

January 20, 2021

UiPath 20.10产品新功能深入解析： 文档理解

崔静

UiPath售前顾问

议程

- 01 文档理解的基本概念
- 02 深入探索UiPath文档理解
- 03 文档理解Demo演示及案例分享
- 04 问答环节

1 文档理解的基本概念

文档处理是RPA的核心，为什么？



流程自动化的主要承诺之一就是捕获文档中的数据



世界上没有公司不处理文件，特别是在银行、金融、保险、制造业、人力资源、公共部门等行业



任何销售流程都涉及重复的单据处理（应付和应收、发票、收据、采购订单、货运跟踪）

文档处理可以是一项挑战

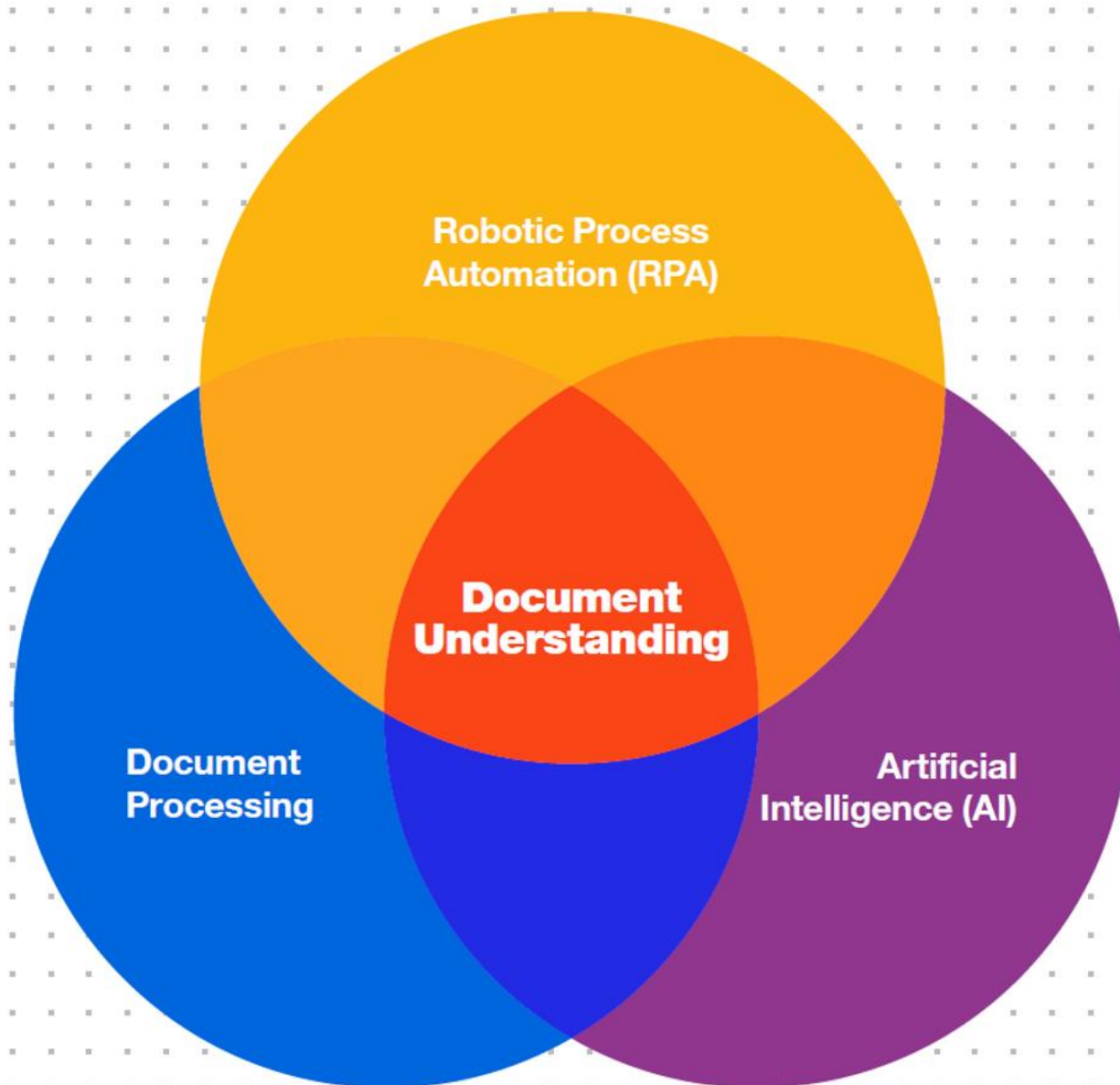
-  手动提取、解释、操作
-  各种各样的文档类型和质量
-  消耗成本和时间
-  人为错误、返工、损失
-  缺乏端到端的复杂解决方案

文档处理可以是一项 解决方案



什么是文档理解?

文档理解是从大量文档中**提取和解释其信息和含义**的能力。
它出现在文档处理、AI和RPA的交叉点上。



不是
OCR

不是
计算机视觉

非结构化

- 没有标准结构，例如文本块
- 在数据一致性、结构、措辞上存在显著差异
- 通常需要“阅读”文本来提取含义，以识别要提取的“正确”信息
- 需要重新训练以保持/改善性能

SALE OF GOODS CONTRACT

Terms and Conditions of Sale

THIS IS A LEGAL DOCUMENT ("SALES CONTRACT") BETWEEN YOU ("BUYER") AND ULTRA NECTAR, INC. ("SELLER"). PLEASE READ THIS AGREEMENT CAREFULLY. BY USING AND ACCESSING THE ULTRANECTARGRAPH WEB SITE YOU INDICATE THAT YOU HAVE READ AND UNDERSTAND THIS AGREEMENT AND AGREE TO BE BOUND BY THIS AGREEMENT. IF YOU DO NOT ACCEPT THIS AGREEMENT, DO NOT ACCESS AND USE THE ULTRA NECTAR, INC. WEB SITE. PLEASE NOTE THAT THE TERMS AND CONDITIONS MAY BE PERIODICALLY UPDATED AND MODIFIED, SO PLEASE BE SURE TO RECHECK THEM BY ACCESSING AND USING THE SITE. YOU ACCEPT, WITHOUT LIMITATION OR QUALIFICATION, THE PRESENT TERMS AND CONDITIONS. YOU ALSO AGREE TO SO ACCEPT FUTURE UPDATES AND MODIFICATIONS OF THE TERMS AND CONDITIONS.

1. PAYMENT TERM: Payment terms are net thirty (30) days from date of invoice. If payment is not received by the due date, invoices are considered past due. First due payments will be subject to a service charge of one and one-half percent (1 1/2%) per month or the maximum amount allowed by law, whichever is less.

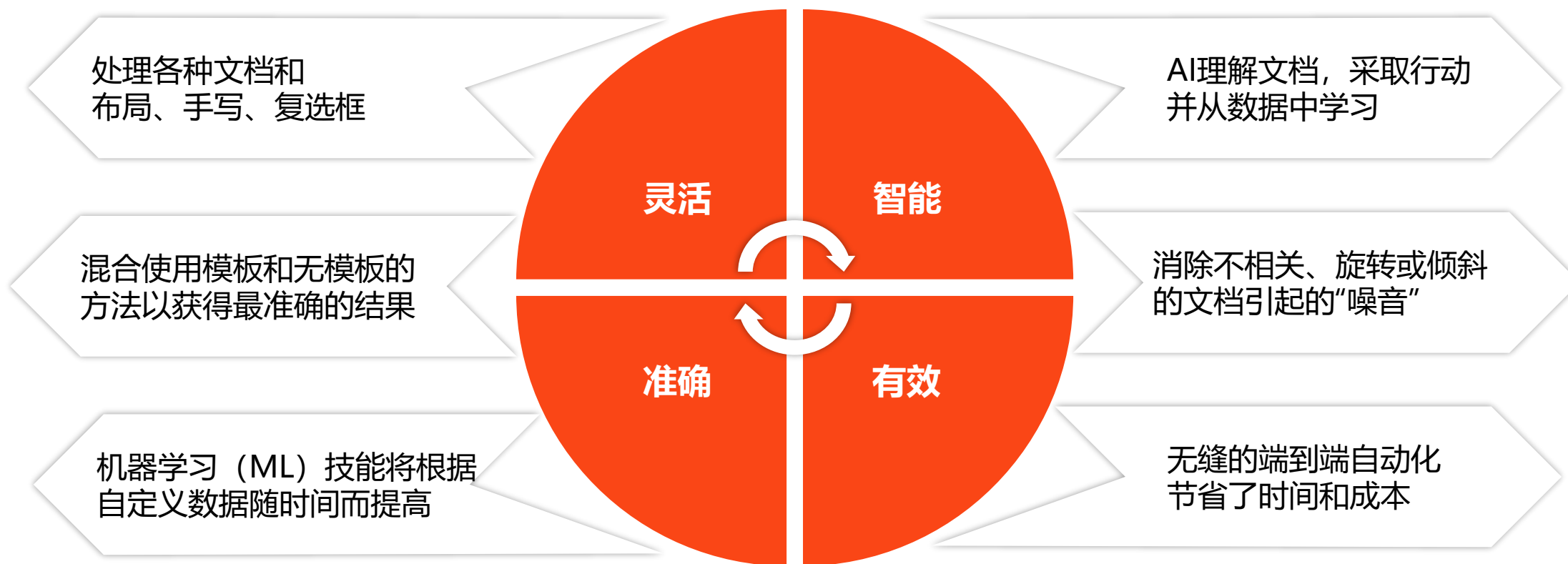
Visa, MasterCard, American Express, Discover, Money Orders, Certified Checks, Company Checks and Personal Checks.

All payments (checks) should be sent to:
Ultra Nectar, Inc.
2734 East 39th St.
Oakland, CA 94601

Your name must be hand imprinted on the check with the correct address and telephone number. Buyer agrees to pay a \$20.00 charge for each returned check and all collection costs, including legal fees, if applicable.

智能处理文档

教机器人如何使用智能拖放技能来处理文档以提取和解释数据



让机器人承担文档处理例程



从**结构化文档**（不同的表格、护照、许可证、时间表等）中提取和解释数据，这些数据可能包含手写文本、签名、复选框。



从包含固定和可变部分（如表格）的**半结构化文档**中提取数据。半结构文档的样本包括发票、收据、采购单、医疗费、银行对账单、水电费等。



分析和提取复杂的**非结构化文档**中的数据，例如各种合同、协议、电子邮件、疾病描述、药物处方、新闻、语音脚本等。



替代**人工验证**并通过使用AI去理解和根据提取的内容采取行动的机器人来处理文档中的数据。

世界上每家公司都在处理文档

通用	金融服务 & 保险	人力资源 (HR)	制造业	公共部门	医疗
• 发票 • 收据 • 采购订单 • 水电费 • 提货单 • 护照 • 许可证	• 应收 & 应付 • 国税局表格 • 贷款申请 • 抵押处理 • 开户和客户准入 • 索赔处理 • 供应商准入 • 合规相关流程	• 员工入职 • 建立筛选 • HR记录处理	• 销售订单处理 • 客户零件请求 • 汇款处理	• 移民申请 • 学校申请 • 护照管理申请	• 医疗表格 • 医疗账单 • 健康记录 • 药物处方

自动化文档处理可以使得.....



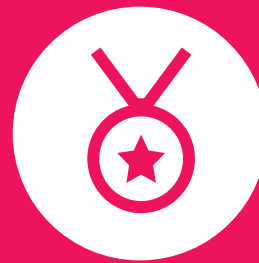
生产率提升

自动执行高度手动的文档处理任务，以提高生产率



效率更高

借助易于部署和维护的自动化，节省了在文书工作上的时间和费用



更好的客户体验

降低错误风险并缩短响应时间，以提供更好的客户体验

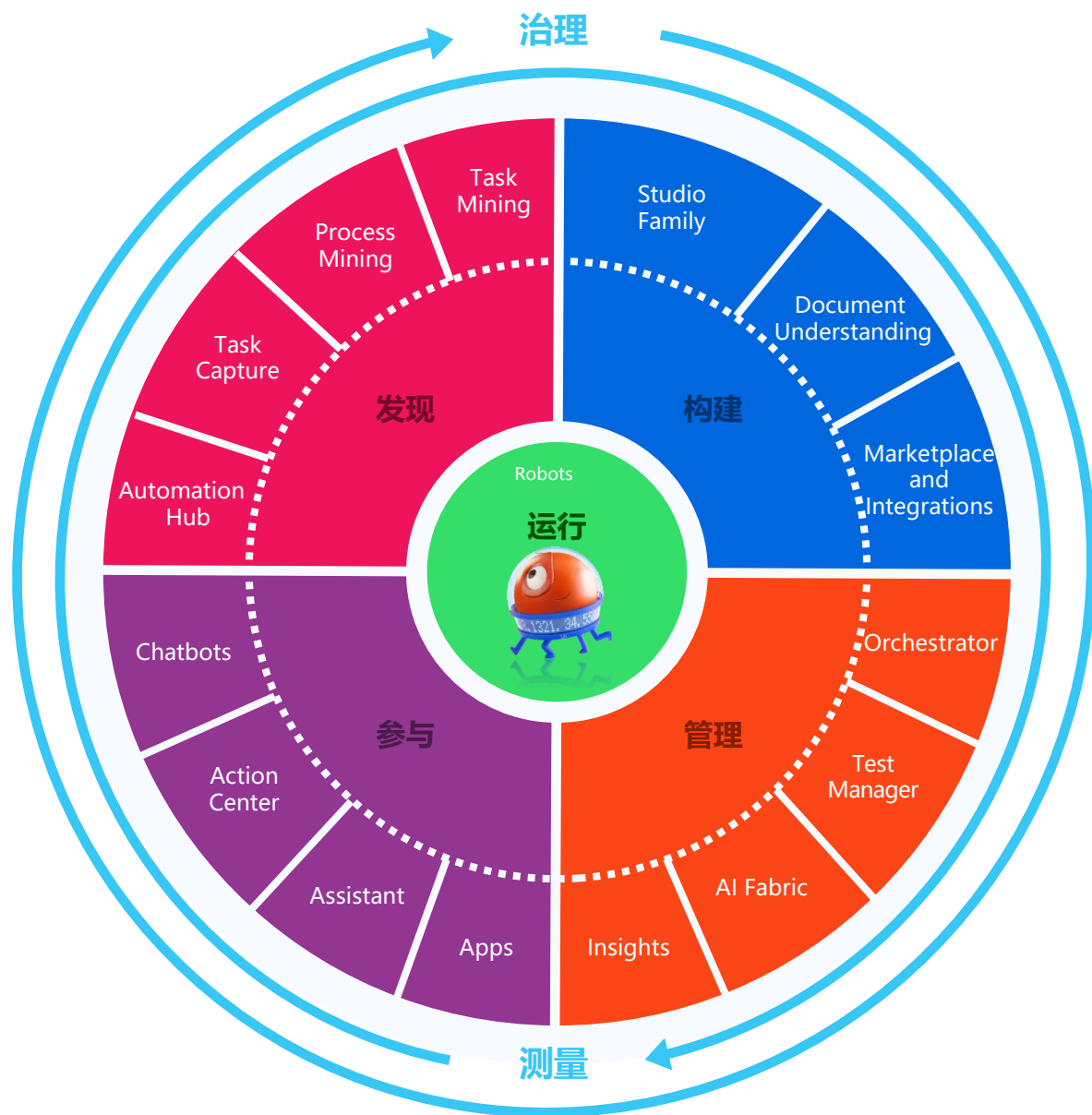


更幸福的员工

帮助员工摆脱日常琐事，专注于高价值任务

2 深入探索 UiPath文档理解

UiPath文档理解和RPA平台



AI Fabric

- 开箱即用的预训练模型
- 引入您自己的模型 – 自定义或第三方
- 重新训练模型



Action Center

- 人工验证



Studio



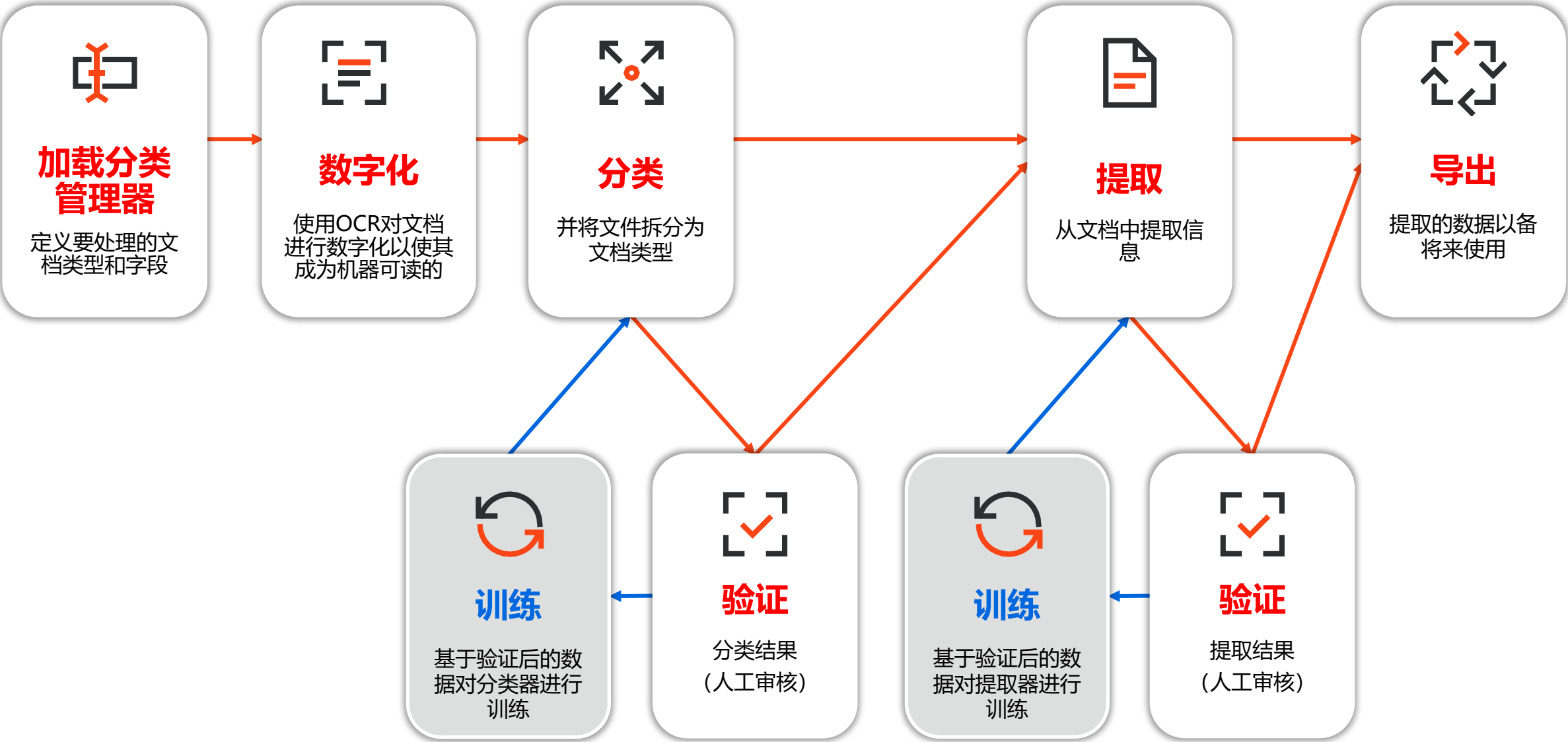
Robots



Orchestrator

- 核心RPA工具

文档理解框架





加载分类管理器



UiPath.IntelligentOCR.Activities 通过 UiPath Core activities that enable the usage of a complete document processing framework, from taxonomy definition, digitization, document classification,...



管理程序包



管理实体



分类管理器



录制



屏幕抓取



数据

UiPath Taxonomy Manager

Document Types

Any Group + Any Category

Search by name

+ Add New Document Type

Click on a document type to edit it.

- Semi-Structured Documents
 - Generalized ML Models
 - Purchase Orders
 - Receipts
 - Invoices
- Structured Documents
 - IRS Forms
 - IRS 4506-T
 - IRS W-9

Document Type Details

Document Type ID: Semi-StructuredDocuments.GeneralizedMLModels.Invoices

Name
Invoices

Group
Semi-Structured Documents

Category
Generalized ML Models

Document Type Code
Optional value

Fields

- Date
- Name
- Invoice No.
- PO No.
- Billing Address
- Shipping Address
- Vendor Address
- Currency
- Tax Amount
- Total

+ New Field

分类管理器一开始仅使用一次，用于定义要处理的文档的集合。

此外，您可以描述要提取的数据。



使用OCR数字化文档中的文本

COMMON

APPLICATION TYPE: ☒ NEW ☐ RENEWAL

DATE OF APPLICATION: 11/06/2020

DATE OF INCORPORATION / FORMATION: 01/01/2021

PLACE OF INCORPORATION / FORMATION: 01/01/2021

ENTRY CONSTITUTION TYPE: ☒ INDIVIDUAL ☐ COMPANY

CIN: 0109412

1. ENTITY DETAILS* (Please refer to the back of the form)

NAME OF THE ENTITY (IN BLOCK LETTERS): T.E.S.

DATE OF COMMENCEMENT OF BUSINESS: 01/01/2021

DATE OF INCORPORATION / FORMATION: 01/01/2021

PLACE OF INCORPORATION / FORMATION: 01/01/2021

ENTRY CONSTITUTION TYPE: ☒ INDIVIDUAL ☐ COMPANY

CIN: 0109412

2. PROOF OF IDENTITY (POI)* (R)

☒ CERTIFICATE OF INCORPORATION / FORMATION
☒ OFFICIALLY VALUED DOCUMENTS
☐ MEMORANDUM AND ARTICLE OF ASSOCIATION

3. DETAILS OF RELATED PERSONS
(An Annexure B to be filled for each related person)

NUMBER OF RELATED PERSONS: 1

NUMBER OF BENEFICIAL OWNERS: 1

4. PROOF OF ADDRESS (POA)* (R)

3.1 CURRENT / PERMANENT OVERSEAS ADDRESS

☒ REGISTERED OFFICE ADDRESS BY REGISTRATION
ADDRESS TYPE: ☐ RESIDENTIAL / ☒ BUSINESS

PROOF OF ADDRESS: ☒ CERTIFICATE OF INCORPORATION / FORMATION
LINE 1: STRYIS
LINE 2:
LINE 3:
DISTRICT: FRANKLIN
STATE / UT NAME CODE:
3.2 CORRESPONDENCE / LOCAL ADDRESS

☐ SAME AS CURRENT / PERMANENT ADDRESS TYPE: ☐ RESIDENTIAL / ☒ BUSINESS

PROOF OF ADDRESS: ☐ CERTIFICATE OF INCORPORATION / FORMATION
LINE 1:
LINE 2:
LINE 3:
DISTRICT:
STATE / UT NAME CODE:
10. PAYMENTS TO CONTRACTORS

Name of person or company the fee is payable to: Name 1

Address: Ukraine, Kyiv, Kolessa str, 3

Average or actual amount paid per month: \$ 20

Do you deduct PAYG tax from this payment? No ☒ Yes ☐

Go to next question

CS4261.1504

Move-For-You Co.
20800 ALMADEN AVE, SUITE 404
SAN JOSE, CA 95120-0520
T: +1 425 555 9876
F: +1 425 555 3456
E: billing@moveforyou-co.com
www.moveforyou-co.com

1 February 2020 **PO: NP74006735**

Bill To: Tony Tzeng
12345 Mango Lane
Seattle, WA 98108

Invoice No: 456200-TZE1

INVOICE DETAILS	FEE
Packing services	\$1,282.00
Storage fees (1 month)	\$1,884.00
House move (white-glove service)	\$5,320.00
Vehicle storage and transport	\$5,186.00
Sales tax 10%	\$1,367.20
Total Fee including Tax	\$15,039.20

PAYABLE WITHIN 15 DAYS OF RECEIPT

Methods of payment

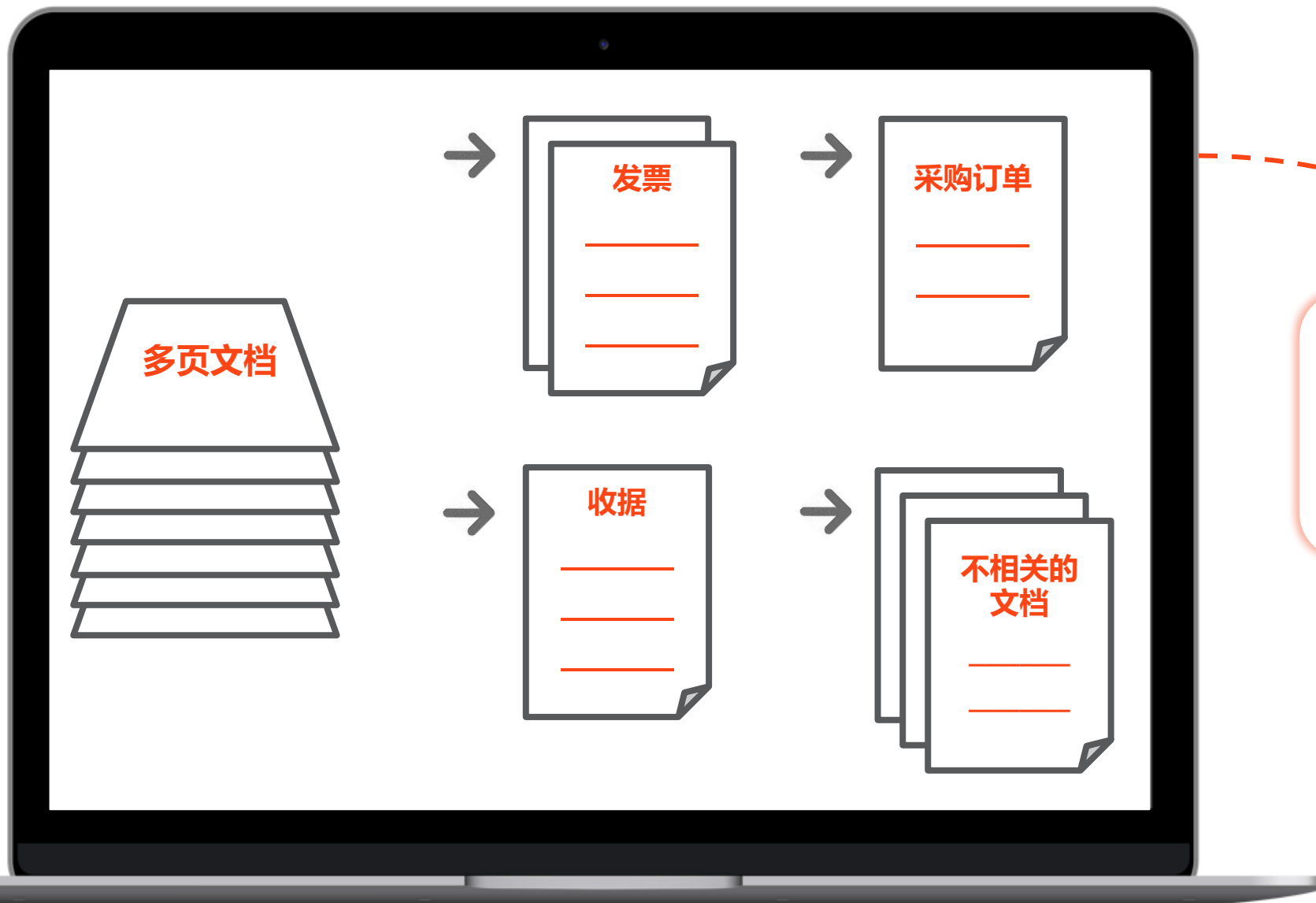
Personal Check: Move-For-You LLC

Wire Transfer: BigBank Co., Account 123456789-0987ABC

Move-For-You Co.
We move so you don't need to move
PO: NP74006735
1 February 2020
PAYABLE WITHIN 15 DAYS OF RECEIPT
20800 ALMADEN AVE, SUITE 404
SAN JOSE, CA 95120-0520
T: +1 425 555 9876
F: +1 425 555 3456
E: billing@moveforyou-co.com
www.moveforyou-co.com
Bill To:
Tony Tzeng
12345 Mango Lane
Seattle, WA 98108
INVOICE DETAILS
Packing services
Storage fees (1 month)
House move (white-glove service)
Vehicle storage and transport
Sales tax 10%
Total Fee including Tax
FEE
\$1,282.00
\$1,884.00
\$5,320.00
\$5,186.00
\$1,367.20
\$15,039.20
Methods of payment
Personal Check: Move-For-You LLC
Wire Transfer: BigBank Co., Account 123456789-0987ABC
Invoice No: 456200-TZE1



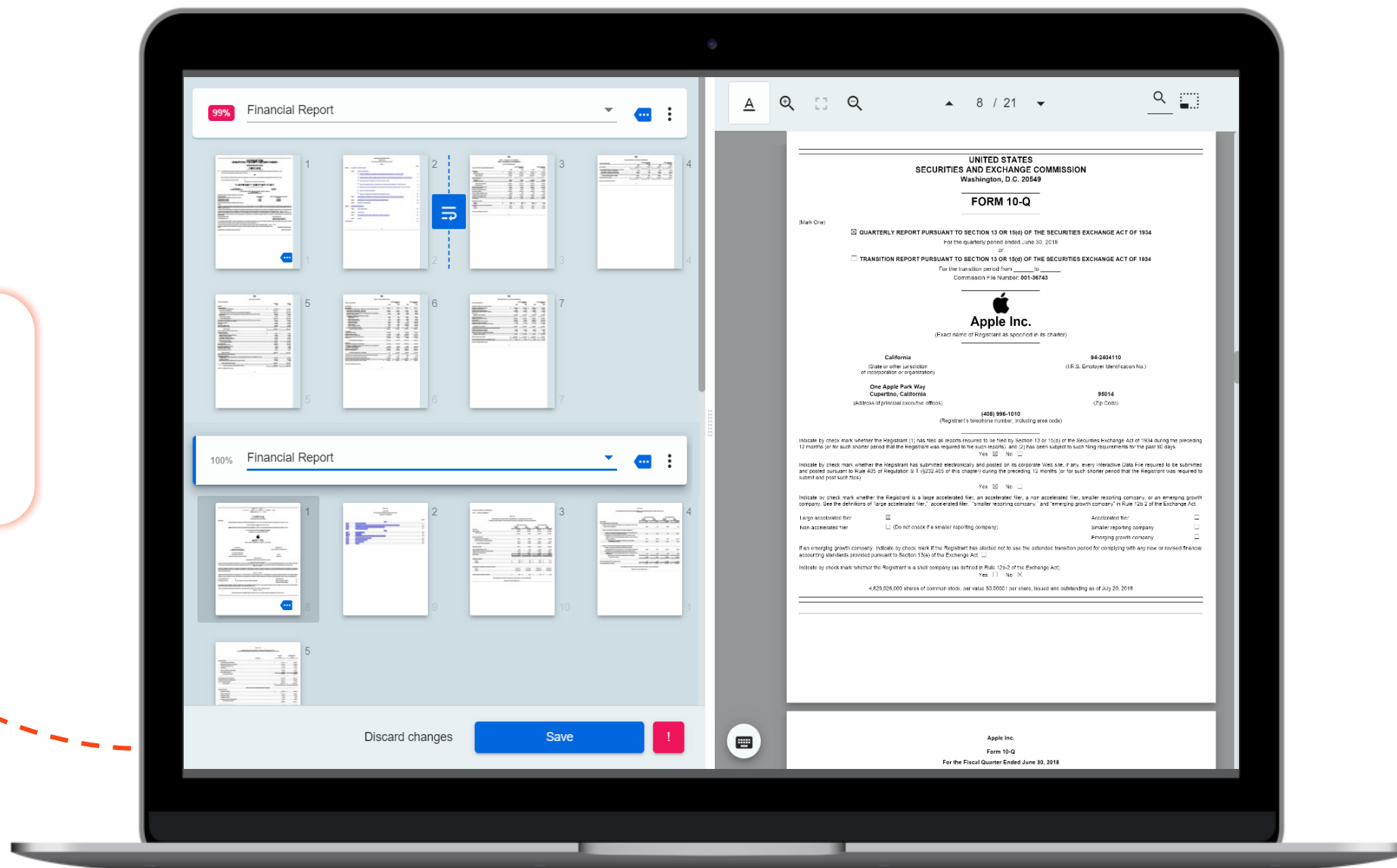
分类和拆分文档



被扫描到一个文件中的文档不是问题 - 由于**分类器**，机器人可以识别文档类型并拆分文件以相应地处理文档。

验证文档的分类

分类站点用于检查、更正和确认文档分类和拆分的结果。





从文档中提取信息

Data Extraction Scope

"Invoice (10).png"

Regex Based Extractor
 Configure Expressions

Form Extractor
 Manage Templates

Intelligent Form Extractor
 Manage Templates

Invoices ML Extractor

Purchase Orders ML Extractor

Configure Extractors

Ui Configure Extractors

Configure Extractors

Configure which extractors you want to apply to each document type and field.
To activate extractors for certain fields, check the appropriate boxes in the configurator.
For extractors that use your defined taxonomy for configuration and data extraction, entering the unique IDs for document types and fields is optional.
For extractors that have their own internal taxonomy, provide the internal taxonomy unique IDs for both document types and fields for which the extractor will be activated.

Document Types and Fields	Regex Based Extractor Minimum Confidence % 90	Form Extractor Minimum Confidence % 90	Intelligent Form Extractor Minimum Confidence % 0	Invoices ML Extractor Minimum Confidence % 90	Receipts ML Extractor Minimum Confidence % 0	Purchase Orders ML Extractor Minimum Confidence % 0
► Purchase Orders	☐	☐	☐	☐ N/A	☐ N/A	☒ N/A
► Receipts	☐	☐	☐	☐ N/A	☒ N/A	☐ N/A
▼ Invoices	☒	☒	☐	☒ N/A	☐ N/A	☐ N/A
Date	☒	☒	☐	☒ date	☐	☐
Name	☐	☒	☐	☒ name	☐	☐
Invoice No.	☐	☒	☐	☒ invoice-no	☐	☐
PO No.	☒	☒	☐	☒ po-no	☐	☐
Billing Address	☐	☒	☐	☒ billing-addr	☐	☐
Shipping Address	☐	☒	☐	☒ shipping-addr	☐	☐
Vendor Address	☐	☒	☐	☒ vendor-addr	☐	☐
Currency	☒	☒	☐	☒ currency	☐	☐
Tax Amount	☐	☒	☐	☒ tax	☐	☐
Total	☐	☒	☐	☒ total	☐	☐

Save Cancel

您可以轻松配置数据提取，以为每个字段选择最合适的提取器。结合使用**基于规则**和**基于模型**的方法，以确保平稳、准确地处理不同文档。

数据提取 – 从规则到AI到混合方法



基于规则或基于模板的方法



依赖规则（如正则表达式）和
模板



处理固定格式的结构化数据



确保已知文件的高准确度

机器学习（ML）模型作为无模板方法



预训练的ML模型

- 发票
- 澳大利亚/印度/日本的发票*
- 收据
- 采购订单*
- 水电费*



引进您自己的模型

- 创建新的自定义模型
- 第三方模型



模型再训练

- AI Fabric
- 验证站点

预训练的ML模型

UiPath™ AI Fabric | UtilityBills

Dashboard

Datasets

Data Labeling

ML Packages

Pipelines

ML Skills

ML Logs

ML Packages / Out of the box Packages / UiPath Document Understanding

UiPath Document Understanding



DocumentUnderstanding

Machine Learning model for extracting commonly occurring data points from Semi-structured or Structured documents, including regular fields...



Invoices

Machine Learning model for extracting commonly occurring data points from Invoices, including header fields and line items. Please see more...



InvoicesAustralia

Machine Learning model for extracting commonly occurring data points from Australian Invoices, including header fields and line items. Please see more...



InvoicesIndia

Machine Learning model for extracting commonly occurring data points from Indian Invoices, including header fields and line items. Please see more...



InvoicesJapan

Machine Learning model for extracting commonly occurring data points from Japanese Invoices, including header fields and line items. Please see more...



PurchaseOrders

Machine Learning model for extracting commonly occurring data points from Purchase Orders, including header fields and line items. Please see more...



Receipts

Machine Learning model for extracting commonly occurring data points from Receipts, including header fields and line items. Please see more...



UtilityBills

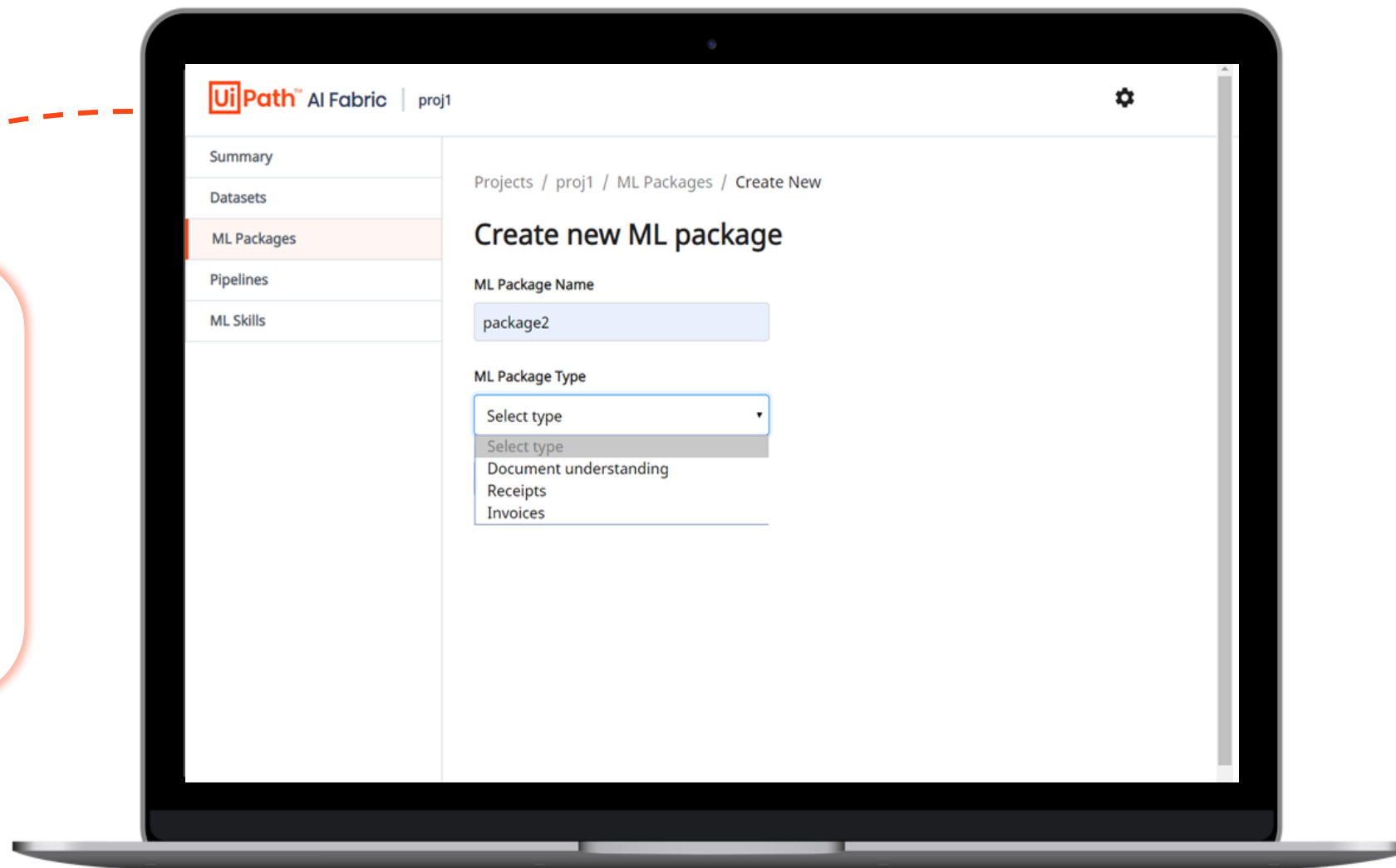
Machine Learning model for extracting commonly occurring data points from Utility Bills, including header fields and line items. Please see more...

利用预先训练的机器学习模型来处理发票、收据、水电费，采购订单（更多模型即将推出）。
引入您自己的模型或第三方模型，并将其合并到您的自动化中。
重新训练模型，以提高其精度！

通过AI Fabric进行机器学习模型训练

您可以使用您的自定义ML模型或重新训练AI Fabric中的现有ML模型，以帮助机器人了解文档的特殊性。

您使用模型的次数越多，它将变得越有效。因此，提取的数据的准确性将随着时间而提高。





验证提取结果

使用验证站点验证提取出来的信息和处理异常。

现在，还可以使用在验证站点中确认或校正的数据来重新训练ML模型。

UiPath Validation Station "MovingInvoice"

96%

12345 Mango Lane Seattle, WA 98108

Address Line 1
12345 Mango Lane

Address Line 2

Address Line 3

City
Seattle

State / County / Province
Washington

Country
United States

Zip Postal Code
98108

123456789

123456789

Discard changes Save 3 / 11 !

Move-For-You Co.

20800 ALMADEN AVE, SUITE 404
SAN JOSE, CA 95120-0520
T: +1 425 555 9876
F: +1 425 555 3456
E: billing@moveforyou-co.com
www.moveforyou-co.com

1 February 2020 PO: NP74006735

Bill To: Invoice No: 456200-TZE1

Tony Tzeng
12345 Mango Lane
Seattle, WA 98108

INVOICE DETAILS	FEE
Packing services	\$1,282.00
Storage fees (1 month)	\$1,884.00
House move (white-glove service)	\$5,320.00
Vehicle storage and transport	\$5,186.00
Sales tax 10%	\$1,367.20
Total Fee including Tax	\$15,039.20

训练 分类器和提取器

让分类器和提取器从分别在
分类站点和验证站点中校正
和验证的数据中学习。

The screenshot displays two overlapping configuration windows from the UiPath Studio interface. The background window is titled 'Train Classifiers Scope' and contains the following fields: 'Document Path' (doc), 'Human Validated Data' (Enter a VB expression), 'Document Text' (text), 'Human Validated Classification Data' (validatedClassification), and 'Document Object Model' (dom). It also features a 'Taxonomy' dropdown set to 'taxonomy' and an 'Intelligent Keyword Classifier Train' section with a 'Learning File Path' field (containing a file icon and the text 'Learning File Path. Text must be q...') and a 'Manage Learning' button. The foreground window is titled 'Train Extractors Scope' and includes: 'Document Path' (doc), 'Human Validated Data' (confirmedExtraction), 'Document Text' (text), and 'Document Object Model' (dom). It also has a 'Machine Learning Extractor Train' section with an 'Endpoint' field (containing three dots), an 'ML Skill' dropdown (showing 'No ML Skill found. Please retry.' with a refresh icon), and an 'Output Folder' field (containing three dots). Both windows have 'Configure Classifiers' and 'Configure Extractors' buttons at the bottom.

导出提取的数据

→ 导出数据以进一步使用或自动化。

例如：

- 到Excel电子表格
- 发送电子邮件
- SAP等



充分利用文档理解生态

光学字符识别 (OCR)	结构化文档	半结构化文档	非结构化文档	自然语言处理 (NLP)
• UiPath Document OCR • Kofax Omnipage • ABBYY FineReader • Google OCR • Microsoft OCR • Tesseract	• UiPath Form Extractor & Intelligent Form Extractor • ABBYY FlexiCapture	• UiPath Machine Learning Extractor • ABBYY FlexiCapture Distributed / for Invoices • Hyperscience • Ephesoft • Vidado • Rossum • Omnius • Microsoft Form Recognizer • Amazon Textract	• Indico • SortSpoke • Botminds AI • Xtracta • Contract Wrangler	• Expert System • Amazon Comprehend • Stanford NLP Group

3 文档理解 Demo演示及案例分享

场景示例： 抵押贷款包结账后审计



财务采购发票处理



IT托管服务和系统集成

Location: Ireland

Size: SME

挑战

- Evros 财务团队每周花费20个小时手动处理每月多达1,600到1,800张的采购发票。
- 手动，重复且耗时的文档处理导致整个团队时间的浪费。
- 他们想要一个可以随业务扩展的自动化解决方案。

解决方案

- 将业务流程映射为自动化方法
- 为主要供应商建立和训练数据模型
- 建立和配置UiPath文档理解
- 设计和实现与Outlook和Microsoft NAV（Evros财务系统）的集成
- 通过机器学习（ML）持续改进
- 必要时需要人工干预

21,000

每年大约21,000张发票

80%

节省80%的时间

6-8 weeks

花费在再训练上，以将准确性从60%提高到80%

80%

随着时间的推移，自动重新训练的准确性仍在不断提高

重要注意事项

- 每个客户都有各自的发票模板
- 在流程开始时还有其他一些文档类型需要过滤掉（例如贷方通知单等）
- 每张发票都经过打印、审核和存储以供审计，这是手工过程的一部分。需要考虑一种新的存储方法

采购订单信用调整的自动化

挑战

- 总部位于纽约的医疗保健提供商有500多个供应商，需要频繁处理订单的退货/取消，因此每月在PeopleSoft ERP中处理800多个贷项通知单（冲销发票）
- 会计团队手动处理供应商贷项通知单并根据ERP中的采购订单进行贷项调整
- 涉及许多业务验证，包括文档和ERP之间的数据比较
- 平均而言，手动处理大约需要8分钟

文档类型

- 65% 半结构化扫描
- 20% 结构化
- 5% 半结构扫描手写
- 10% 数字化PDFs

为什么自动化？

- 及时调整采购订单（PO）的信贷额度可实现高效的营运资金
- 通过AI自动化避免人为错误以解决更复杂的情况

为什么UiPath文档理解？

- 开箱即用的ML模型用于发票处理
- 与OCR引擎的内置集成，包括UiPath Document OCR和OmniPage OCR
- 用于人工验证的行动中心

医疗保健提供商

位置：USA

200-250 训练数据集

代表生产人口的要素，是构建成功的ML模型的核心

SME 教育/培训

需要SMEs学习如何准备数据集

2-3 分钟

用12个元素标记1个文档

~40% 时间节省

和CPU（17小时）相比，在GPU（10小时）中进行了400个文档模型训练，

2-4 周

从首次交付到生产，并具有预期的置信度

文档修订

Location: USA

挑战

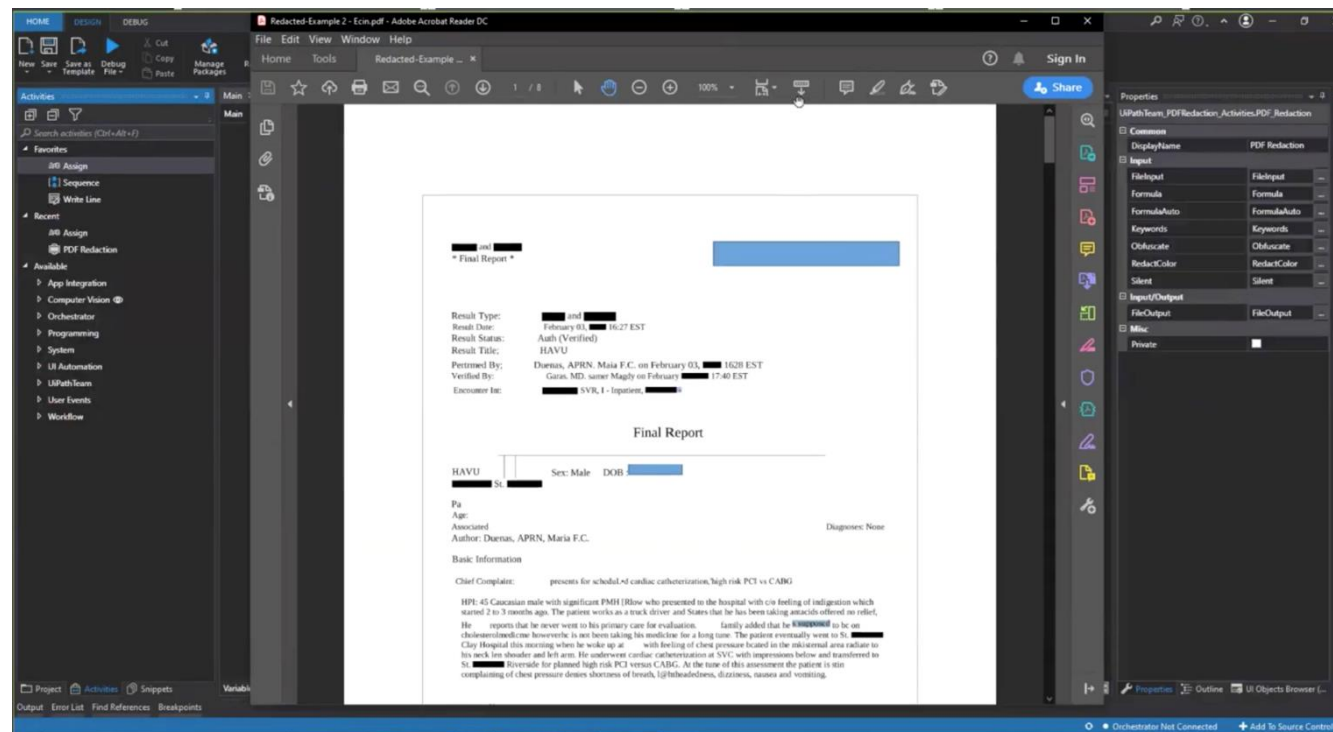
- 客户需要合规性要求，在将文件传送给第三方以验证患者索赔之前，需要对文件进行编辑
- 手动完成耗时的过程

文档类型

- 平均10页的文档

重要注意事项

- 优先处理覆盖80%文档的前5种文档类型
- 10-11 PII字段 + 被视为修订范围的医学术语列表
- 人工要审查100%的修订，从头开始并逐渐减少（从有人参与的协助过渡到STP路径）



为什么UiPath文档理解？

- 自动编辑PII和PHI覆盖范围
- 机器学习持续改进
- 位于美国的生产支持部门
- 必要时需要人工验证

如何试用?

JC

Home

Actions

Apps

Processes

Data Service

Admin

UiPath™ Automation Cloud™

Tenants

Users and Groups

Licenses

Audit Logs

Organization Settings

Licenses Community Plan

Subscription Expiry: Nov 14, 2022

Users

Robots & Services

Unattended Robot

The maximum number of Unattended Robots that are able to execute back-office processes simultaneously.

1

Data Service Units

Each Data Service Unit grants you 100 MB of data storage, 500 MB of attachment storage and 1,000 API calls/day.

1

Computer Vision

Throughput limit (megapixels/min)

API Key: 2a8VYQ3Lh8Ei18novvN+N...

30

[Generate new](#)

[Copy API key](#)

Testing Robot

The maximum number of Testing Robots that are able to run simultaneously.

1

Document Understanding

Documents limited to 2 pages and 4MB

API Key: vjikkjyseioIf0YXvZJwif/XZ...

177Bysta

[Generate new](#)

[Copy API key](#)

Request Enterprise Trial

如何找到更多资料?



- [UiPath Document Understanding webpage](#) ← for **trial** and more resources
- [How AI Can Continuously Improve and Scale Automations](#) (webinar with customers sharing case studies)
- [Guide on Document Understanding](#) (white paper)
- [Academy training](#)
- Documentation on [Document Understanding framework & extractors](#) and [ML model training via AI Fabric](#)
- [AI & RPA](#) webpage

Q & A

崔静

UiPath售前顾问

Jing.Cui@uipath.com

