ComPDFKit

数据提取助力AI

创新科技 引领未来

01.

目录

ComPDFKit

ComPDFKit

01. AI大语言模型训练的挑战

The Challenges for Training AI Large Language Models

01. AI 大语言模型训练数据来源

自 ChatGPT 问世以来，大语言模型在解决自然语言处理任务、改善客户服务、增强营销策略等方面的潜力被广泛认可，因此，训练高质量大语言模型的需求与日俱增。

训练 AI 大语言模型所需的海量数据来源包括但不限于网页、社交媒体帖子、电子版书籍、公司内部的文档、邮件、聊天记录等。这些数据分为结构化和非结构化两类。若无法对非结构化数据进行有效解析，其潜在的巨大价值将无法得以实现。在非结构化数据中，PDF 文档占据主要比重，因此有效处理 PDF 文档对于管理其他类型的非结构化文档具有重要意义。

01. AI大语言模型训练的挑战

在训练 AI 大语言模型时，

非结构化数据

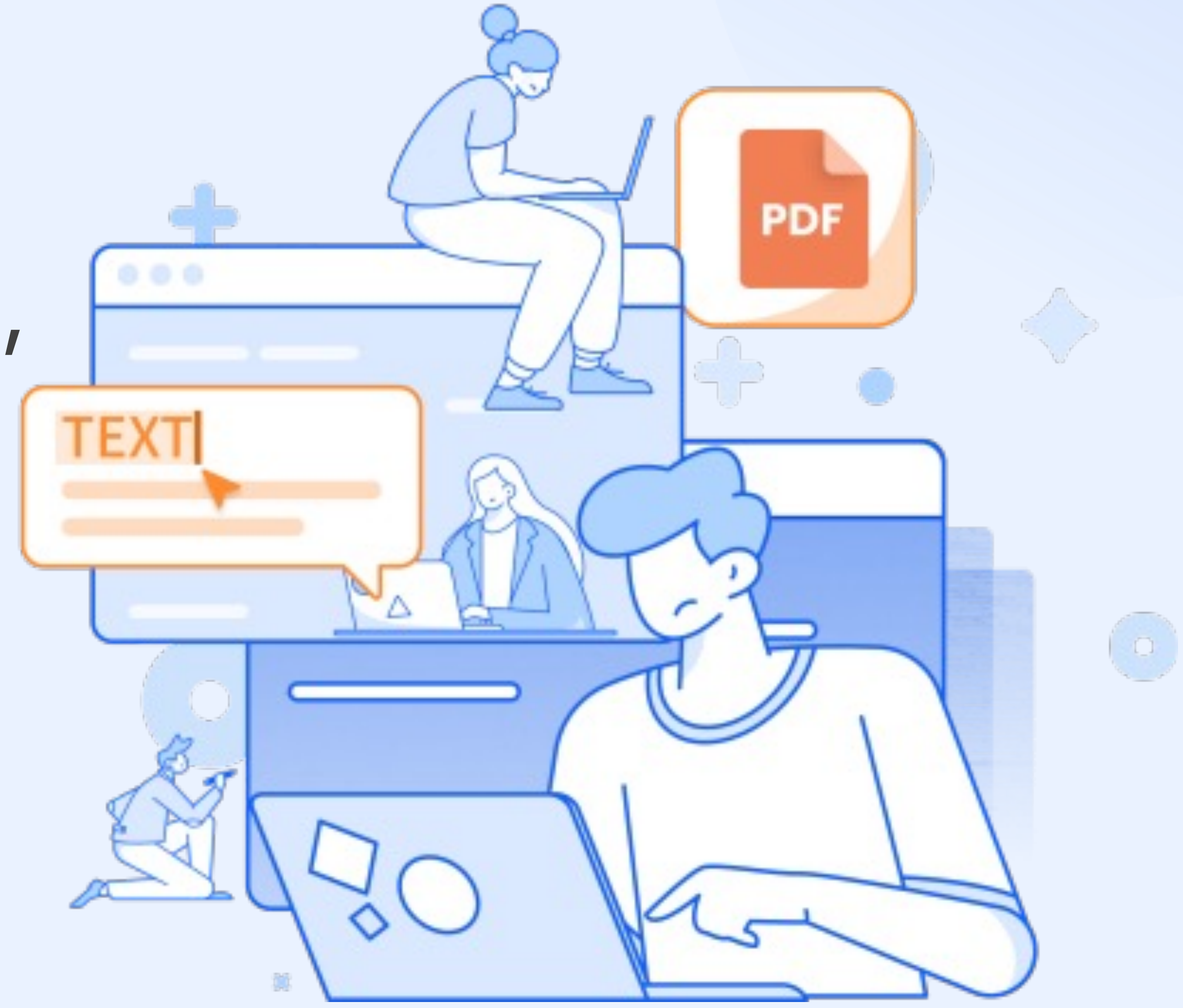
是否让您感到苦恼?

01. AI大语言模型训练的挑战

手动提取 PDF 文档的数据，

既耗费时间又耗费人力，

准确率还低?



01. AI大语言模型训练的挑战

使用传统算法识别
和提取扫描档PDF，
版面乱、效率低?

01. AI大语言模型训练的挑战

担心数据安全和隐私，
需要训练企业内部的
AI 大语言模型？

ComPDFKit

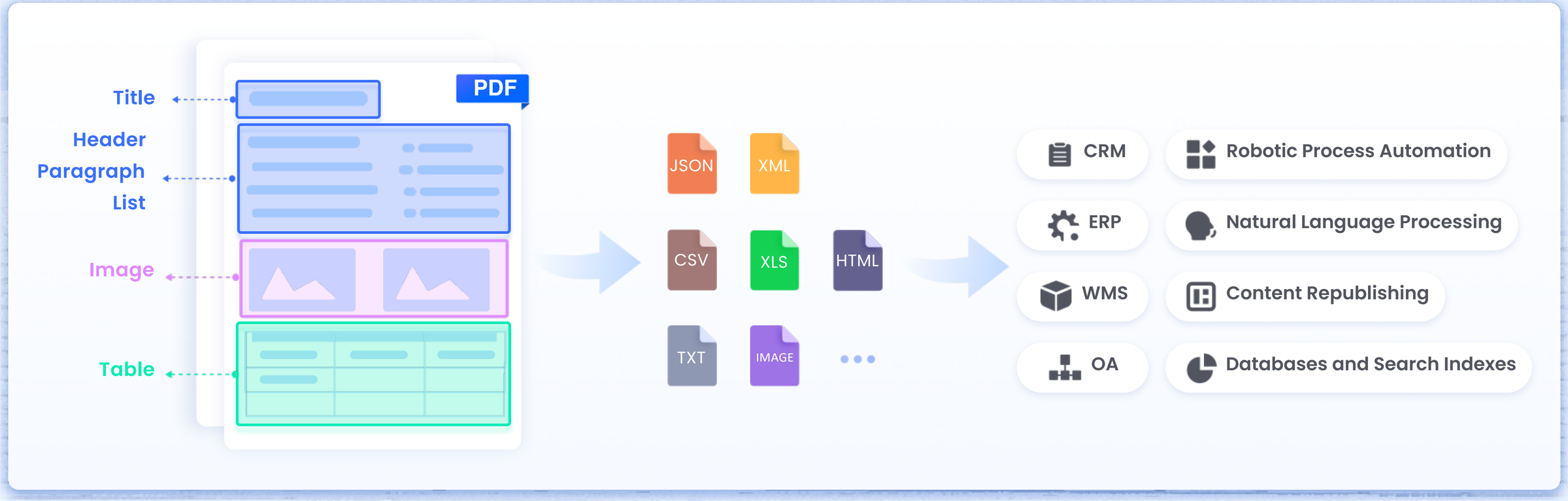
02. ComPDFKit 的应对措施

ComPDFKit’s Solutions

02. ComPDFKit 助力数据提取

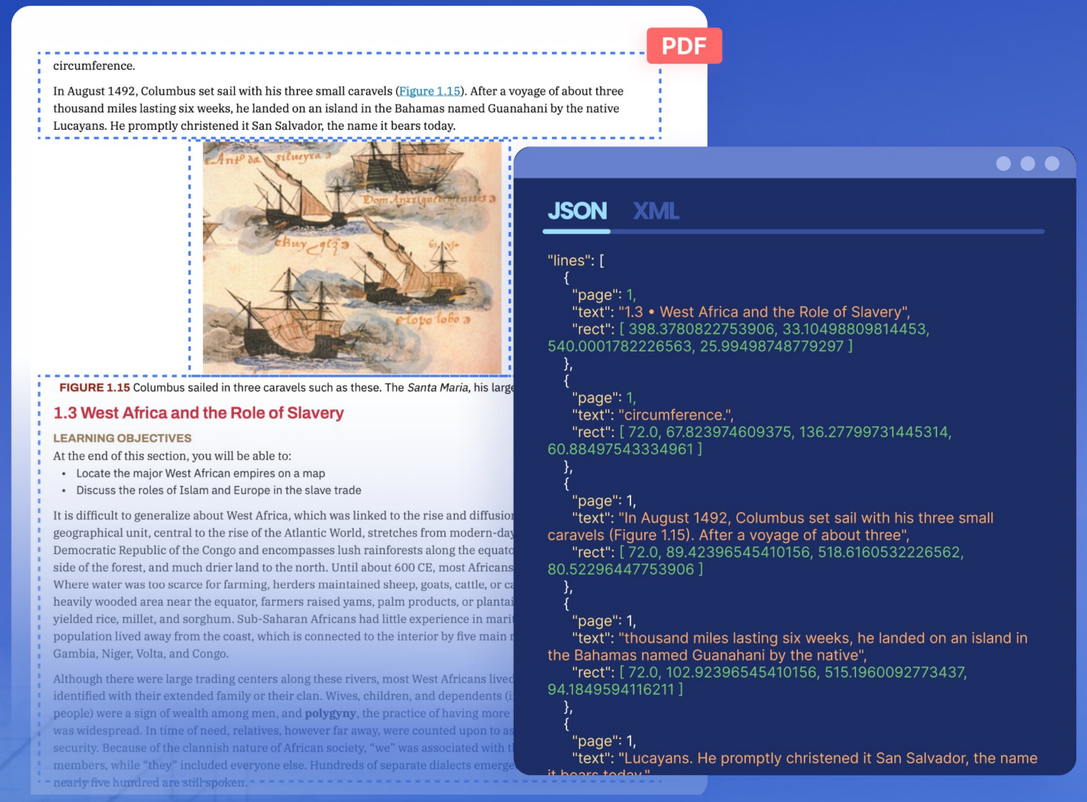
ComPDFKit

以 PDF 文档为代表的非结构化数据，包含各种页面元素、样式等，其结构化的过程涉及文本提取、图像识别以及表格识别。无论是训练 AI 大语言模型，还是建立数据库，抑或是将数据录入系统，将非结构化文档结构化无疑是众多开发者遇到的一大难点。因此，PDF 文档的多样性和复杂性使准确提取数据变得异常困难且具有挑战性。



ComPDFKit

准确度



02. 措施一：提升准确度

- 准确度是数据提取的基础。数据提取的准确度受多方面因素影响，其中包括复杂的语境、模型的复杂性、数据的多样性、标注的准确性、数据中的不规则结构、错误、缺失值或噪声。此外，缺乏足够的领域知识、数据过拟合或不足等问题也会对数据提取的准确度产生负面影响。

- ComPDFKit 结合 Document AI 与传统算法，提高了识别和提取 PDF 文档及扫描档的准确度。精确的数据为用户提供了分析和决策的坚实基础，有利于促进业务运营。

ComPDFKit

02. 措施二：识别版面布局

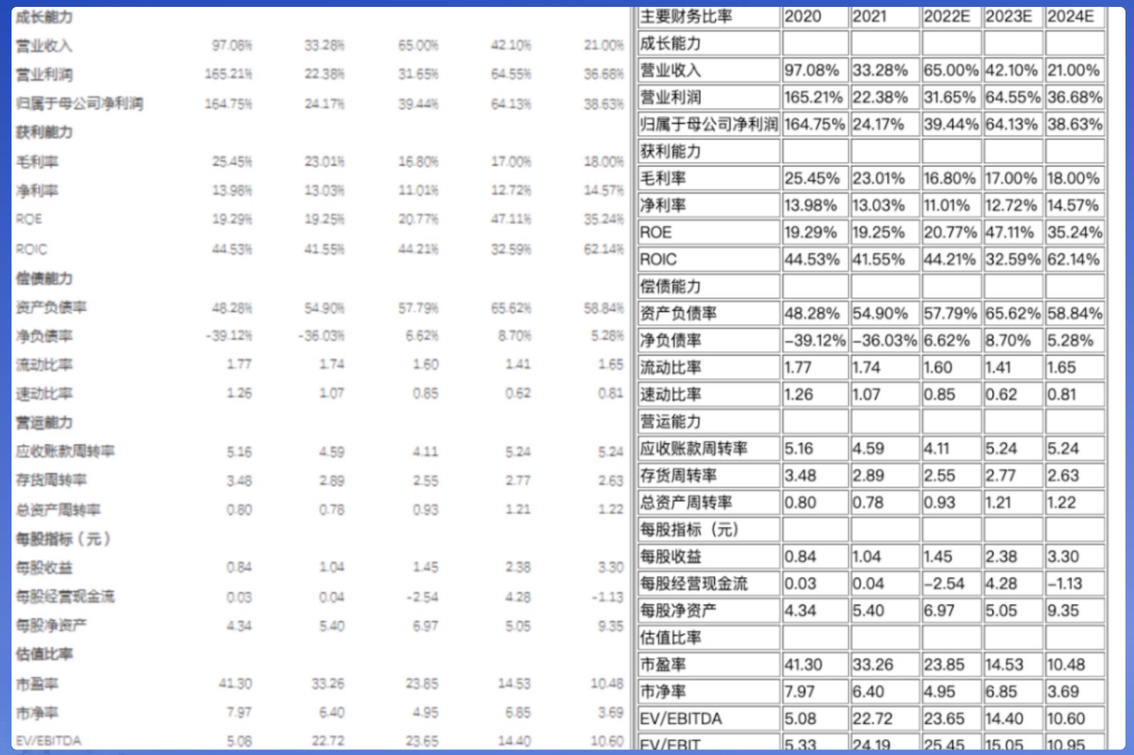
- 因格式和布局复杂、文本图像混合、语言和字体多样、图片质量和分辨率限制、以及复杂噪声等，准确识别 PDF 文档的结构和版面难度较大。正确提取 PDF 文档的版面有利于确保文档中信息的完整性、一致性以及易读性。

- ComPDFKit 使用通用模型和行业模型，针对不同行业、不同类型的文档进行版面识别，准确分析 PDF 的结构、文本、段落、图表、页眉、页脚和表格格式，可以保持文档的原始呈现形式，为进一步的数据提取打下基础。

版面识别

ComPDFKit

表格结构



成长能力					
营业收入	97.08%	33.28%	65.00%	42.10%	21.00%
营业利润	165.21%	22.38%	31.65%	64.55%	36.68%
归属于母公司净利润	164.75%	24.17%	39.44%	64.13%	38.63%
获利能力					
毛利率	25.45%	23.01%	16.80%	17.00%	18.00%
净利率	13.98%	13.03%	11.01%	12.72%	14.57%
ROE	19.29%	19.25%	20.77%	47.11%	35.24%
ROIC	44.53%	41.55%	44.21%	32.59%	62.14%
偿债能力					
资产负债率	48.28%	54.90%	57.79%	65.62%	58.84%
净负债率	-39.12%	-36.03%	6.62%	8.70%	5.28%
流动比率	1.77	1.74	1.60	1.41	1.65
速动比率	1.26	1.07	0.85	0.62	0.81
营运能力					
应收账款周转率	5.16	4.59	4.11	5.24	5.24
存货周转率	3.48	2.89	2.55	2.77	2.63
总资产周转率	0.80	0.78	0.93	1.21	1.22
每股指标（元）					
每股收益	0.84	1.04	1.45	2.38	3.30
每股经营现金流	0.03	0.04	-2.54	4.28	-1.13
每股净资产	4.34	5.40	6.97	5.05	9.35
估值比率					
市盈率	41.30	33.26	23.85	14.53	10.48
市净率	7.97	6.40	4.95	6.85	3.69
EV/EBITDA	5.08	22.72	23.65	14.40	10.60

主要财务比率	2020	2021	2022E	2023E	2024E
成长能力					
营业收入	97.08%	33.28%	65.00%	42.10%	21.00%
营业利润	165.21%	22.38%	31.65%	64.55%	36.68%
归属于母公司净利润	164.75%	24.17%	39.44%	64.13%	38.63%
获利能力					
毛利率	25.45%	23.01%	16.80%	17.00%	18.00%
净利率	13.98%	13.03%	11.01%	12.72%	14.57%
ROE	19.29%	19.25%	20.77%	47.11%	35.24%
ROIC	44.53%	41.55%	44.21%	32.59%	62.14%
偿债能力					
资产负债率	48.28%	54.90%	57.79%	65.62%	58.84%
净负债率	-39.12%	-36.03%	6.62%	8.70%	5.28%
流动比率	1.77	1.74	1.60	1.41	1.65
速动比率	1.26	1.07	0.85	0.62	0.81
营运能力					
应收账款周转率	5.16	4.59	4.11	5.24	5.24
存货周转率	3.48	2.89	2.55	2.77	2.63
总资产周转率	0.80	0.78	0.93	1.21	1.22
每股指标（元）					
每股收益	0.84	1.04	1.45	2.38	3.30
每股经营现金流	0.03	0.04	-2.54	4.28	-1.13
每股净资产	4.34	5.40	6.97	5.05	9.35
估值比率					
市盈率	41.30	33.26	23.85	14.53	10.48
市净率	7.97	6.40	4.95	6.85	3.69
EV/EBITDA	5.08	22.72	23.65	14.40	10.60
EV/EBIT	5.33	24.19	25.45	15.05	10.95

02. 措施三：还原表格结构

- PDF 文档中，由于表格格式复杂、边框不清晰列宽不规则等，识别并提取表格中的数据十分具有挑战性。如果用户手动处理PDF 文档中的表数据，不仅增加了工作负担，同时难以保障数据提取的准确性。

- ComPDFKit 利用 Document AI ，结合自主研发的表格算法专利，准确还原表格结构，按需提取关键信息，提升数据提取中表格识别的准确性，确保所得数据的可靠性。

ComPDFKit

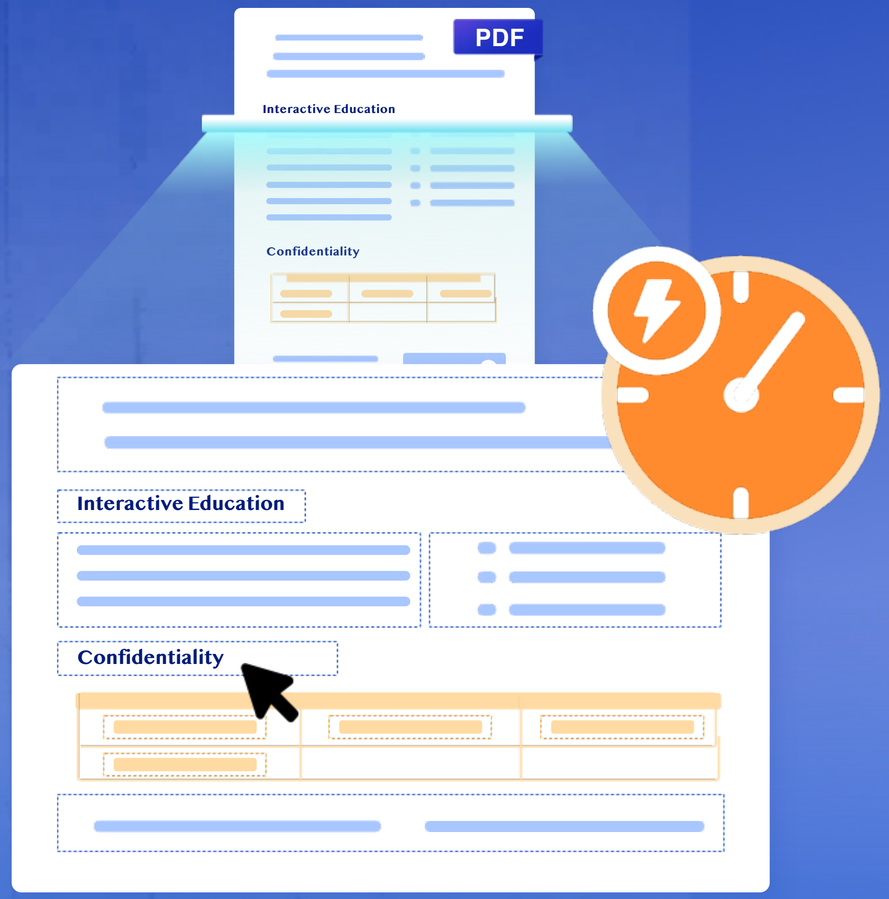
02. 措施四：提取数学公式

- 数学公式的复杂性、多样性、语境依赖性，以及图形表达和格式差异增加了数据提取难度。在面对从大量数据中提取特定数学公式的挑战时，手动检索、识别和提取耗费时间和精力。

- ComPDFKit 采用专用数学模型，运用排版感知技术、注重符号识别以及不同字体和大小的识别，以确保在复杂 PDF 文档中准确理解和提取各种格式的数学公式。这优化了数学公式的提取性能，使从 PDF 文档中提取数学公式方面更加简便、准确。

ComPDFKit

处理效率



02. 措施五：提高处理效率

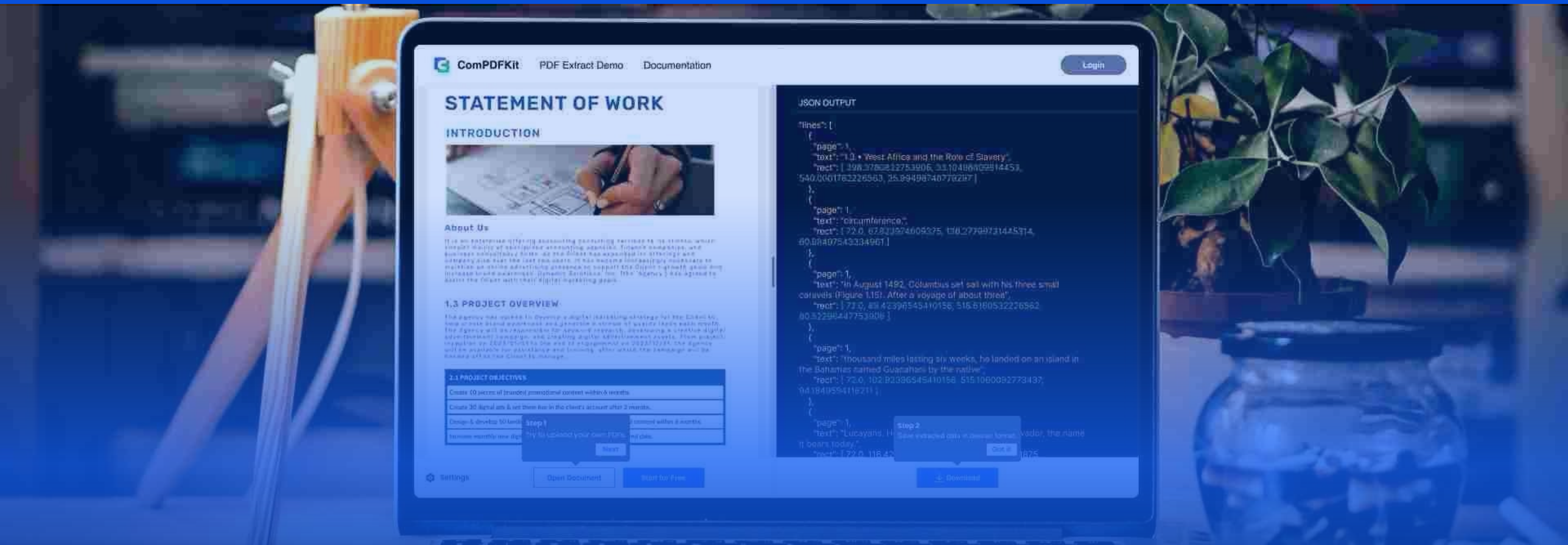
- 数据规模庞大、类型多样、结构复杂、质量参差不齐等问题造成数据提取的效率低下。而人工处理可能导致耗时长、错误率高，限制了数据提取的规模和速度。

- ComPDFKit 运用先进的算法、AI 技术，综合考虑数据的复杂性和多样性，努力克服技术难点，支持 GPU 以优化性能，显著提高了提取速度和准确性，从而提升了整体效率，确保提取过程既准确又高效。

ComPDFKit

03. 产品介绍

Products Introduction



03. 产品优势

ComPDFKit

ComPDFKit 数据提取解决方案致力于为用户提供高效、准确的数据提取服务，满足多样化的需求。

PDF 原生数据提取

字符、单词、字体、表单字段、图像、甚至位置，可以被完全识别并提取为结构化的 JSON/ XML 等格式，以便在后续工作中进行二次处理。

01.

PDF 数据组合分类

按照自然阅读顺序对 PDF 文件中的标题、表格、页眉、页脚和段落进行分类，分析其结构，并保持与原始文档一致的结构连贯性。

02.

AI 助力精准提取

经过深度训练的文档智能（Document AI）在识别元素标识符、位置等方面脱颖而出，显著提升了对 PDF 结构的分析、信息提取和图像分类的准确度，从而实现更加自然的文档阅读体验。

03.

24h 竭诚服务

提供 5*24 小时的专业服务保障及技术支持，快速响应，及时解答你的疑问。您可以通过邮件、微信、或者其他任意方式联系我们！

04.

03. 部署方式

ComPDFKit

数据提取 SDK

ComPDFKit SDK 方案提供离线解决方案，即使在没有网络的情况下，也能够进行数据提取。这一解决方案适用于一些安全性要求高或者需要在无网络环境下进行工作的情况。

数据提取 API

API 方案通过云端进行数据处理，无需在本地存储大量的数据。这降低了客户端的数据储存资源需求，尤其适用于资源受限的设备。同时，无需进行复杂的配置，通过简单的 HTTP 请求集成 API，即可快速集成数据提取功能。

数据提取 Processor

ComPDFKit Processor 是一款为 Linux 平台设计的 SDK，旨提供私有化部署方案，用于转换 PDF 文件。它为开发人员提供了丰富的 API，包括强大的数据提取功能。通过将其部署在您的私有服务器上，以确保数据安全，满足您企业的安全需求。

03. 竞品对比

竞品对比			
	ComPDFKit	**Adobe**	**Apryse**
部署方式	• SDK (Java, .NET, C++) • RESTful • On-premises	• RESTful	• SDK(Java, .NET) • RESTful • CLI
支持格式	• JSON, XML • XLSX, CSV • DOCX • HTML • TXT • PNG, JPG	• JSON	• JSON • XML
提取元素	• 文字 • 图片 • 表格 • 页眉，页脚，标题 • 列表，公式，代码 • 等等	• 文字 • 图片 • 表格	• 文字 • 图片 • 表格

ComPDFKit

是替代 Adobe 和 Apryse 的理想之选

- 部署方式全面，价格优势突出
- 导出格式丰富，满足不同需求
- 自主产权，保障数据安全
- 体积小，准确率高，速度快，兼容性高
- 一对一技术支持，及时解决问题
- 持续更新迭代，支持免费升级
- 授权灵活，便于企业统一管理

ComPDFKit

04. 客户案例

User Cases

04. 客户案例

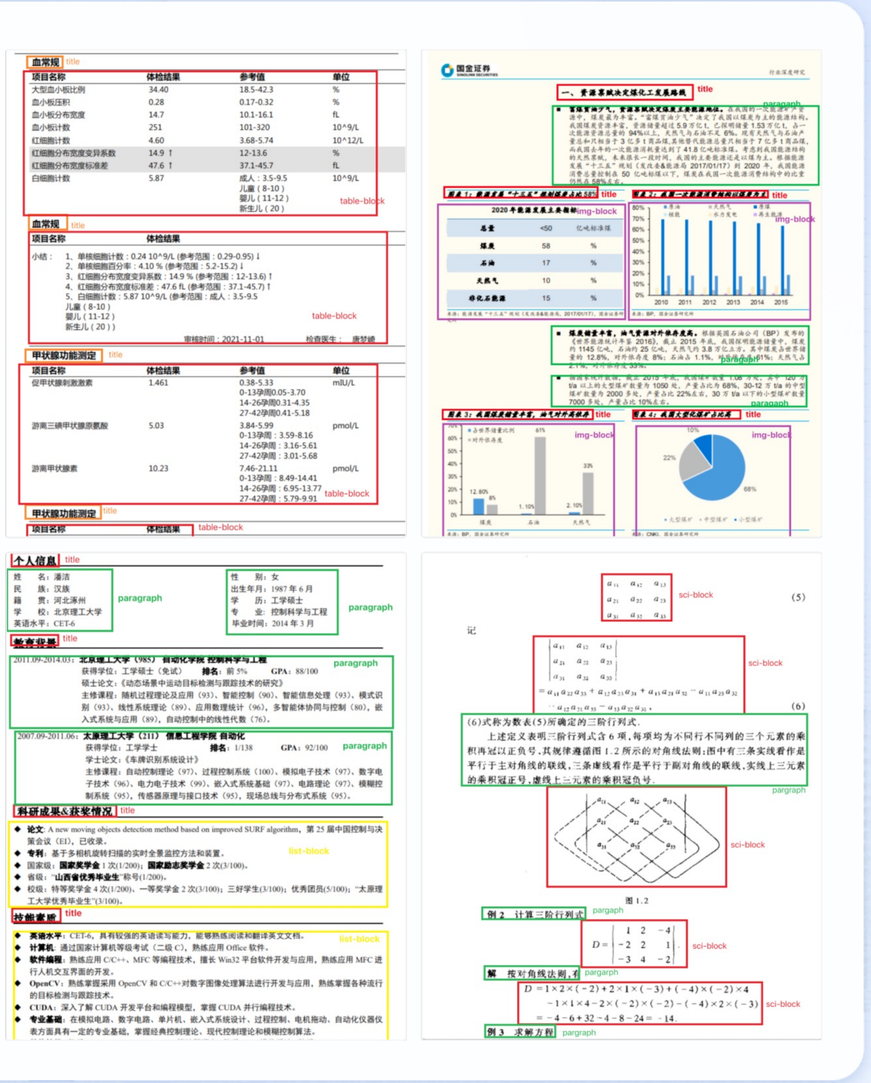
某AI大数据模型公司

挑战：

为了训练和优化机器学习模型，AI大数据模型公司需要对海量的文档进行注释、分类和标记。然而，高昂的标注成本、效率不足、标注一致性难以保证以及隐私安全问题都是重要挑战。

解决方案：

ComPDFKit 帮助该企业集成数据提取功能，创建一个全自动化标注工具，实现数据标注的自动化。这显著提高了工作效率，将原本需要2天完成的标注工作缩短至仅需1小时。这不仅节省了客户大量宝贵的资源，还提升了数据标注质量，扩充了AI训练的数据来源，为其 AI 大语言模型的进一步发展提供了有力支持。



04. 客户案例

某金融证券公司

挑战：

不同的财务报表在排版和格式上都各有差异，这使得准确提取财报表格数据变得相当困难。特别是对于扫描文档，其底稿中的数据难以用常规技术提取，只能依赖人工查找和搬运，导致处理效率较低。这种多样性和复杂性极大地增加了从公司财报中提取数据的挑战。

二、联系人和联系方式

	董事会秘书（信息披露境内代表）	证券事务代表
姓名	李蔚宾	林飞静
联系地址	福州市鼓楼区软件大道89号福州软件园5区5号楼	福州市鼓楼区软件大道89号福州软件园5区5号楼
电话	0591-38509866	0591-38509866
传真	0591-38509869	0591-38509869
电子信箱	boardoffice@foxitsoftware.cn	boardoffice@foxitsoftware.cn

三、信息披露及备置地点

公司披露年度报告的媒体名称及网址	《中国证券报》网址：https://www.cs.com.cn/ 《上海证券报》网址：https://www.cnstock.com/ 《证券时报》网址：http://www.stcn.com/ 《证券日报》网址：http://www.zqrb.cn/
公司披露年度报告的证券交易所网址	上海证券交易所网站（https://www.sse.com.cn）
公司年度报告备置地点	福州市鼓楼区软件大道89号福州软件园5区5号楼

解决方案：

ComPDFKit 针对不同财报表格样式逐一构建算法模型，实现智能的定位和抽取披露内容。利用 AI 技术从财务报表中提取关键数据，打通海量扫描件的数据孤岛。有效的数据帮助该公司更迅速、准确地分析财报数据，为业务决策提供了有力的支持。


```
{
    "info": [
        {
            "pageNumber": 1,
            "tables": [
                {
                    "rowSize": 6,
                    "columnSize": 3,
                    "tableCells": [
                        [
                            {
                                "text": ""
                            },
                            {
                                "text": "董事会秘书（信息披露境内代表）"
                            },
                            {
                                "text": "证券事务代表"
                            }
                        ],
                        [
```

04. 客户案例

某论文平台

挑战：

不同学科的论文存在多种图表数据、结构、类型（如线图、柱状图），增加了识别的难度。同时，某些领域的术语和特定上下文的依赖性也使得数据提取更具挑战性。

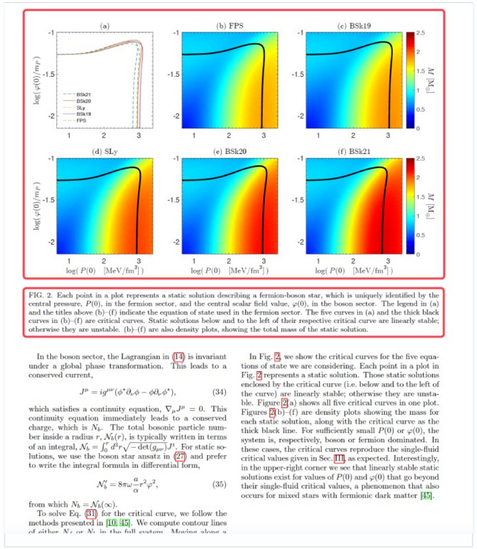


FIG. 2. Each point in a plot represents a static solution describing a fermion-boson star, which is uniquely identified by the central pressure, $P(0)$, in the fermion sector, and the central scalar field value, $\varphi(0)$, in the boson sector. The legend in (a) and the titles above (b)–(f) indicate the equation of state used in the fermion sector. The five curves in (a) and the thick black curves in (b)–(f) are critical curves. Static solutions below and to the left of their respective critical curve are linearly stable, otherwise they are unstable. (b)–(f) are also density plots, showing the total mass of the static solution.

In the boson sector, the Lagrangian in (14) is invariant under a global phase transformation. This leads to a conserved current,

$$J^\mu = ig^{\mu\nu}(\phi^*\partial_\nu\phi - \phi\partial_\nu\phi^*), \tag{34}$$

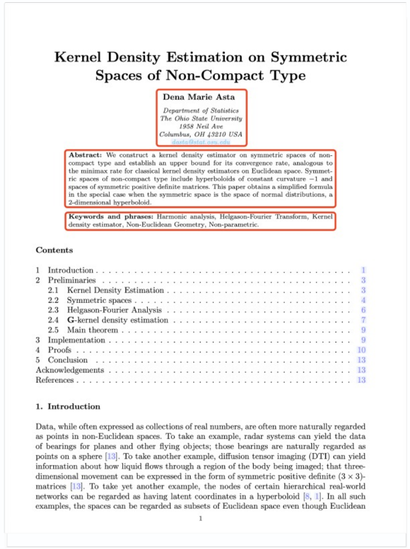
which satisfies a continuity equation, $\nabla_\mu J^\mu = 0$. This continuity equation immediately leads to a conserved charge, which is N_k. The total bosonic particle number inside a radius r, $\mathcal{N}_k(r)$, is typically written in terms of an integral, $\mathcal{N}_k = \int_0^r d^3r\sqrt{-\det(g_{\mu\nu})}J^t$. For static solutions, we use the boson star ansatz in (27) and prefer to write the integral formula in differential form,

$$\mathcal{N}_k' = 8\pi\alpha^{-\frac{a}{\alpha}}r^2\varphi^2, \tag{35}$$

from which $N_k = \mathcal{N}_k(\infty)$.

To solve Eq. (31) for the critical curve, we follow the methods presented in [10, 45]. We compute contour lines of either N_F or N_B in the full system. Moving along a

In Fig. 2, we show the critical curves for the five equations of state we are considering. Each point in a plot in Fig. 2 represents a static solution. Those static solutions enclosed by the critical curve (i.e. below and to the left of the curve) are linearly stable; otherwise they are unstable. Figure 2(a) shows all five critical curves in one plot. Figures 2(b)–(f) are density plots showing the mass for each static solution, along with the critical curve as the thick black line. For sufficiently small $P(0)$ or $\varphi(0)$, the system is, respectively, boson or fermion dominated. In these cases, the critical curves reproduce the single-fluid critical values given in Sec. II, as expected. Interestingly, in the upper-right corner we see that linearly stable static solutions exist for values of $P(0)$ and $\varphi(0)$ that go beyond their single-fluid critical values, a phenomenon that also occurs for mixed stars with fermionic dark matter [45].



Kernel Density Estimation on Symmetric Spaces of Non-Compact Type

Dena Marie Asta

Department of Statistics
The Ohio State University
1958 Neil Ave
Columbus, OH 43210 USA

Abstract: We construct a kernel density estimator on symmetric spaces of non-compact type and establish an upper bound for its convergence rate, analogous to the minimax rate for classical kernel density estimators on Euclidean space. Symmetric spaces of non-compact type include hyperboloids of constant curvature -1 and spaces of symmetric positive definite matrices. This paper obtains a simplified formula in the special case when the symmetric space is the space of normal distributions, a 2-dimensional hyperboloid.

Keywords and phrases: Harmonic analysis, Helgason-Fourier Transform, Kernel density estimator, Non-Euclidean Geometry, Non-parametric.

Contents

1. Introduction

Data, while often expressed as collections of real numbers, are often more naturally regarded as points in non-Euclidean spaces. To take an example, radar systems can yield the data of bearings for planes and other flying objects; those bearings are naturally regarded as points on a sphere [13]. To take another example, diffusion tensor imaging (DTI) can yield information about how liquid flows through a region of the body being imaged; that three-dimensional movement can be expressed in the form of symmetric positive definite (3×3)-matrices [13]. To take yet another example, the nodes of certain hierarchical real-world networks can be regarded as having latent coordinates in a hyperboloid [8, 1]. In all such examples, the spaces can be regarded as subsets of Euclidean space even though Euclidean

1

解决方案：

ComPDFKit 结合 AI 技术和计算机视觉技术，解析不同格式和排版的 PDF 论文，解决了不同学科论文中多样表数据问题。同时引入术语识别模型处理专业术语，引入领域专家知识优化算法。在高效率的前提下，提高了对复杂学科差异的识别和解释能力，数据准确率高达95%以上。

[4] arXiv:2401.10895 [pdf , source] today

AI in Supply Chain Risk Assessment: A Systematic Literature Review and Bibliometric Analysis

Md Abrar Jahin,Saleh Akram Naife,Anik Kumar Saha,M. F. Mridha

Subjects: Machine Learning (cs.LG) ; Computational Engineering, Finance, and Science (cs.CE)

Supply chain risk assessment (SCRA) has witnessed a profound evolution through the integration of artificial intelligence (AI) and machine learning (ML) techniques, revolutionizing predictive capabilities and risk mitigation strategies. The significance of this evolution stems from the critical role of robust risk management strategies in ensuring operational resilience and continuity within modern supply chains. Previous reviews have outlined established methodologies but have overlooked emerging AI/ML techniques, leaving a notable research gap in understanding their practical implications within SCRA. This paper conducts a systematic literature review combined with a comprehensive bibliometric analysis. We meticulously examined 1,717 papers and derived key insights from a select group of 48 articles published between 2014 and 2023. The review fills this research gap by addressing pivotal research questions, and exploring existing AI/ML techniques, methodologies, findings, and future trajectories, thereby providing a more encompassing view of the evolving landscape of SCRA. Our study unveils the transformative impact of AI/ML models, such as Random Forest, XGBoost, and hybrids, in substantially enhancing precision within SCRA. It underscores adaptable post-COVID strategies, advocating for resilient contingency plans and aligning with evolving risk landscapes. Significantly, this review surpasses previous examinations by accentuating emerging AI/ML techniques and their practical implications within SCRA. Furthermore, it highlights the contributions through a comprehensive bibliometric analysis, revealing publication trends, influential authors, and highly cited articles.

收起

赞 讨论 笔记 收藏 全屏 图表 我来讲解

pic_0 pic_1 pic_2 pic_3 pic_4 pic_5 pic_6 pic_7 pic_8 pic_9 pic_10 table 1 table 2

table 3 table 4 table 5 table 6 table 7 table 8 table 9 table 10 table 11 table 12 table 13 table 14 table 15

table 16 table 17 table 18 table 19 table 20 table 21

04. 客户案例

2022年某国内高校项目

纸质文件

将纸质文件扫描或拍照，建立数字化的文档数据库，方便随时检索和管理。

学习资料

识别论文的章节结构、论文内容，标记图像、表格、标题等不同版面，分类和合并本地数据，并将其保存为各种所需的格式。

数字图书馆

提取关键信息，如课程名称、教材列表、作者、出版社等，建立图书馆数据库，方便图书检索、借还管理以及统计分析。

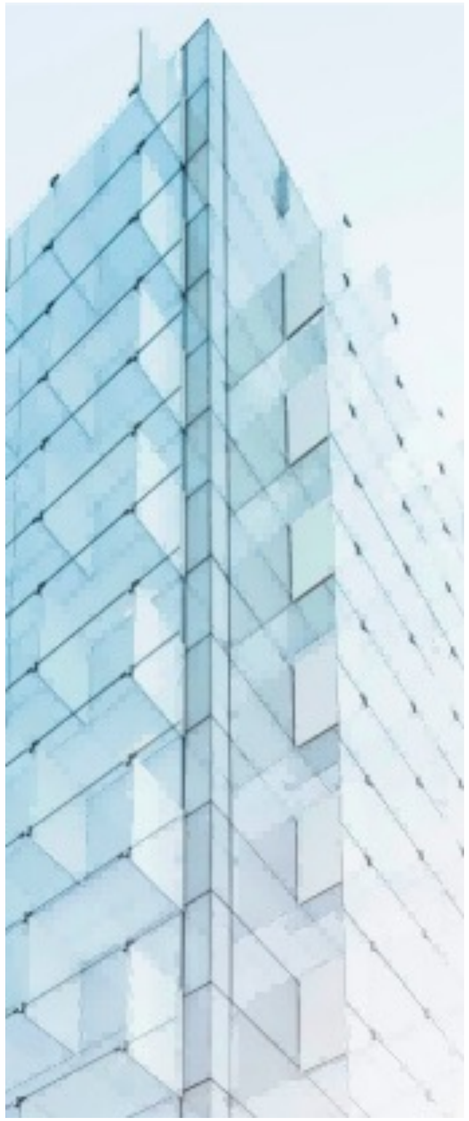
档案管理

通过接入高校教务系统，管理教资、人员档案、财务文件等，精准识别手写文本内容，提供高效率的整合与分析。

ComPDFKit

05. 关于我们

About Us



PDF Technologies, Inc. × ComPDFKit

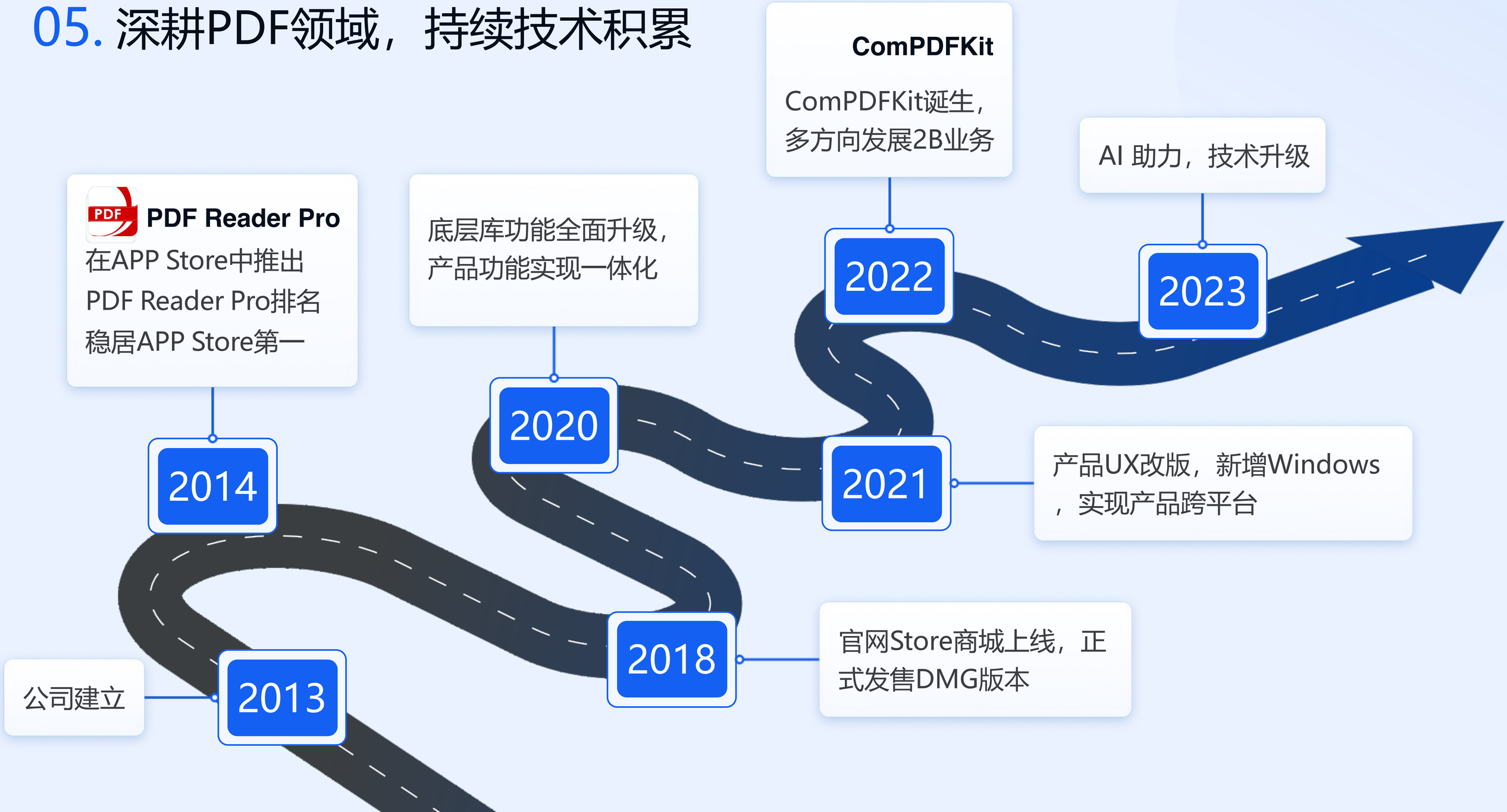
ComPDFKit 是由 PDF Technologies 公司研发，专为个人开发者和初创企业精心打造的尖端 PDF SDK。PDF Technologies, Inc. 成立于2013年，是一家国际化的软件服务厂商。PDF Technologies, Inc. 起始于 PDF 文档处理，致力于提供高效办公的软件及服务。旗下产品 PDF Reader Pro、ComPDFKit、Document AI 等，深受全球各地 9 千万用户的喜爱和支持。

公司提供的软件和服务涵盖了 PDF 办公领域、SDK 技术授权、SaaS 公有化/ 私有化部署、智能文档等。不论是 C 端的个人用户，还是企业、政府等 B 端客户，都可以通过我们的软件和服务从传统纸质文档转向数字化，提高办公效率!

PDF Technologies, Inc. 将不断提升产品研发和创新的能力，建设团队高效文化，为产品和服务提供源源不断的动力和活力。

官网: https://www.compdf.com
邮箱: support@compdf.com

05. 深耕PDF领域，持续技术积累



05. ComPDFKit

ComPDFKit，是我司专为开发人员设计的最先进的 PDF 解决方案。开发人员可“一次开发,多平台适配”，快速、高效地将软件扩展到多个平台。

平台

Windows | Web | UWP | Android | iOS | Mac | Linux

语言

C/C++ | C# | Python | JavaScript | Java | PHP | Ruby | Kotlin | Swift | Objective-C | VB

框架

.Net Framework | .Net Core | WPF | Vanilla JavaScript | React | Angular | Vue | Svelte | Next.js | Nuxt.js

Blazor | jQuery | React Native | Flutter | Electron | MAUI | Xamarin | Cordova | Lonic | Node.js

集成

SharePoint | Microsoft Teams | Microsoft OneDrive | Salesforce

API库

API for Java | API for PHP | API for Python | API for .Net | API for Swift

ComPDFKit

PDF Reader Pro

PDF Reader Pro由ComPDFKit提供技术支持，是一款用户必备的集管理、编辑、转换、阅读功能于一体的专业的全能PDF阅读器。快速、易用、强大，让您出色完成 PDF 工作。

1

APP Store排名第一

30+

获得多家平台奖项

10m+

年度下载量超千万次

200M+

全球 2亿活跃用户

05. 合作伙伴

ComPDFKit

SETAPP

APPSUMO

stacksocial

BITLINK
LEAD YOU TO THE TECHNOLOGY

Seller Solutions

Cult of Mac Store

MacUpdate

Macworld

PCWorld

9TO5Mac

ComPDFKit

技术支持 & 联系我们

官网: https://www.compdf.com

邮箱: support@compdf.com