Data Science & AI for Economists

Lecture 8: OCR and Text Recognition - From Theory to Practice

Zhaopeng Qu
Business School, Nanjing University

October 28 2025

Roadmap

Today's and next week's Agenda

Part I: Foundations

- Introduction to OCR
  - Core Concepts
  - Applications in Economics

Part II: OCR in Practice:

- Process of OCR
- Ways to access OCR services

OCR in Practice with PaddleOCR

- Local Installation
- Baidu AI Cloud
- LLM models via API

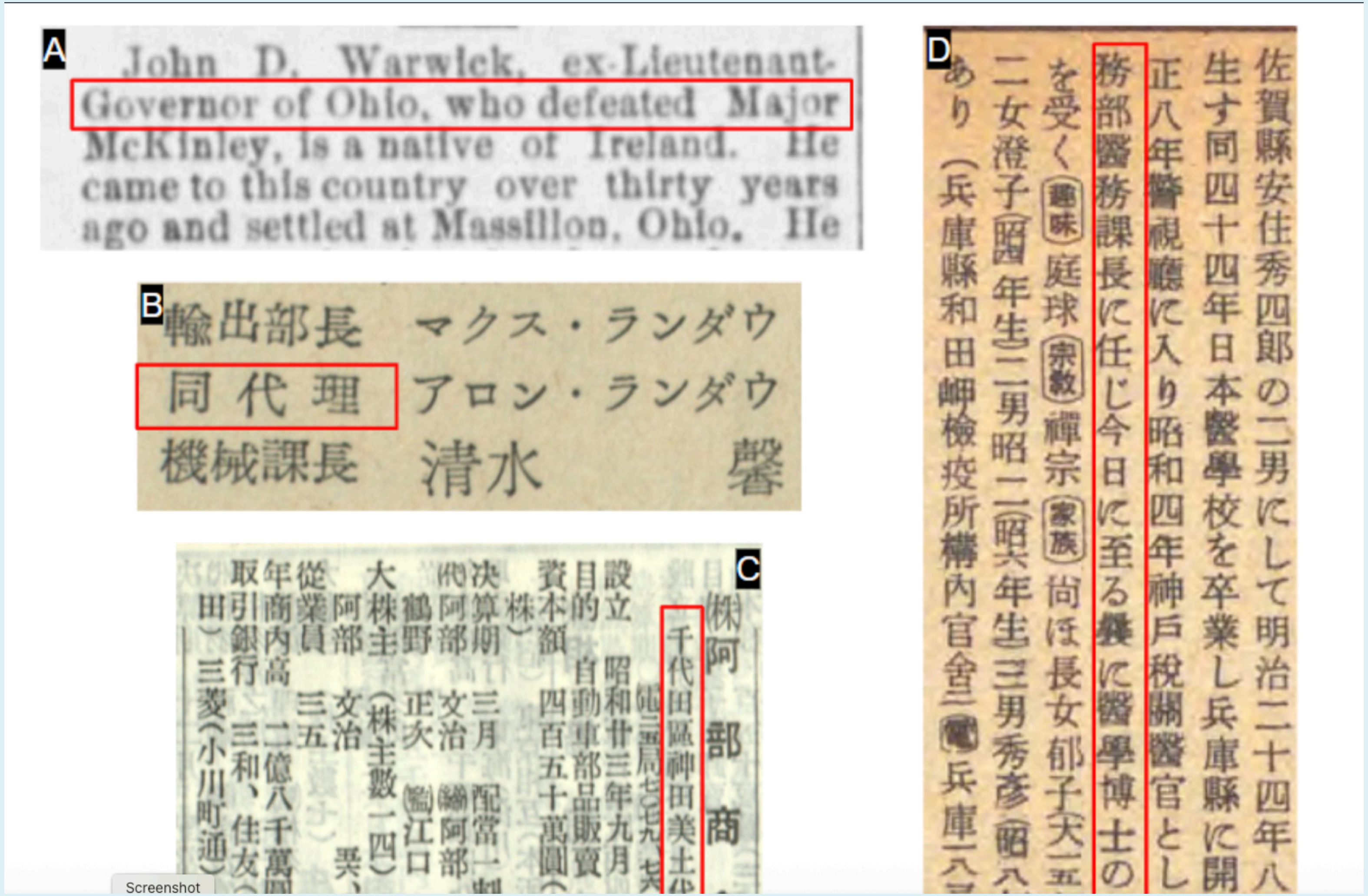
Part I: OCR Foundations

Introduction to OCR

What is OCR?

- OCR (Optical Character Recognition) is a technology that converts images of text into machine-readable text.
- OCR is a powerful tool for economists to collect data from the internet and data from documents.
- The process of OCR is related with some advanced image processing and machine learning techniques.

Introduction to OCR


A
B
C
D
Screenshot

John D. Warwick, ex-Lieutenant-
Governor of Ohio, who defeated Major
McKinley, is a native of Ireland. He
came to this country over thirty years
ago and settled at Massillon, Ohio. He

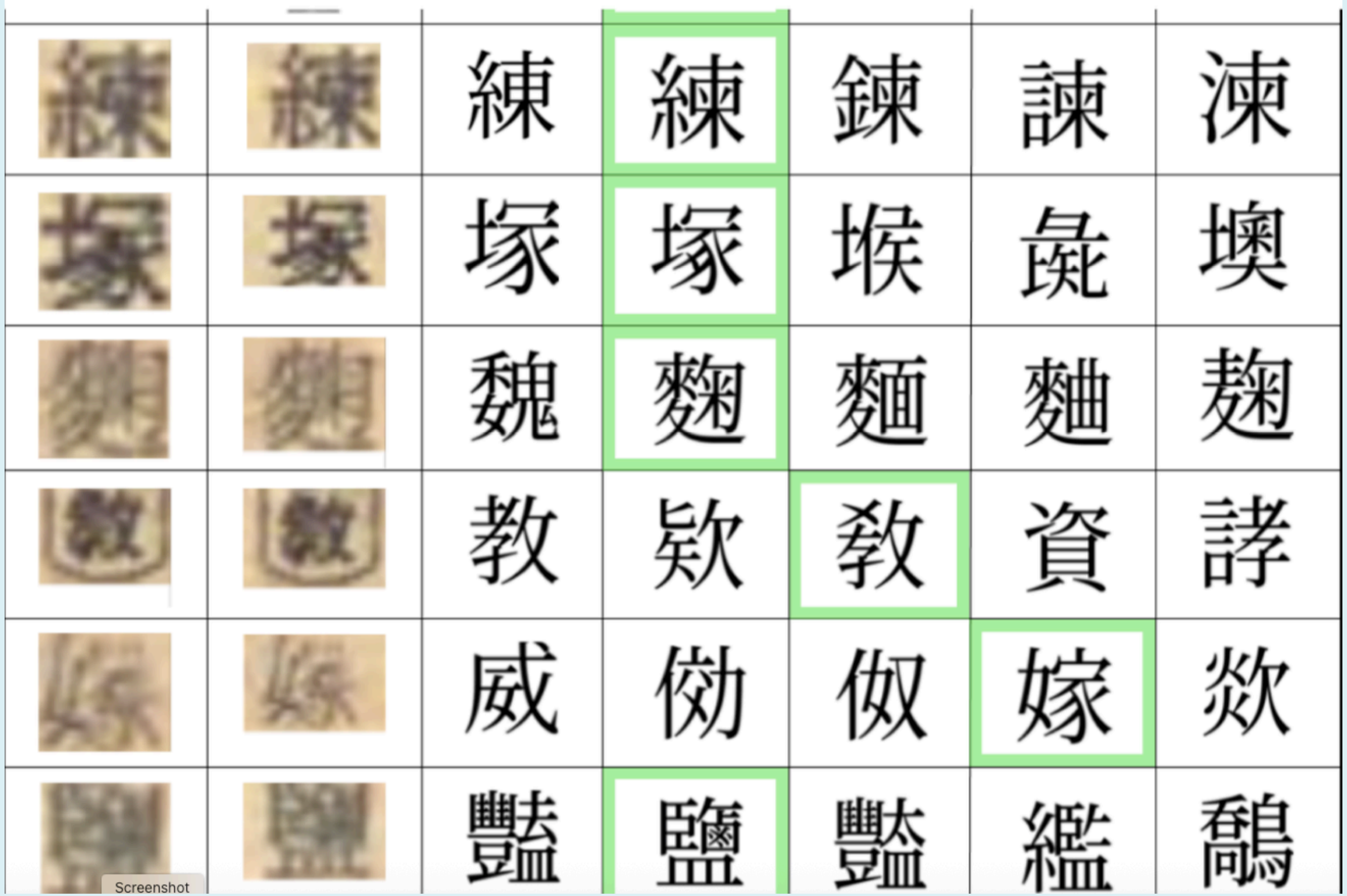
輸出部長　マクス・ランダウ
同　代　理　アロン・ランダウ
機械課長　清水　馨

㈱阿　部　商
千代田區神田美土代
電三局七七九、七
設立　昭和廿三年九月
目的　自動車部品販賣
資本額　四百五十萬圓（
株）
決算期　三月　配當一割
㈹阿部　文治　締阿部
鶴野　正次　監江口
大株主　（株主數一四）
阿部　文治　巽、
從業員　三五
年商内高　二億八千萬圓
取引銀行　三和、住友（
田）三菱（小川町通）

佐賀縣安住秀四郎の二男にして明治二十四年八
生す同四十四年日本醫學校を卒業し兵庫縣に開
正八年警視廳に入り昭和四年神戸税關醫官とし
務部醫務課長に任じ今日に至る曩に醫學博士の
を受く（趣味）庭球（宗教）禪宗（家族）尚ほ長女郁子（大二五
二女澄子（昭四年生）二男昭一（昭六年生）三男秀彦（昭八
あり（兵庫縣和田岬檢疫所構内官舍（電）兵庫一八

source: Hinz et al. (2023)

Introduction to OCR

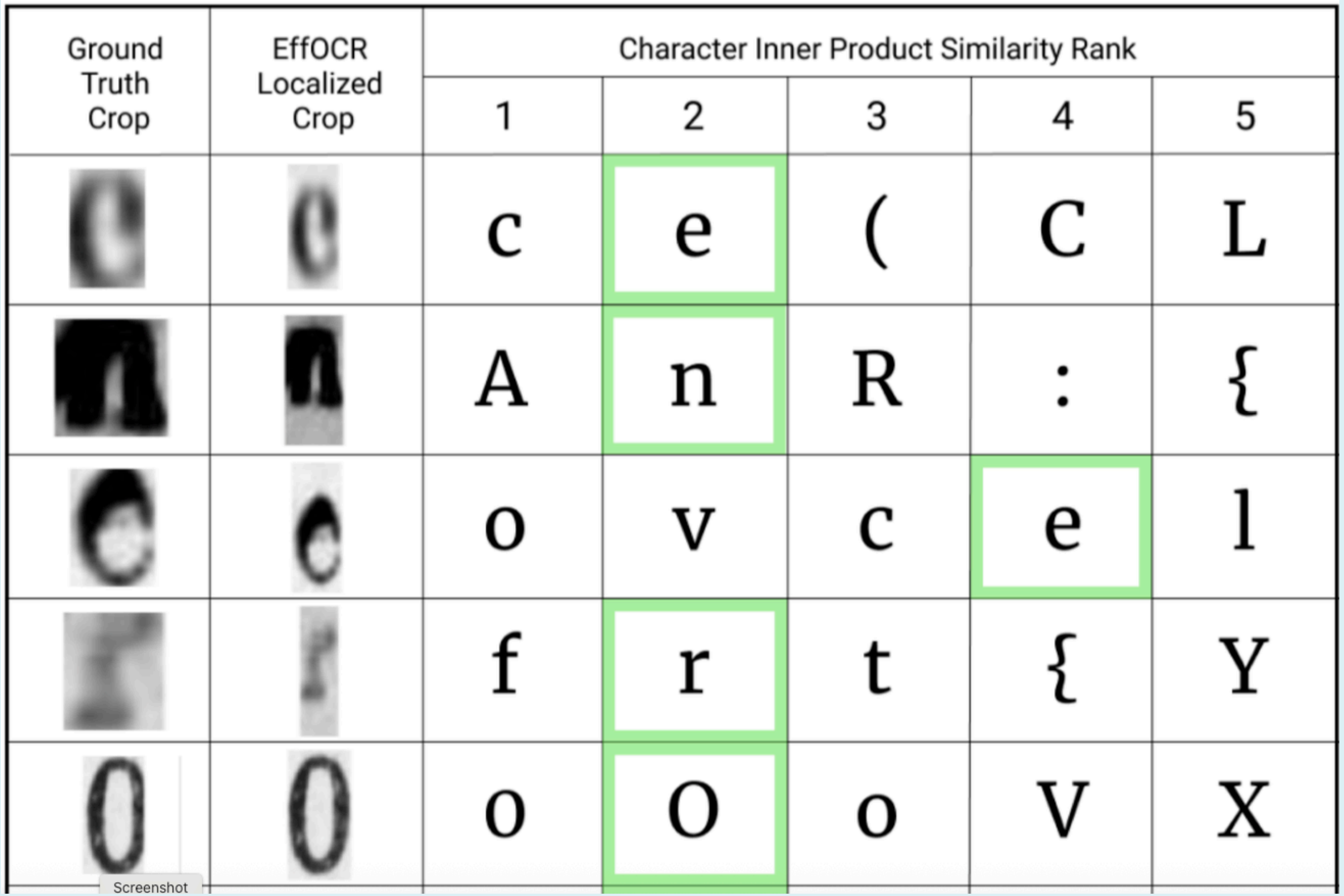


		練	練	鍊	諫	湅
		塚	塚	堠	彘	墺
		魏	麴	麵	麯	麹
		教	欵	敎	資	誟
		威	俲	伮	嫁	欻
		豔	鹽	豓	繿	鷮

Screenshot

source: Dell et al. (2024)

Introduction to OCR



Ground Truth Crop	EffOCR Localized Crop	Character Inner Product Similarity Rank				
		1	2	3	4	5
		c	e	(	C	L
		A	n	R	:	{
		o	v	c	e	l
		f	r	t	{	Y
		o	O	o	V	X

Screenshot

source: Dell et al. (2024)

Introduction to OCR

Accuracy of OCR

- OCR accuracy measured using character error rate (CER)
  - Levenshtein distance measures the minimum number of single-character edits (insertions, deletions, or substitutions) needed to change one string into another.
  - This distance is then divided by the length of the correct text ("ground truth") to get a normalized error rate.
- Example:
  - Ground truth: "hello"
  - OCR result: "helo"
  - Levenshtein distance = 1 (one deletion)
  - CER = $1/5 = 0.2$ or 20% error rate

Applications of OCR in Social Science

1. Extracting Historical Data:
   - Digitizing economic reports, newspapers, and archival documents.
2. Automating Data Entry:
   - Converting scanned documents into structured data for analysis.
3. Analyzing Financial Documents:
   - Invoices, contracts, and receipts for economic research.

OCR Applications in Top Economics Journals

- Dell, M., & Querubin, P. (2017). *Nation Building Through Foreign Intervention: Evidence from Discontinuities in Military Strategies*. The Quarterly Journal of Economics, 133(2), 701–764.

- They use OCR to digitize the declassified US Air Force historical documents to extract the data and newly discovered algorithm component of bombing strategy.

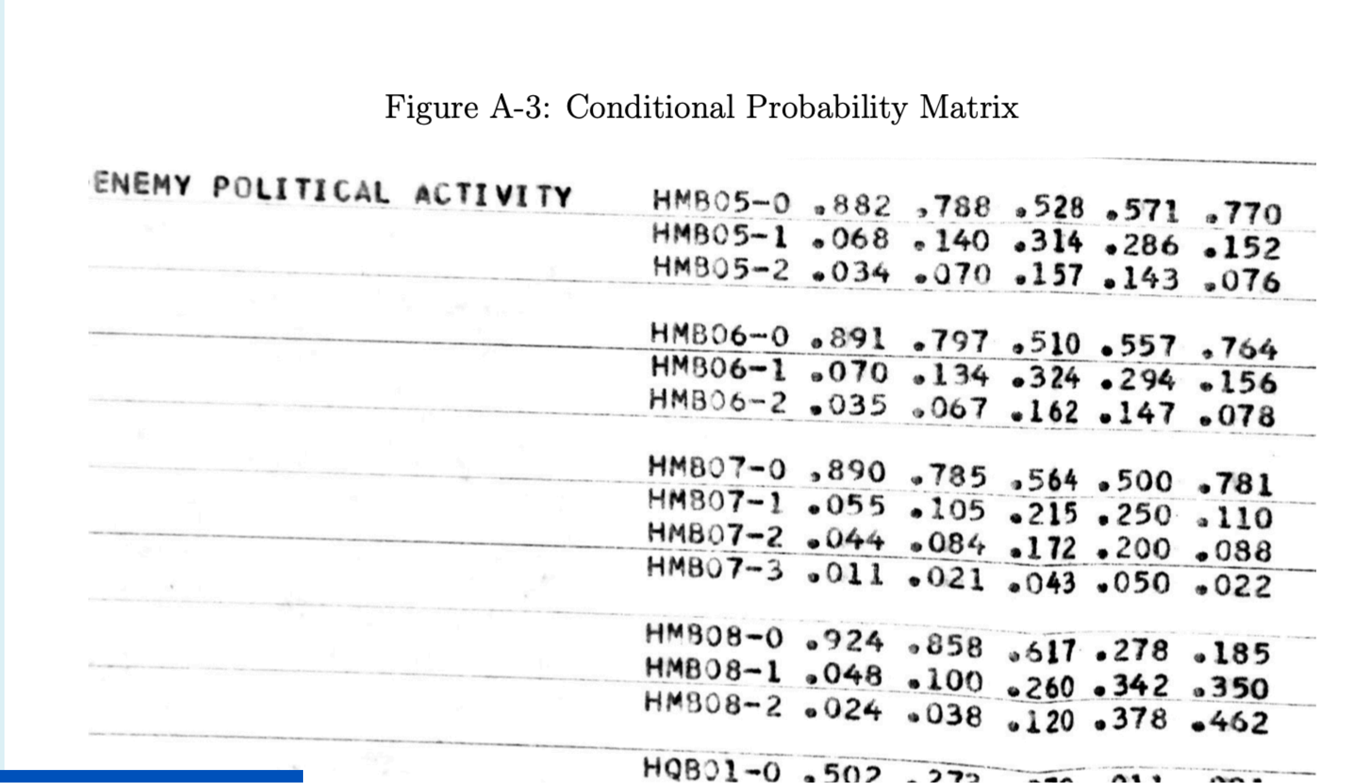


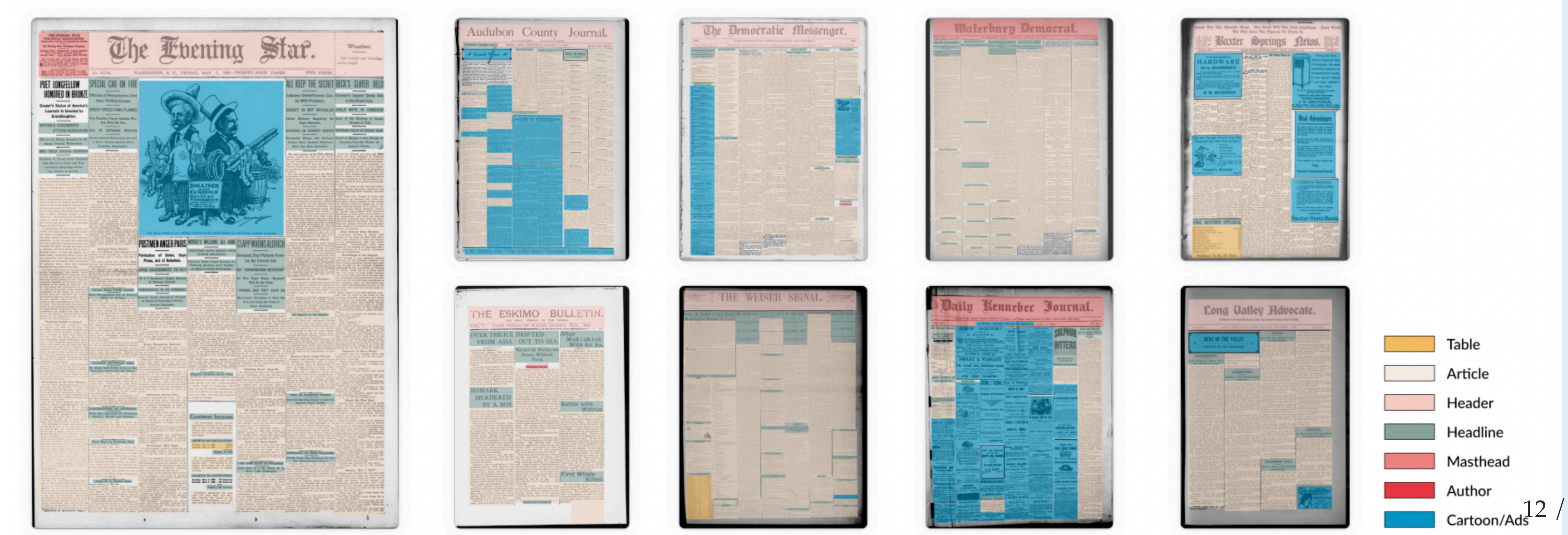
Figure A-3: Conditional Probability Matrix

ENEMY POLITICAL ACTIVITY	HMB05-0	.882	.788	.528	.571	.770
	HMB05-1	.068	.140	.314	.286	.152
	HMB05-2	.034	.070	.157	.143	.076
	HMB06-0	.891	.797	.510	.557	.764
	HMB06-1	.070	.134	.324	.294	.156
	HMB06-2	.035	.067	.162	.147	.078
	HMB07-0	.890	.785	.564	.500	.781
	HMB07-1	.055	.105	.215	.250	.110
	HMB07-2	.044	.084	.172	.200	.088
	HMB07-3	.011	.021	.043	.050	.022
	HMB08-0	.924	.858	.617	.278	.185
	HMB08-1	.048	.100	.260	.342	.350
	HMB08-2	.024	.038	.120	.378	.462
	HQB01-0	.502	[illegible]	[illegible]	[illegible]	[illegible]

OCR Applications in Working Papers

- Dell et al (2024). *American Stories: A Large-Scale Structured Text Dataset of Historical U.S. Newspapers*.

> We detect 1.14 billion individual content regions in around 20M newspaper scans from Library of Congress's Chronicling America collection. Headlines, articles, bylines, and captions are custom-OCRed. The dataset contains 438 million structured article texts.



OCR in Practice

Process of OCR

1. Identify the content type: Document or Image
2. Preprocessing: Image enhancement, noise removal, and binarization.
3. Feature Extraction: Identify and extract features from the image.
4. Character Recognition: Identify and recognize characters in the image.
5. Postprocessing: Correct errors and format the recognized text.

- Others steps are also involved in the process of OCR, such as:
    - locating the text or the image you want to extract.
    - layout analysis, etc.

Ways to access OCR services

1. Business or commercial services: Adobe Acrobat Pro

2. Open-source engines and local engines:

    - **Tesseract**: An open-source OCR engine developed by HP and english is the default language.

    - **PaddleOCR**: A Chinese OCR engine developed by Baidu,excellent for Chinese text recognition.

    - Many others... `olmocr`,

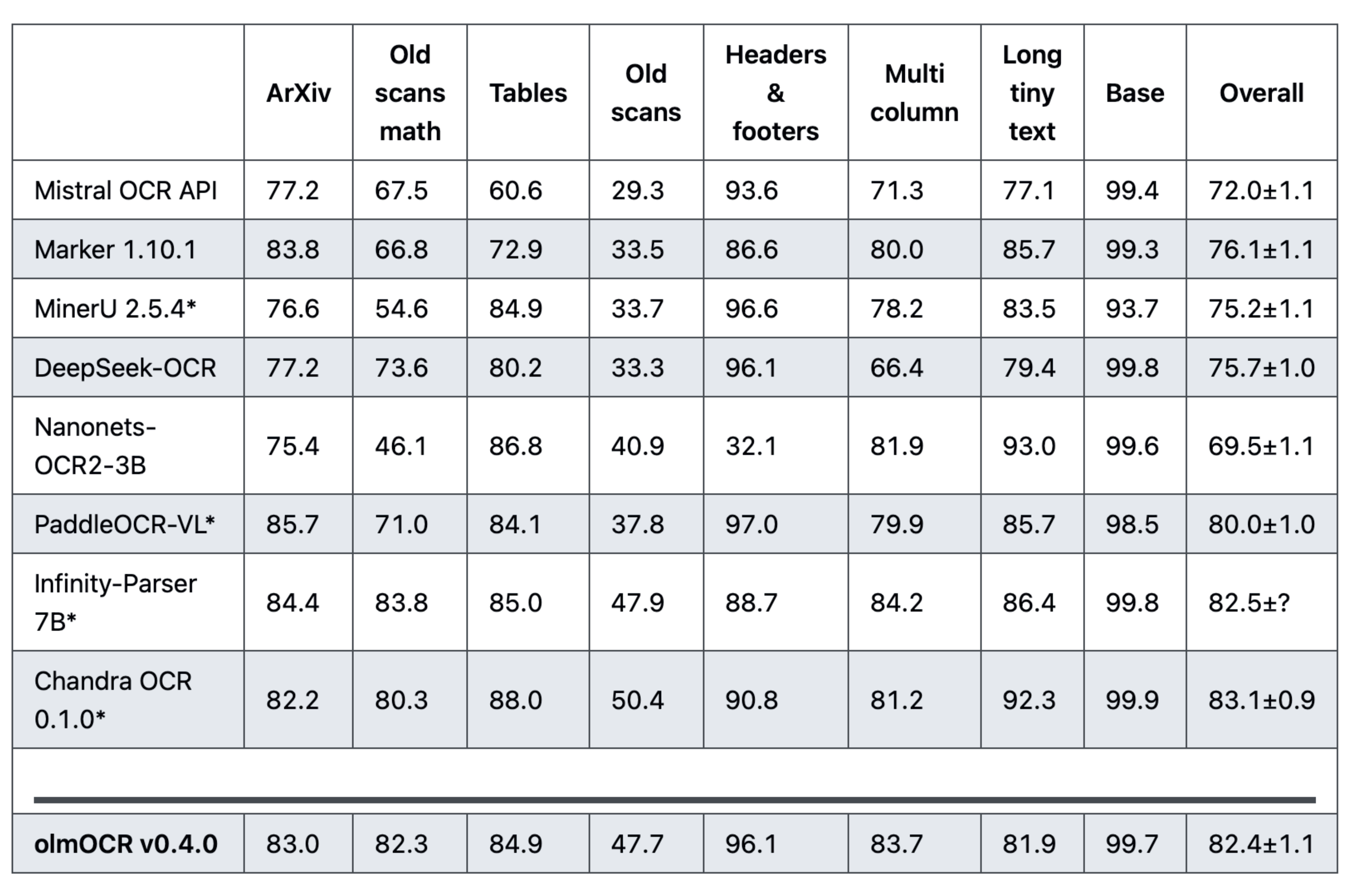
3. Cloud services:

    - **Baidu AI Cloud**: A cloud service platform provided by Baidu, including OCR services.

4. LLM models:

    - 千帆大模型服务平台: A LLM model developed by Baidu
    - **Qwen**: A LLM model developed by Alibaba Cloud
    - **DeepSeek**: A LLM model developed by DeepSeek

Different OCR Engines



	ArXiv	Old scans math	Tables	Old scans	Headers & footers	Multi column	Long tiny text	Base	Overall
Mistral OCR API	77.2	67.5	60.6	29.3	93.6	71.3	77.1	99.4	72.0±1.1
Marker 1.10.1	83.8	66.8	72.9	33.5	86.6	80.0	85.7	99.3	76.1±1.1
MinerU 2.5.4*	76.6	54.6	84.9	33.7	96.6	78.2	83.5	93.7	75.2±1.1
DeepSeek-OCR	77.2	73.6	80.2	33.3	96.1	66.4	79.4	99.8	75.7±1.0
Nanonets-OCR2-3B	75.4	46.1	86.8	40.9	32.1	81.9	93.0	99.6	69.5±1.1
PaddleOCR-VL*	85.7	71.0	84.1	37.8	97.0	79.9	85.7	98.5	80.0±1.0
Infinity-Parser 7B*	84.4	83.8	85.0	47.9	88.7	84.2	86.4	99.8	82.5±?
Chandra OCR 0.1.0*	82.2	80.3	88.0	50.4	90.8	81.2	92.3	99.9	83.1±0.9
olmOCR v0.4.0	83.0	82.3	84.9	47.7	96.1	83.7	81.9	99.7	82.4±1.1

Ways to access OCR services

- Which one should we use? Depends on your needs.
  - Accuracy
  - Cost
  - Language support
  - Speed
  - Privacy
- Local engines: Privacy-sensitive documents, offline processing, no API costs.
- Cloud services: Higher accuracy, support for complex layouts, less setup required.
- LLM models: Higher accuracy, support for complex layouts, less setup required, but more expensive.

OCR in Practice with PaddleOCR(I): Local Installation

OCR in Practice with PaddleOCR(II): Baidu AI Cloud

Introduction to Baidu AI Cloud

What is Baidu AI Cloud?

- Baidu AI Cloud is a platform that provides a wide range of AI services, including OCR (Optical Character Recognition) and many other AI services.
    - 人工智能 (Artificial Intelligence)
    - 云计算 (Cloud Computing)
    - 应用服务 (Application Services)
- It is one of the most popular AI platforms in China, which is similar to AWS and Google Cloud.
- Several Other AI companies in China, like Alibaba Cloud (阿里云) and Tencent Cloud (腾讯云), are also very popular.

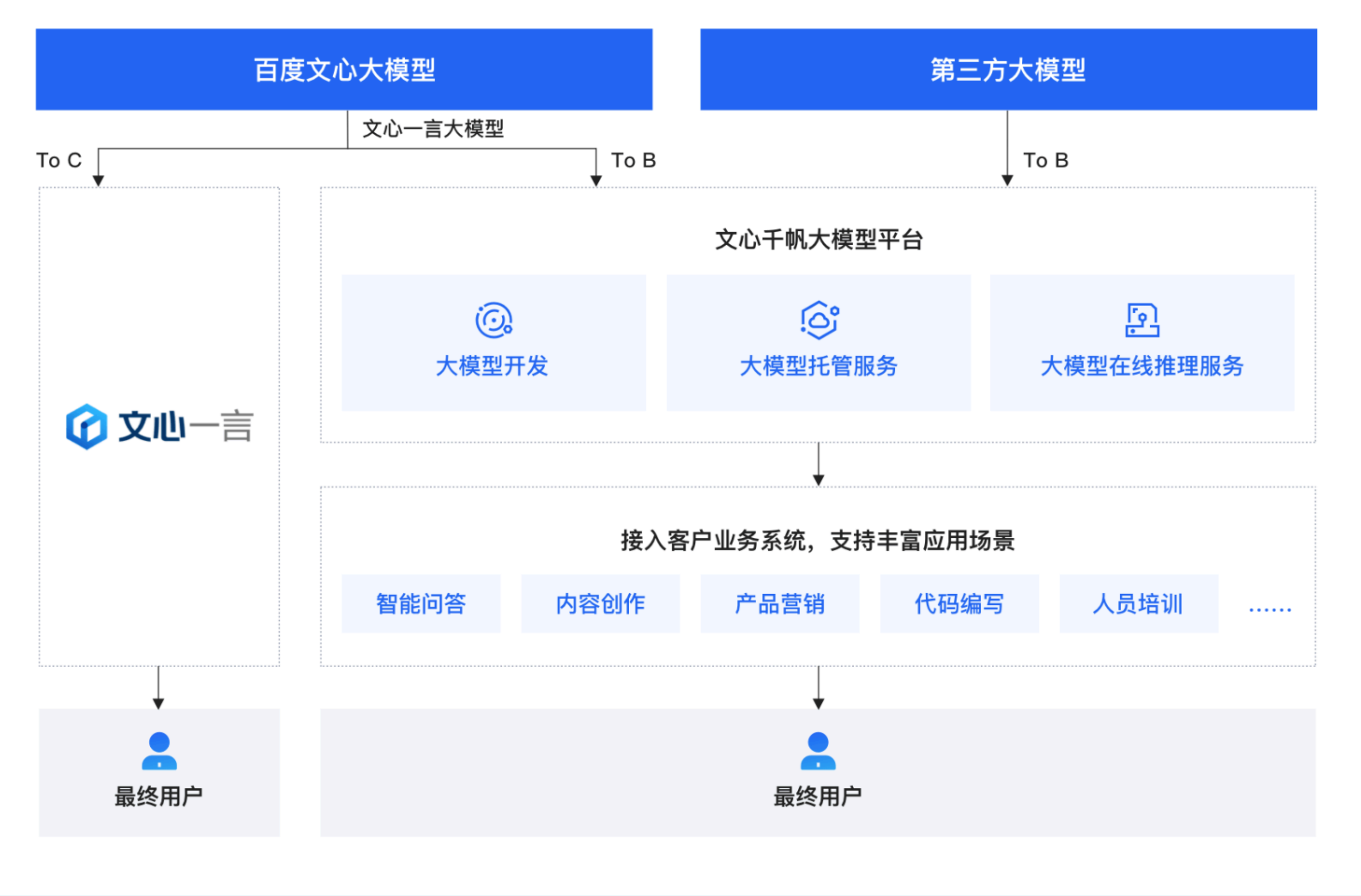
Introduction to Baidu AI Cloud

Which services we will use?

- Specific services: 百度智能云提供的 OCR 服务
- 文字识别服务 (Text Recognition Services):
    - 通用文字识别 (General Text Recognition)
    - 表格文字识别 (Table Text Recognition)
    - 网络图片文字识别 (Web Image Text Recognition)
    - 手写文字识别 (Handwriting Recognition)
    - 文档图像增强 (Document Image Enhancement)
- General AI services:
    - 千帆大模型服务平台 (Qianfan LLM Platform)
    - 文心一言 (ERNIE Bot)

Introduction to Baidu AI Cloud

千帆大模型服务平台 VS 文心一言大模型



百度文心大模型

第三方大模型

文心一言大模型

To C

To B

To B

文心一言

文心千帆大模型平台

大模型开发

大模型托管服务

大模型在线推理服务

接入客户业务系统，支持丰富应用场景

智能问答

内容创作

产品营销

代码编写

人员培训

......

最终用户

最终用户

Introduction to Baidu AI Cloud

Which services we will use?

- Question: 千帆大模型服务平台 VS 表格文字识别 VS 文心一言大模型
  - 千帆大模型服务平台：提供多种 AI 的综合服务，适合商用用户及通用类场景等。
  - 表格文字识别：专门用于表格文字识别的 OCR 服务。
  - 文心一言大模型：针对用户提供自然语言处理和生成服务的 LLM 模型。
- Answer: For structured data extraction, use specialized table OCR service. For more complex tasks involving understanding and generation, try ERNIE Bot (文心一言).

PaddleOCR with Baidu AI Cloud

- There are three ways to use Baidu AI Cloud:

1. Web Console
2. API(Application Programming Interface)
3. SDK(Software Development Kit)

Introduction to Baidu AI Cloud

项目	Web Console	API	SDK
访问	浏览器访问	通过 HTTP 请求直接调用服务，需要自行构建请求和解析响应	通过预编译的库或包集成服务，提供更高层次的抽象和封装
便捷性	无需编程知识，直接通过网页界面进行操作，非常便捷	需要网络编程知识，处理请求和响应的细节，相对繁琐	简化调用过程，提供示例代码和文档，降低使用难度
功能	提供全面的云资源管理和监控功能，以及特定的网页应用	提供基础服务调用功能，可能需要组合多个 API 来满足复杂需求	可能包含额外功能和工具，如异常处理、异步调用等
开发维护	无需编写和维护代码，降低了开发成本，只需关注业务逻辑	需要自行处理细节，管理不同版本的 API，维护成本较高	降低维护成本，提供跨平台和语言的支持，更新和维护更高效
适用用户	适用于需要快速上手、无需编程知识的用户	适用于需要高度定制化和自动化处理的开发者	适用于快速集成和调用的开发者，或希望在不同平台间共享代码的团队

Introduction to Baidu AI Cloud

workflow for using Baidu AI Cloud

1. Start with Web Console to understand capabilities
2. Review API documentation for your use case
3. Test with small dataset using SDK/API
4. Scale up to full research dataset

- Documentation:
  - 百度智能云 OpenAPI
  - 百度智能云 SDK 中心

Getting Started with Baidu AI Cloud

Step-by-Step: Using Baidu AI Cloud

Setup Process

1. 注册百度账号 (Register Baidu Account)
   - Visit cloud.baidu.com
   - Create account (may require Chinese phone number)
2. 创建应用 (Create Application)
   - Navigate to OCR service console
   - Create new application for your project
3. 获取 API Key 和 Secret Key (Get API Credentials)
   - Copy your API Key and Secret Key
   - Important: Keep these credentials secure!

Step-by-Step: Using Baidu AI Cloud

Implementation Process

1. 调用 API/SDK (Call API/SDK)
   - Study relevant documentation
   - Choose between API (HTTP requests) or SDK (Python/Java libraries)
   - Implement authentication
   - Send OCR requests
2. 查看调用情况和结果 (Monitor Usage and Results)
   - Check API quota and usage
   - Verify OCR accuracy
   - Process results into structured data