



# 数据提取助力AI

创新科技 引领未来

## AI大语言模型 训练的挑战

The Challenges for  
Training AI Large  
Language Models

01.

---

## ComPDFKit 的 应对措施

ComPDFKit's  
Solutions

02.

---

## 产品介绍

Product Introduction

03.

---

## 客户案例

User Cases

04.

---

## 关于我们

About Us

05.

---

# 01. AI大语言模型训练的挑战

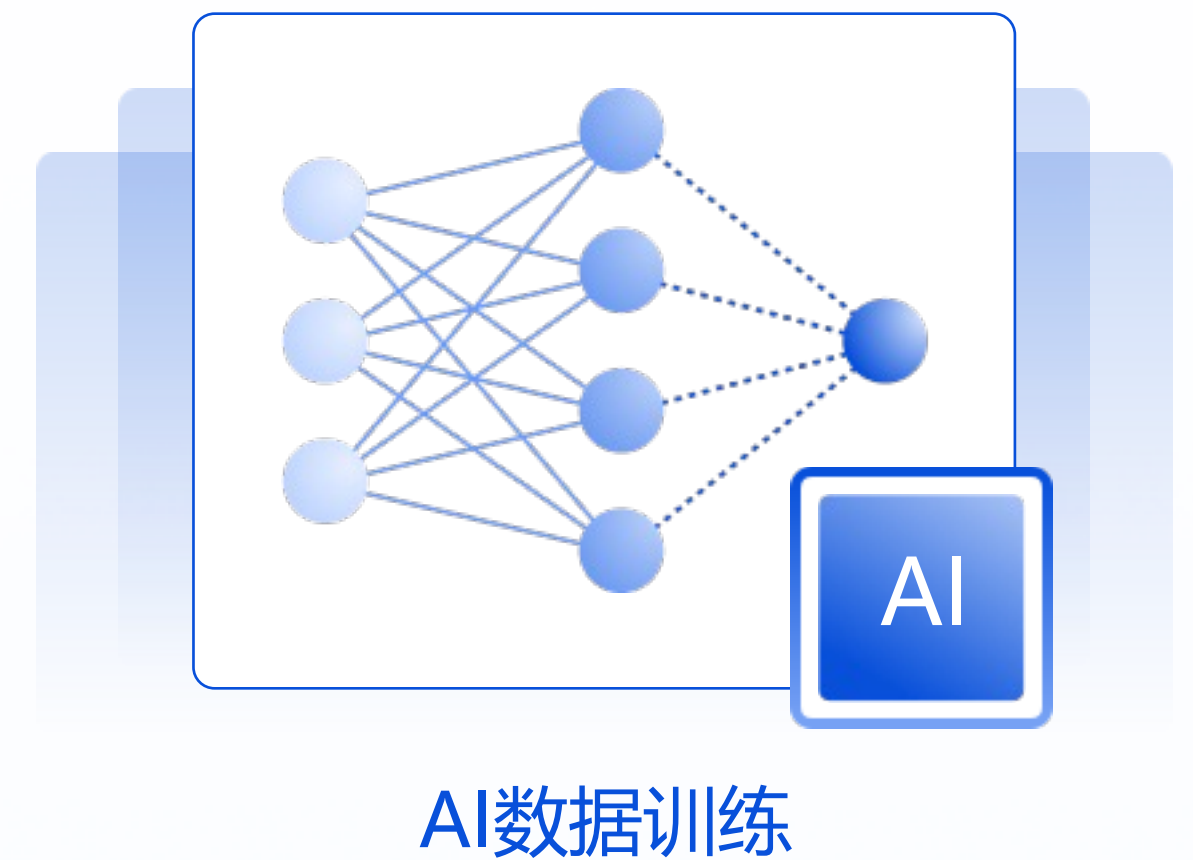
## The Challenges for Training AI Large Language Models



# 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，

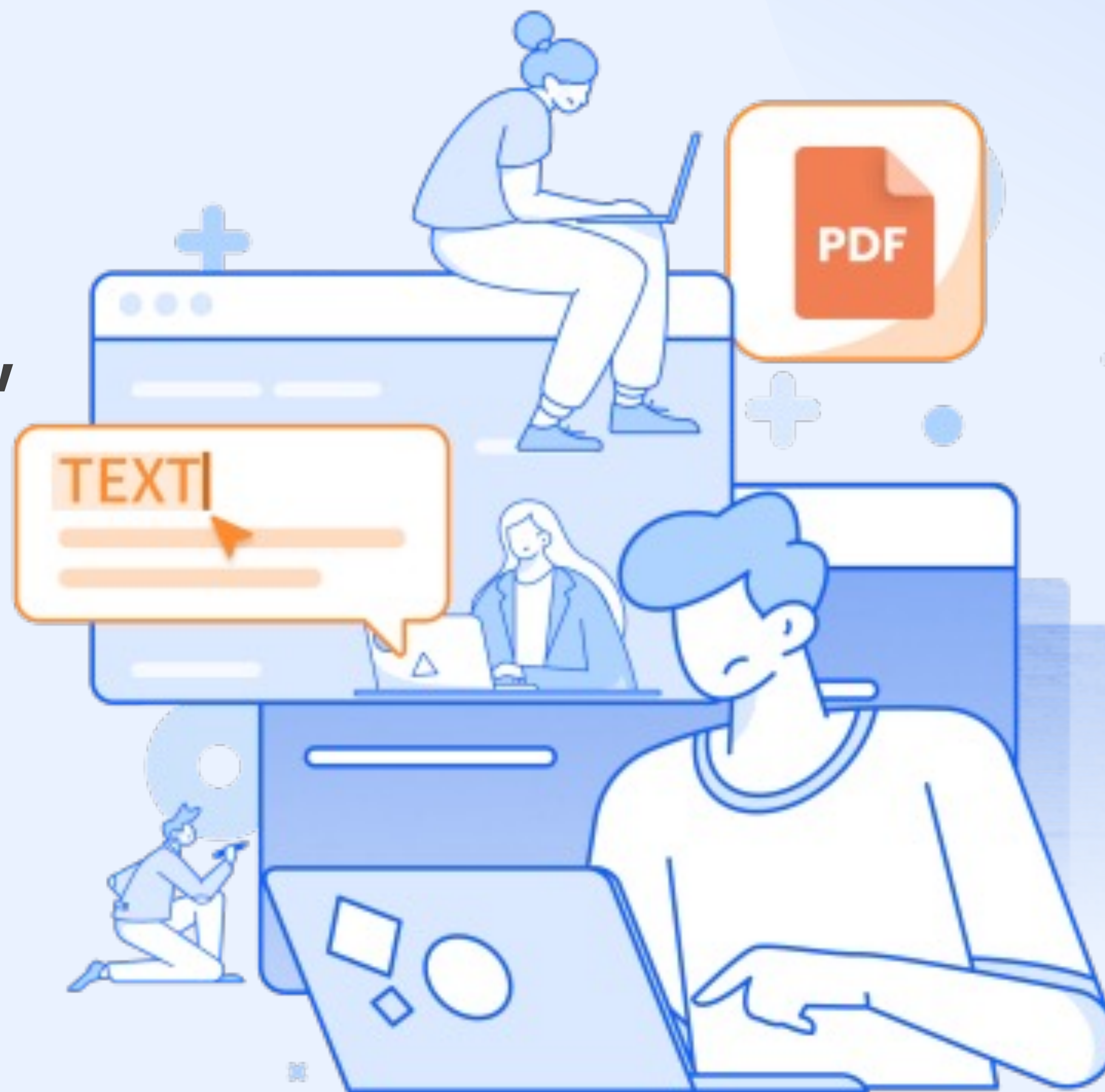
**非结构化数据**

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？



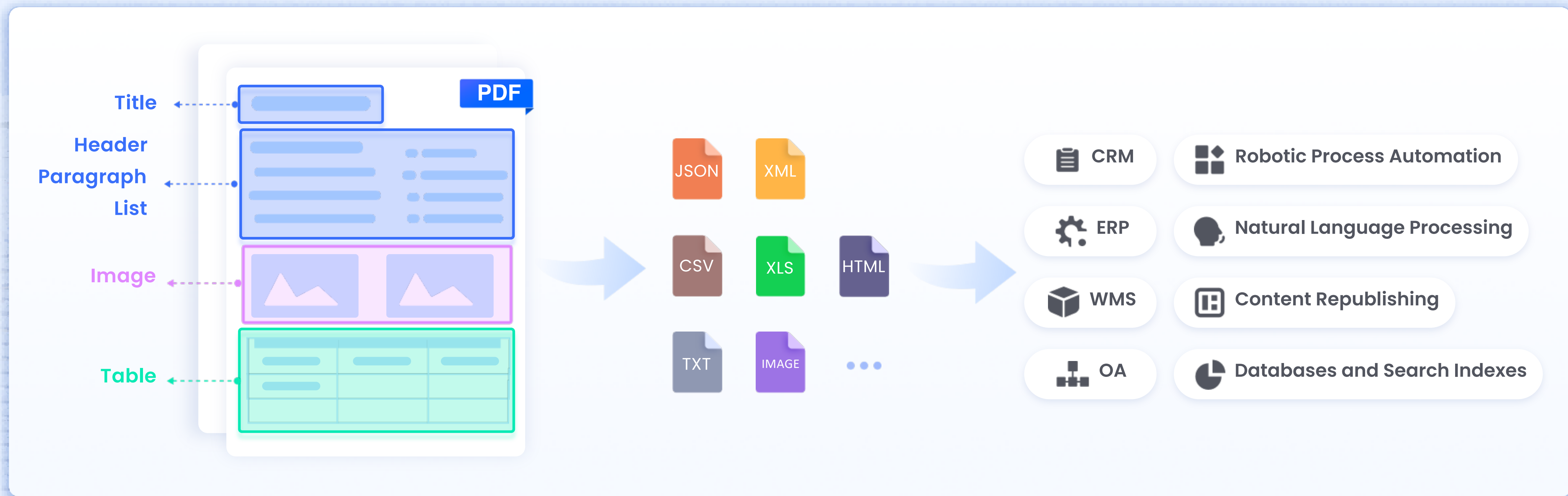
# 02. ComPDFKit 的应对措施

ComPDFKit's Solutions



## 02. ComPDFKit 助力数据提取

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。



“ 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。

circumference.

In August 1492, Columbus set sail with his three small caravels (Figure 1.15). After a voyage of about three thousand miles lasting six weeks, he landed on an island in the Bahamas named Guanahani by the native Lucayans. He promptly christened it San Salvador, the name it bears today.



**FIGURE 1.15** Columbus sailed in three caravels such as these. The *Santa Maria*, his large

### 1.3 West Africa and the Role of Slavery

**LEARNING OBJECTIVES**

At the end of this section, you will be able to:

- Locate the major West African empires on a map
- Discuss the roles of Islam and Europe in the slave trade

It is difficult to generalize about West Africa, which was linked to the rise and diffusion of the Atlantic World, stretches from modern-day Democratic Republic of the Congo and encompasses lush rainforests along the equator to the heavily wooded area near the equator, farmers raised yams, palm products, or plantain. Where water was too scarce for farming, herders maintained sheep, goats, cattle, or camels. In the heavily wooded area near the equator, farmers raised yams, palm products, or plantain. Sub-Saharan Africans had little experience in maritime trade. Most of the population lived away from the coast, which is connected to the interior by five main rivers: the Niger, Volta, and Congo.

Although there were large trading centers along these rivers, most West Africans lived in small villages. They identified with their extended family or their clan. Wives, children, and dependents (including slaves) were a sign of wealth among men, and polygyny, the practice of having more than one wife, was widespread. In time of need, relatives, however far away, were counted upon to assist. Because of the clannish nature of African society, "we" was associated with the clan, while "they" included everyone else. Hundreds of separate dialects emerged. Nearly five hundred are still spoken.

```
JSON XML
"lines": [
  {
    "page": 1,
    "text": "1.3 • West Africa and the Role of Slavery",
    "rect": [ 398.3780822753906, 33.10498809814453,
    540.00017822226563, 25.99498748779297 ]
  },
  {
    "page": 1,
    "text": "circumference.",
    "rect": [ 72.0, 67.823974609375, 136.27799731445314,
    60.88497543334961 ]
  },
  {
    "page": 1,
    "text": "In August 1492, Columbus set sail with his three small
    caravels (Figure 1.15). After a voyage of about three",
    "rect": [ 72.0, 89.42396545410156, 518.6160532226562,
    80.52296447753906 ]
  },
  {
    "page": 1,
    "text": "thousand miles lasting six weeks, he landed on an island in
    the Bahamas named Guanahani by the native",
    "rect": [ 72.0, 102.92396545410156, 515.1960092773437,
    94.1849594116211 ]
  },
  {
    "page": 1,
    "text": "Lucayans. He promptly christened it San Salvador, the name
    it bears today."
  }
]
```

# 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。



图例

- 段落
- 表格
- 图表
- 页眉
- 页脚

“ 表格结构

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					

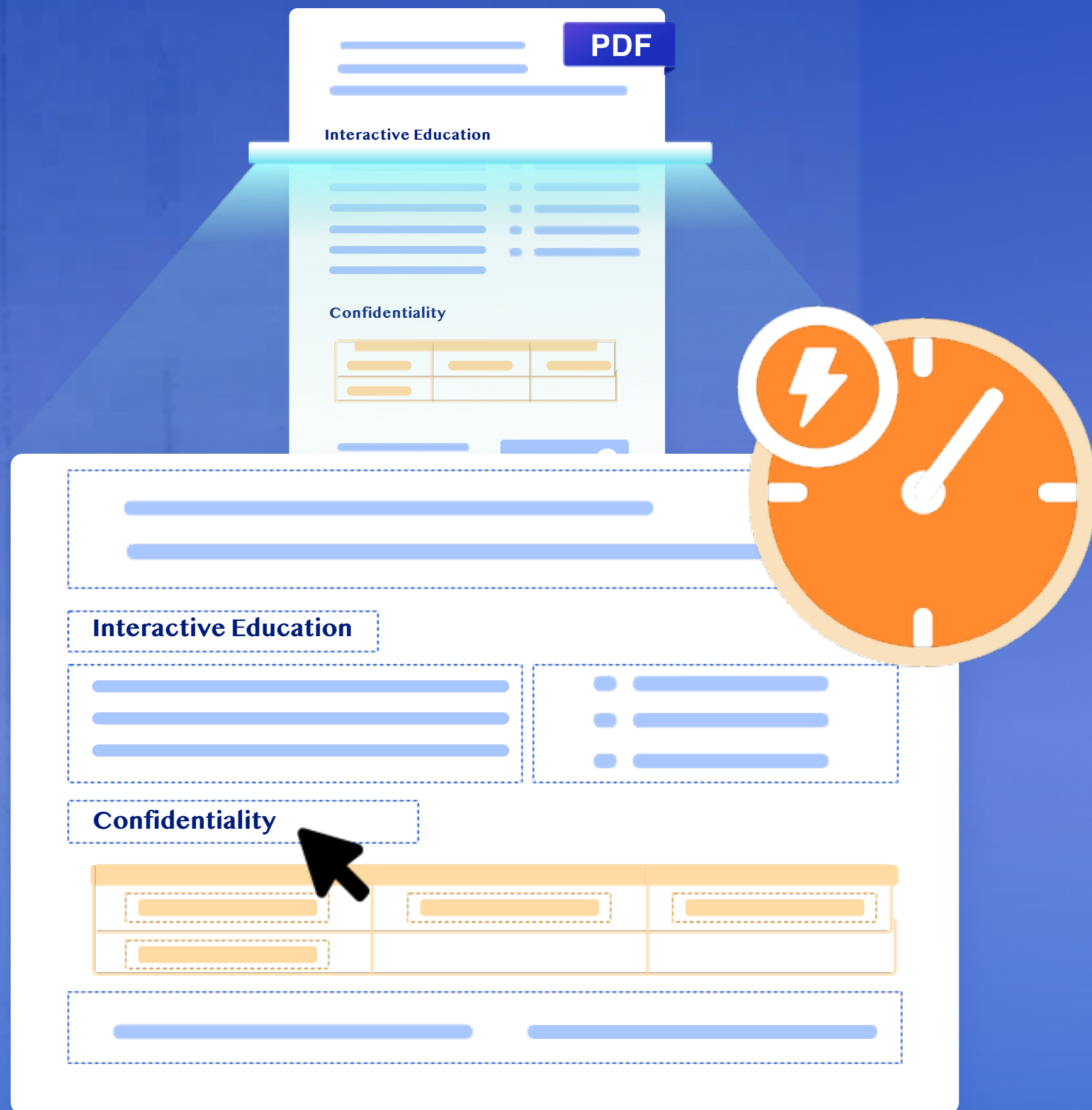
## 02. 措施三：还原表格结构

- 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。

## 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。



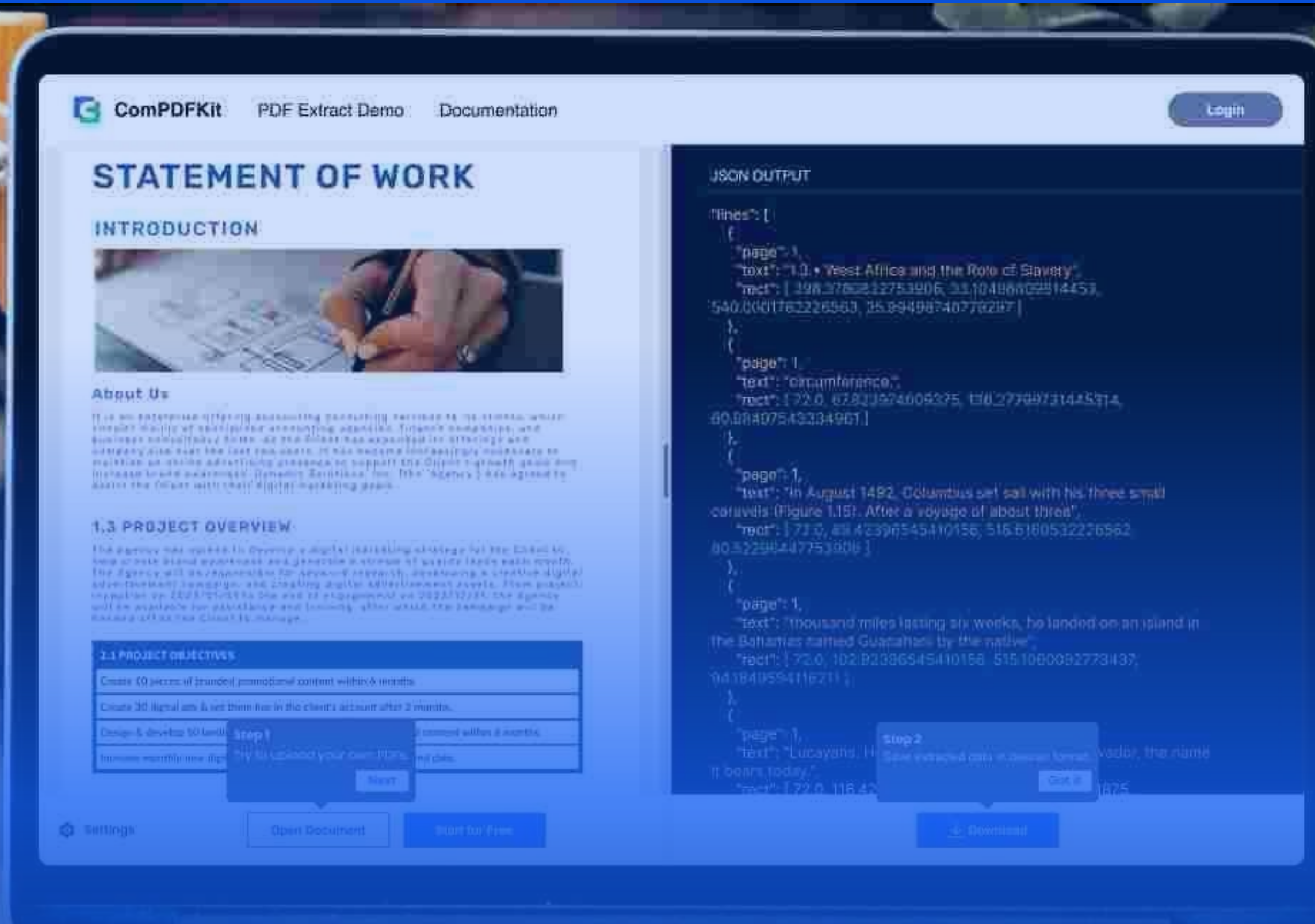


## 02. 措施五：提高处理效率

- 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。

# 03. 产品介绍

## Products Introduction



# 03. 产品优势

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

 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	<ul style="list-style-type: none"><li>• SDK (Java, .NET, C++)</li><li>• RESTful</li><li>• On-premises</li></ul>	<ul style="list-style-type: none"><li>• RESTful</li></ul>	<ul style="list-style-type: none"><li>• SDK(Java, .NET)</li><li>• RESTful</li><li>• CLI</li></ul>
支持格式	<ul style="list-style-type: none"><li>• JSON, XML</li><li>• XLSX, CSV</li><li>• DOCX</li><li>• HTML</li><li>• TXT</li><li>• PNG, JPG</li></ul>	<ul style="list-style-type: none"><li>• JSON</li></ul>	<ul style="list-style-type: none"><li>• JSON</li><li>• XML</li></ul>
提取元素	<ul style="list-style-type: none"><li>• 文字</li><li>• 图片</li><li>• 表格</li><li>• 页眉, 页脚, 标题</li><li>• 列表, 公式, 代码</li><li>• 等等</li></ul>	<ul style="list-style-type: none"><li>• 文字</li><li>• 图片</li><li>• 表格</li></ul>	<ul style="list-style-type: none"><li>• 文字</li><li>• 图片</li><li>• 表格</li></ul>



## 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
- 自主知识产权, 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

# 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。

血常规

项目名称	体检结果	参考值	单位
大型血小板比例	34.40	18.5-42.3	%
血小板压积	0.28	0.17-0.32	%
血小板分布宽度	14.7	10.1-16.1	fl
血小板计数	251	101-320	10 <sup>9</sup> /L
红细胞计数	4.60	3.68-5.74	10 <sup>12</sup> /L
红细胞分布宽度变异系数	14.9	12-13.6	%
红细胞分布宽度标准差	47.6	37.1-45.7	fl
白细胞计数	5.87	成人: 3.5-9.5 儿童(8-10): 婴儿(11-12): 新生儿(20)	10 <sup>9</sup> /L

小结: 1. 单核细胞计数: 0.24 10<sup>9</sup>/L (参考范围: 0.29-0.95) ↓  
2. 单核细胞百分率: 4.10 % (参考范围: 5.2-15.2) ↓  
3. 红细胞分布宽度变异系数: 14.9 % (参考范围: 12-13.6) ↓  
4. 红细胞分布宽度标准差: 47.6 fl (参考范围: 37.1-45.7) ↓  
5. 白细胞计数: 5.87 10<sup>9</sup>/L (参考范围: 成人: 3.5-9.5  
儿童(8-10):  
婴儿(11-12):  
新生儿(20))

审核时间: 2021-11-01 检查医生: 康梦楠

个人信息

姓名: 潘清 性别: 女  
民族: 汉族 出生年月: 1987年6月  
籍贯: 河北涿州 学历: 工学硕士  
学校: 北京理工大学 专业: 控制科学与工程  
英语水平: CET-6 毕业时间: 2014年3月

教育经历

2011.09-2014.03: 北京理工大学(985) 自动化学院 控制科学与工程  
获得学位: 工学硕士(免试) 排名: 前5% GPA: 88/100  
硕士论文: 《动态场景中运动目标检测与跟踪技术的研究》  
主修课程: 随机过程理论及应用(93)、智能控制(90)、智能信息处理(93)、模式识别(93)、线性系统理论(89)、应用数理统计(96)、多智能体协同与控制(80)、嵌入式系统与应用(89)、自动控制中的线性代数(76)。

2007.09-2011.06: 太原理工大学(211) 信息工程学院 自动化  
获得学位: 工学学士 排名: 1/138 GPA: 92/100  
学士论文: 《车辆识别系统设计》  
主修课程: 自动控制理论(97)、过程控制系统(100)、模拟电子技术(97)、数字电子技术(96)、电力电子技术(99)、嵌入式系统基础(97)、电路理论(97)、模糊控制系统(95)、传感器原理与接口技术(95)、现场总线与分布式系统(95)。

科研成果&获奖情况

- 论文: A new moving objects detection method based on improved SURF algorithm, 第25届中国控制与决策会议(ED), 已收录。
- 专利: 基于多相机旋转扫描的实时全景监控方法和装置。
- 国家级: 国家奖学金1次(1/200); 国家励志奖学金2次(3/100)。
- 省级: 山西省优秀毕业生称号(1/200)。
- 校级: 特等奖学金4次(1/200); 一等奖学金2次(3/100); 三好学生(3/100); 优秀团员(5/100); "太原理工大学优秀毕业生"(3/100)。

技能掌握

- 英语水平: CET-6, 具有较强的英语读写能力, 能够熟练阅读和翻译英文文档。
- 计算机: 通过国家计算机等级考试(二级C), 熟练应用 Office 软件。
- 软件编程: 熟练应用 C/C++, MFC 等编程技术, 擅长 Win32 平台软件开发与应用, 熟练应用 MFC 进行人机交互界面的开发。
- OpenCV: 熟练掌握采用 OpenCV 和 C/C++ 对数字图像处理算法进行开发与应用, 熟练掌握各种流行的目标检测与跟踪技术。
- CUDA: 深入了解 CUDA 开发平台和编程模型, 掌握 CUDA 并行编程技术。
- 专业基础: 在模拟电路、数字电路、单片机、嵌入式系统设计、过程控制、电机拖动、自动化仪器仪表等方面具有一定的专业基础, 掌握经典控制理论、现代控制理论和模糊控制算法。

资源禀赋决定工业化发展路径

富煤贫油少气, 资源禀赋决定工业化发展路径。我国在一次能源生产和消费中, 煤炭占主导地位。"富煤贫油少气"决定了我国以煤炭为主的能源结构。我国煤炭资源丰富, 资源储量超过50万亿吨, 已探明储量1.53万亿吨, 占一次能源总量的84%以上。天然气与石油不足。煤炭与石油产量之和仅相当于3亿多桶油当量, 其他替代能源总量仅相当于1亿多桶油当量。而我国去年的一次能源消费量达到了41.8亿吨标准煤。考虑到我国能源结构的天然禀赋, 未来很长一段时间, 我国的主要能源还是以煤为主。根据能源发展"十三五"规划(发改能源局2017/01/17)到2020年, 我国能源消费总量控制在50亿吨标准煤以下, 煤炭在我国一次能源消费结构中的比重控制在50%左右。

2020年能源发展主要指标

总量	占比
煤炭	58%
石油	17%
天然气	10%
核能	15%

煤炭储量丰富, 油气资源对外依存度高。根据石油公司(BP)发布的《世界能源统计年鉴2016》, 截至2015年底, 我国探明石油储量中, 煤炭约1145亿吨, 石油约25亿吨, 天然气约3万亿立方米。其中煤炭占世界总量的12.8%, 对外依存度8%; 石油占1.1%, 对外依存度81%; 天然气占2.1%, 对外依存度100%。

我国能源结构中, 油气对外依存度高

我国一次能源消费构成

能源类型	占比
煤炭	61%
石油	12.8%
天然气	2.1%
核能	22%

记

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \quad (5)$$
$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ -a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31} \end{pmatrix} \quad (6)$$

(6)式称为数表(5)所确定的三阶行列式。

上述定义表明三阶行列式含6项, 每项均为不同行不同列的三个元素的乘积再冠以正负号, 其规律遵循图1.2所示的对角线法则: 图中有三条实线看作是平行于主对角线的连线, 三条虚线看作是平行于副对角线的连线, 实线上三元素的乘积冠正号, 虚线上三元素的乘积冠负号。

例2 计算三阶行列式

$$D = \begin{vmatrix} 1 & 2 & -4 \\ -2 & 2 & 1 \\ -3 & -4 & -2 \end{vmatrix}$$

解 按对角线法则, 有

$$D = 1 \times 2 \times (-2) + 2 \times 1 \times (-3) + (-4) \times (-2) \times 4 - 1 \times 1 \times 4 - 2 \times (-2) \times (-2) - (-4) \times 2 \times (-3) = -4 - 6 + 32 - 4 - 8 - 24 = -14$$

例3 求解方程

# 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	李蔚真	林飞静
联系地址	福州市鼓楼区软件大道89号福州软件园G区5号楼	福州市鼓楼区软件大道89号福州软件园G区5号楼
电话	0591-38509866	0591-38509866
传真	0591-38509809	0591-38509809
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: <a href="https://www.cs.com.cn/">https://www.cs.com.cn/</a> 《上海证券报》网址: <a href="https://www.cnstock.com/">https://www.cnstock.com/</a> 《证券时报》网址: <a href="http://www.stcn.com/">http://www.stcn.com/</a> 《证券日报》网址: <a href="http://www.zqrb.cn/">http://www.zqrb.cn/</a>
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 ( <a href="https://www.sse.com.cn">https://www.sse.com.cn</a> )
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G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

## 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,  $\phi(0)$ , 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,

$$J^\mu = i\psi^\dagger \partial^\mu \psi - \psi \partial^\mu \psi^\dagger, \quad (34)$$

which satisfies a continuity equation,  $\nabla_\mu J^\mu = 0$ . This continuity equation immediately leads to a conserved charge, which is  $N_b$ . The total bosonic particle number inside a radius  $r$ ,  $N_b(r)$ , is typically written in terms of an integral,  $N_b = \int_0^r 4\pi r'^2 \sqrt{-g} \rho_{bb} dr'$ . For static solutions, we use the boson star ansatz in (27) and prefer to write the integral formula in differential form,

$$N_b = 8\pi \int_0^r r'^2 \rho^2 dr', \quad (35)$$

from which  $N_b = N_b(\infty)$ .

To solve Eq. (31) for the critical curve, we follow the methods presented in [10, 12]. We compute contour lines of either  $N_b$  or  $N_b$  in the full system. Notice that in

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  $\phi(0)$ , the system is, respectively, boson or fermion dominated. In these cases, the critical curves reproduce the single-fluid critical values given in Sec. III as expected. Interestingly, in the upper-right corner we see that linearly stable static solutions exist for values of  $P(0)$  and  $\phi(0)$  that go beyond their single-fluid critical values, a phenomenon that also occurs for mixed stars with fermionic dark matter [12].

**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 . . . . . 1
- 2 Preliminaries . . . . . 3
- 2.1 Kernel Density Estimation . . . . . 3
- 2.2 Symmetric spaces . . . . . 4
- 2.3 Helgason-Fourier Analysis . . . . . 6
- 2.4 G-kernel density estimation . . . . . 7
- 2.5 Main theorem . . . . . 9
- 3 Implementation . . . . . 9
- 4 Proofs . . . . . 10
- 5 Conclusion . . . . . 13
- Acknowledgements . . . . . 13
- References . . . . . 13

**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 [1]. 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  $(3 \times 3)$ -matrices [2]. To take yet another example, the nodes of certain hierarchical real-world networks can be regarded as having latent coordinates in a hyperboloid [3, 4]. In all such examples, the spaces can be regarded as subsets of Euclidean space even though Euclidean

## 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

点赞 讨论 笔记 收藏 全屏 图表

我来讲解

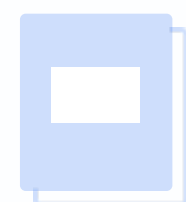
pic\_0 pic\_1 pic\_2 pic\_3 pic\_4 pic\_5 pic\_6 pic\_7 pic\_8 pic\_9 pic\_10 table 1 table 2  
table 3 table 4 table 5 table 6 table 7 table 8 table 9 table 10 table 11 table 12 table 13 table 14 table 15  
table 16 table 17 table 18 table 19 table 20 table 21

# 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。





# 05. 关于我们

About Us





## PDF Technologies, Inc. × ComPDFKit

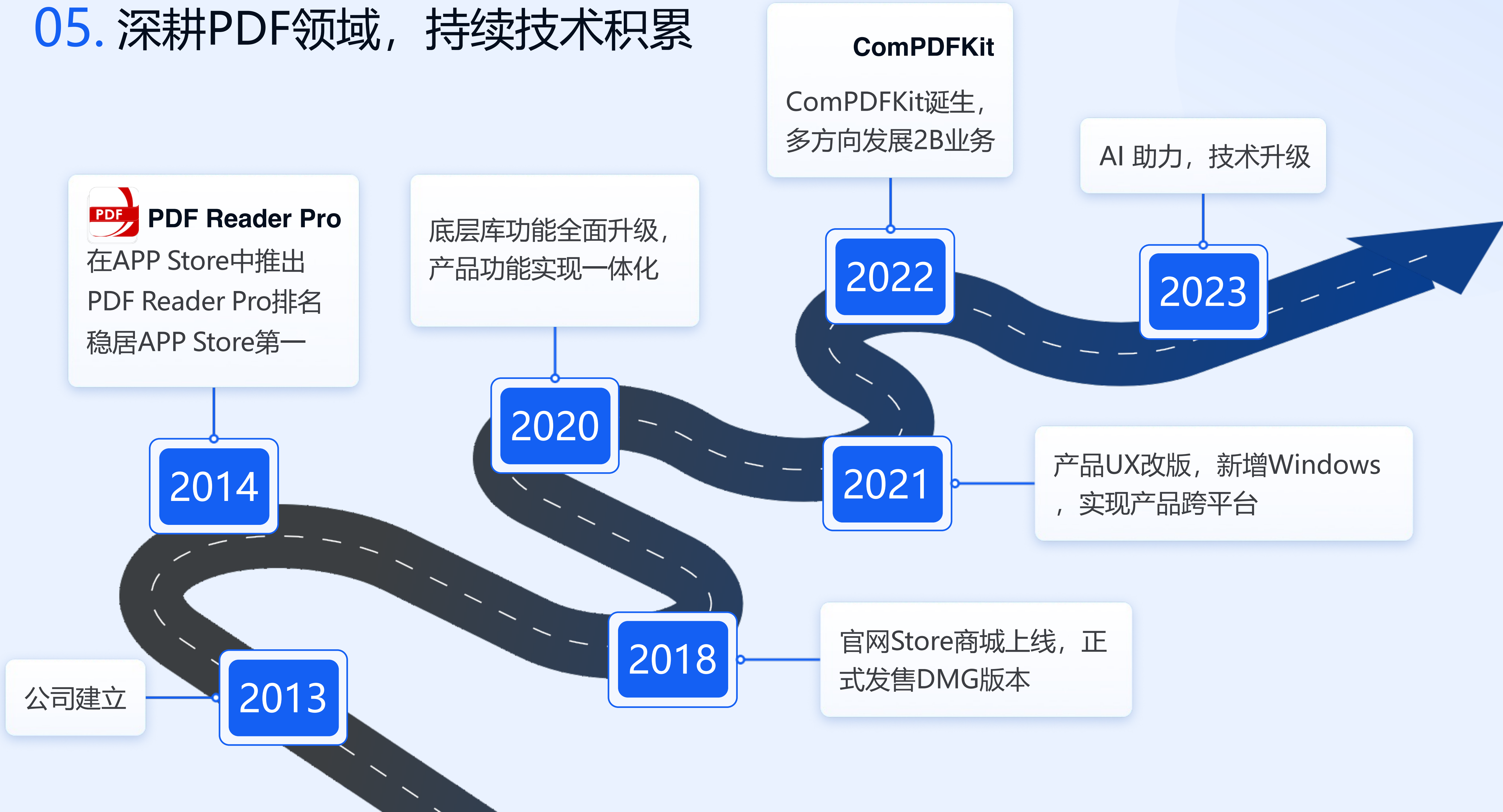
ComPDFKit 是由 PDF Technologies 公司研发，专为个人开发者和初创企业精心打造的尖端 PDF SDK。PDF Technologies, Inc. 成立于2013年，是一家国际化的软件服务厂商。PDF Technologies, Inc. 起始于 PDF 文档处理，致力于提供高效办公的软件及服务。旗下产品 PDF Reader Pro、ComPDFKit、Document AI 等，深受全球各地 9 千万用户的喜爱和支持。

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](mailto: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 | Ionic | 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



## 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



**APPSUMO**



Cult of Mac Store



**Macworld**

**PCWorld**

9TO5Mac

**ComPDFKit**

---

# 技术支持 & 联系我们

官网: <https://www.compdf.com>

邮箱: [support@compdf.com](mailto:support@compdf.com)