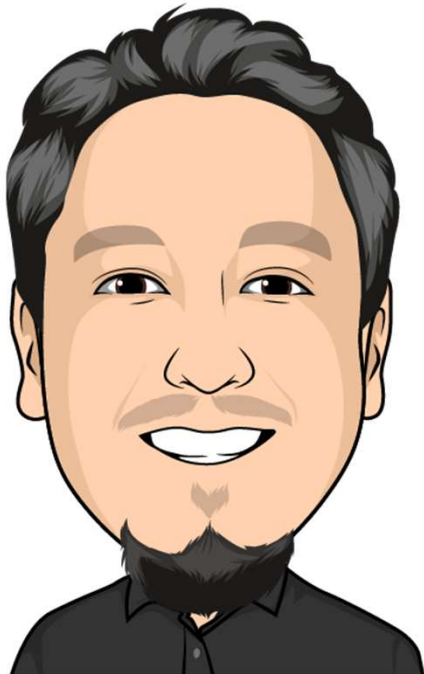# Automated information extraction with Azure AI Document Intelligence.

Originally from Mexico.

Tech Lead at Geneca.

Spend time with family, movies, videogames, soccer.

@thesoccerdev

drkclw

samueljgomez

Thank you to our sponsors!

Agenda.

- Azure AI offerings
    - Cognitive Services
    - Applied AI services

- Azure AI Document Intelligence
    - Model options
    - Input requirements
    - Data privacy

- Demo

Cognitive services.

💬 Speech

🗨 Language

👁 Vision

◆ Decision

Applied Services API.

Services for common business problems

Document intelligence (Form recognizer)

Immersive reader

Metrics advisor

Video indexer

Applied Services API.

Cognitive search

Bot service

Model options.

- Document analysis models.
  - Read
  - Layout
  - General document

- Prebuilt models.

- Custom models.

Document analysis models

Read model.

- Extract text from documents (print and handwritten).
  - Detects paragraphs, text lines, words, locations and languages.

- Higher resolution than Vision Read API.

- Underlying OCR engine for other models.

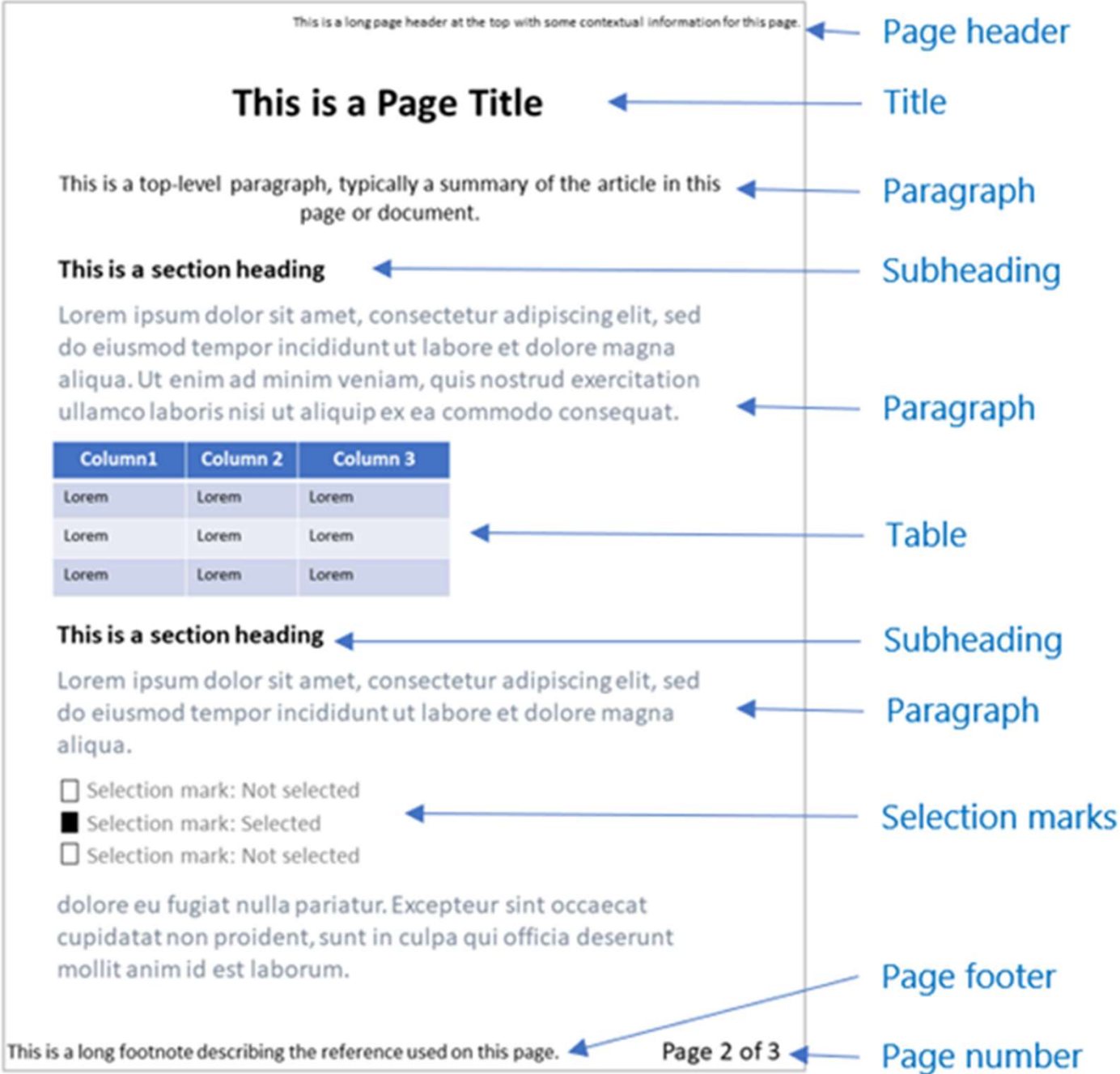# Read model development options.

Model	Resources	Model ID
Read model	• **Document Intelligence Studio** • **REST API** • **C# SDK** • **Python SDK** • **Java SDK** • **JavaScript**	**prebuilt-read**

Layout model.

- Extract text **and layout** to return structured data representations.

- Text roles in documents.
    - **Geometric roles**: Text, tables, and selection marks are examples of geometric roles.
    - **Logical roles**: Titles, headings, and footers are examples of logical roles.

- Combines OCR with deep learning models.

This is a long page header at the top with some contextual information for this page. → **Page header**

This is a Page Title
→ **Title**

This is a top-level paragraph, typically a summary of the article in this page or document. → **Paragraph**

This is a section heading
→ **Subheading**

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. → **Paragraph**

Column1	Column 2	Column 3
Lorem	Lorem	Lorem
Lorem	Lorem	Lorem
Lorem	Lorem	Lorem

→ **Table**

This is a section heading
→ **Subheading**

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. → **Paragraph**

☐ Selection mark: Not selected
■ Selection mark: Selected
☐ Selection mark: Not selected
→ **Selection marks**

dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

This is a long footnote describing the reference used on this page. → **Page footer**

Page 2 of 3 → **Page number**

Layout model development options.

Feature	Resources	Model ID
Layout model	• **Document Intelligence Studio** • **REST API** • **C# SDK** • **Python SDK** • **Java SDK** • **JavaScript SDK**	**prebuilt-layout**

General document model.

- Extract text, layout **and key-value** pairs from documents.

- Pretrained model.

- Supports structured, semi-structured and unstructured documents.

Key-value pairs

- For structured document the label and value for a field.

- For unstructured documents they are based on the text in the paragraph (like date in a contract).

General document model development options.

Feature	Resources	Model ID
General document model	•**Document Intelligence Studio** •**REST API** •**C# SDK** •**Python SDK** •**Java SDK** •**JavaScript SDK**	**prebuilt-document**

Prebuilt models

GENECA

Available models.

- Invoice.
  - Extract customer and vendor details.

- Receipt.
  - Extract sales transaction details.

- Identity.

- Insurance card.

Available models.

- W2.
  - Extract taxable compensation details.

- Business card.
  - Extract business contact details.

- Contract.
  - Extract agreement and party details.

- US Tax 1098-E Form.
  - Extract student loan interest details.

Available models.

- US Tax 1098 form.
  - Extract mortgage interest details.

- US Tax 1098-T form.
  - Extract qualified tuition details.

Development options.

Feature	Resources	Model ID
Prebuilt models	• **Document Intelligence Studio** • **REST API** • **C# SDK** • **Python SDK** • **Java SDK*** • **JavaScript SDK***	**prebuilt-invoice** **prebuilt-receipt** **prebuilt-idDocument** **prebuilt-healthInsuranceCard.us** **prebuilt-tax.us.w2** **prebuilt-contract** **prebuilt-businessCard** **prebuilt-tax.us.1098** **prebuilt-tax.us.1098E** **prebuilt-tax.us.1098T**

Custom models

GENECA

Custom template.

- Extract data from static layouts.

- Capabilities.
  - Form fields.
  - Selection marks.
  - Tabular fields (tables).
  - Signature.
  - Selected regions.

Custom neural.

- Extract data from mixed-type documents.

- Supports structured, semi-structured and unstructured data.

Custom composed.

- Extract data using a collection of models.

- .

Classification model.

- Identify document type before calling extraction model.

- .

Input requirements.

File formats.

Model	PDF	Image (JPEG/JPG, PNG, BMP, and TIFF)	Microsoft Office: Word (DOCX), Excel (XLS), PowerPoint (PPT), and HTML
Read	Yes	Yes	Yes
Layout	Yes	Yes	No
General Document	Yes	Yes	No
Prebuilt	Yes	Yes	No
Custom	Yes	Yes	No

Model restrictions.

- For PDF and TIFF, up to 2000 pages can be processed (2 for free tier).

- .File size less than 500 MB (4 MB for free tier).

- Image dimensions must be between 50 x 50 pixels and 10,000 px x 10,000 pixels.

Model restrictions.

- PDFs can't be password protected.

- Minimum size of text is 8-point text at 150 DPI.

Data privacy.

- Authentication.

- Data secured in transit.

- Encrypted input data.

Data privacy.

- Data stored.
  - Data and extracted results are stored temporarily.
  - Output for custom trained models and models themselves are stored temporarily as well.
  - Input data and results are deleted within 24 hours and not used for other purposes.

- Containers.

DEMO

> Let this code compile,
>
> In not an hour, but a little while.
>
> Allow this application to run with speed and zest,
>
> as this will keep users glued to my test.
>
> If my code blows up in my face,
>
> do it with style and grace.
>
> And last but not least, when all is said and run,
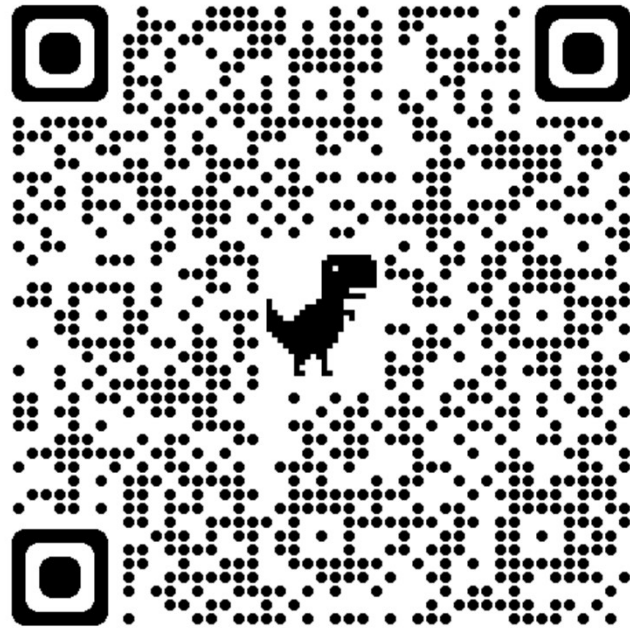>
> let thy requirements have been correctly done.

Questions?

Useful links.

- https://learn.microsoft.com/en-us/azure/ai-services/document-intelligence/?view=doc-intel-3.1.0

- https://learn.microsoft.com/en-us/legal/cognitive-services/document-intelligence/data-privacy-security

- https://formrecognizer.appliedai.azure.com/studio

Useful links.

@thesoccerdev

drkclw

samueljgomez