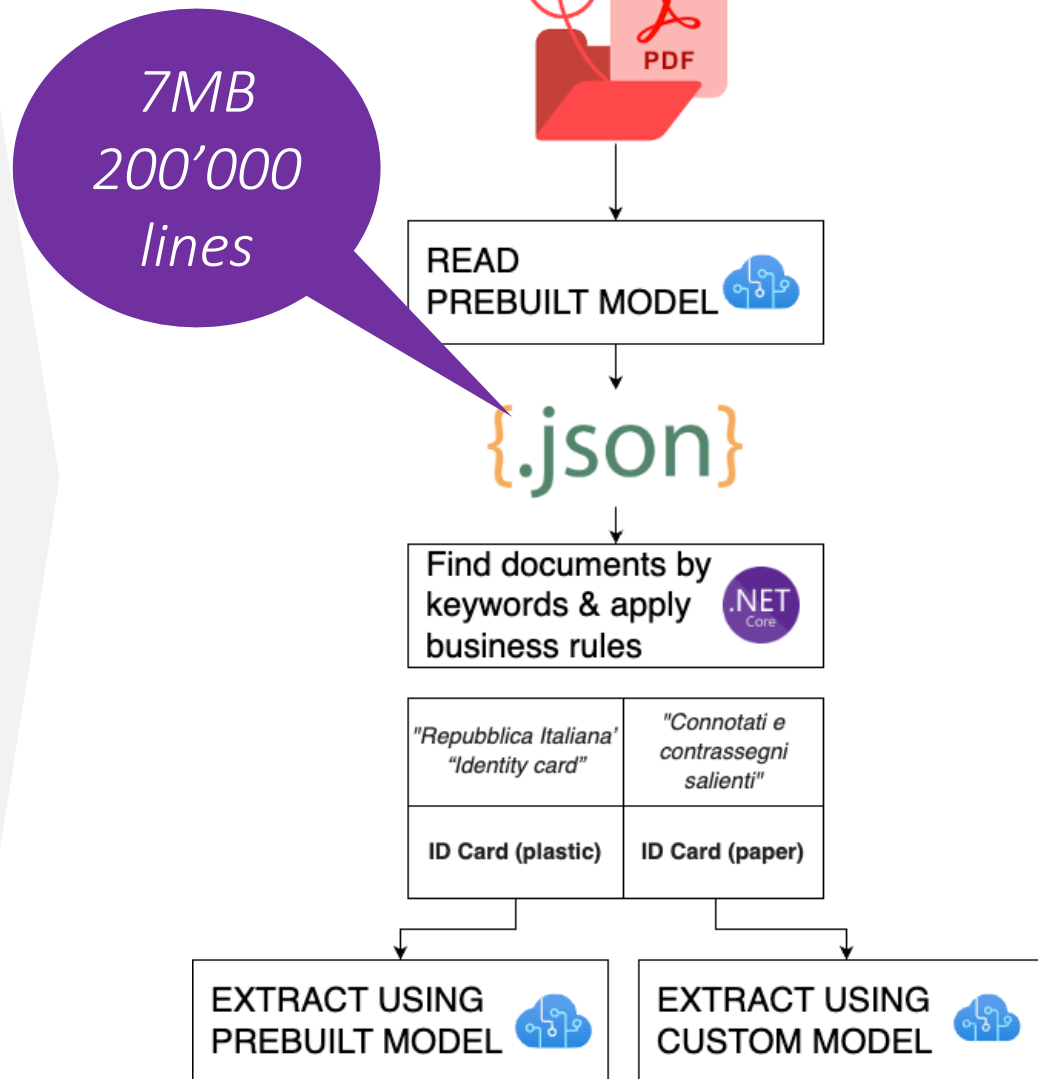


Final design

200 applicants per business day,
up to 10 pages per applicant:
40'000 pages/month

Model	Costs, \$
Read	60 (pay as you go)
Prebuilt to extract IDs	40 (pay as you go)
Custom to extract data from all other documents	1440 (commitment + overage)
Total	1540

Actually much less, because READ model gives enough data to avoid any further extraction



What we have build – Features

- Is any ID present?

Easy with READ model, but also cheap & easily done with classifier.

EU modern cards are well supported by prebuilt model.

For legacy IDs train a custom model or reuse PII found elsewhere to search in READ output.

- Are all questionnaires (depend on business rules, e.g. applicant age) fully filled in?

Easy with custom model; very high confidence level; train for every checkbox/mark.

- Are all papers signed?

- Easy, high confidence rate, but unable to detect similarity of signatures, only presence.

- Are all documents belong to the same person?

- Solved by business logic: we rely on Codice Fiscale and check its correctness algorithmically.

- Are there any expired documents?

- Solved by business logic.

What we have build – Challenges

- Handwritten input → No problem, very good quality for majority of cases
- Variety of layouts → Explore “Composed Model” feature.
If forms vary slightly by design/layout yet deliver same content – combine them as “Composed Model”, but avoid this approach for forms with different content in similar design. Also avoid “Composed Models” to combine forms and tables for similar content.
- Too small fields → Confidence suffers. When possible – improve printed forms.
- Corrections → Confidence suffers / unreliable result. Needs human validation.
- Multiple docs in one PDF file → No problem, READ model handles it well, also Classifier.
- Multi-page forms with pages put in random order within PDF -> READ model handles well
- Date format DD/MM vs MM/DD → set Field Type (Date) & Field Subtype/Format (e.g. DMY)
- Signature comparison → not solved now
- Poor quality / bad angle photos → No problem, very good quality for majority of cases

Final obstacle: how to copy custom model to another tenant?

- Document Intelligence Studio / Azure Portal GUI does not have a feature to copy/move models between Document Intelligence instances.
- The workaround (or “true way”) is to use REST API for this (you may build a pipeline in Azure DevOps for this)

Lessons learned

01

Azure Document Intelligence Studio is easy & fun to use, so that all model training & testing were accomplished by business users, without involvement of engineers.

02

Azure Document Intelligence V1, when becomes Generally Available, should be an amazing product capable to solve hardest tasks of data extraction.

03

Being already reasonably priced, one can reduce the bill even more by opting for Commitment Tier or going offline (with Container Instances) or by using their creativity.

04

As of today, AI is far from taking humans jobs – mostly because its outcome requires validation and verification, making AI a great tool to augment humans at attention-intensive tasks.

Thank you!

Zahhar Kirillov

Senior Project Manager

Zahhar_kirillov@epam.com