AWS re:Invent

SVS210-2

Automatically scale a serverless app with Amazon Textract & MongoDB

Diana Esteves
Senior Engineer
MongoDB

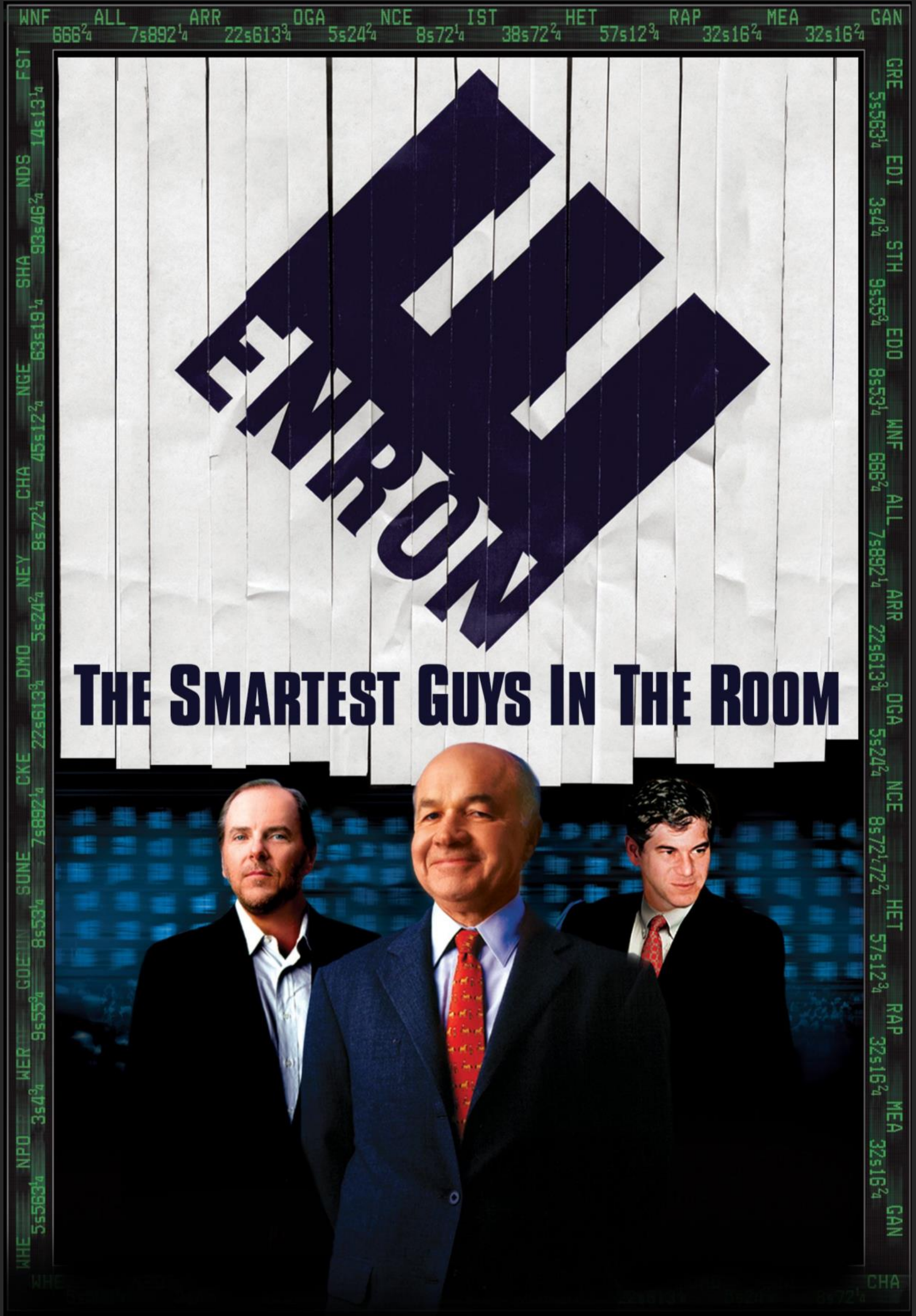
Ralph Capasso
Lead Engineer
MongoDB

Agenda

AWS machine learning (ML) and optical character recognition (OCR) services

MongoDB data platform

Building the app



ENRON
WNF 666 ALL 7s892 ARR 22s613 OGA 5s24 NCE 8s72 IST 38s72 HET 57s12 RAP 32s16 MEA 32s16 GAN

THE SMARTEST GUYS IN THE ROOM

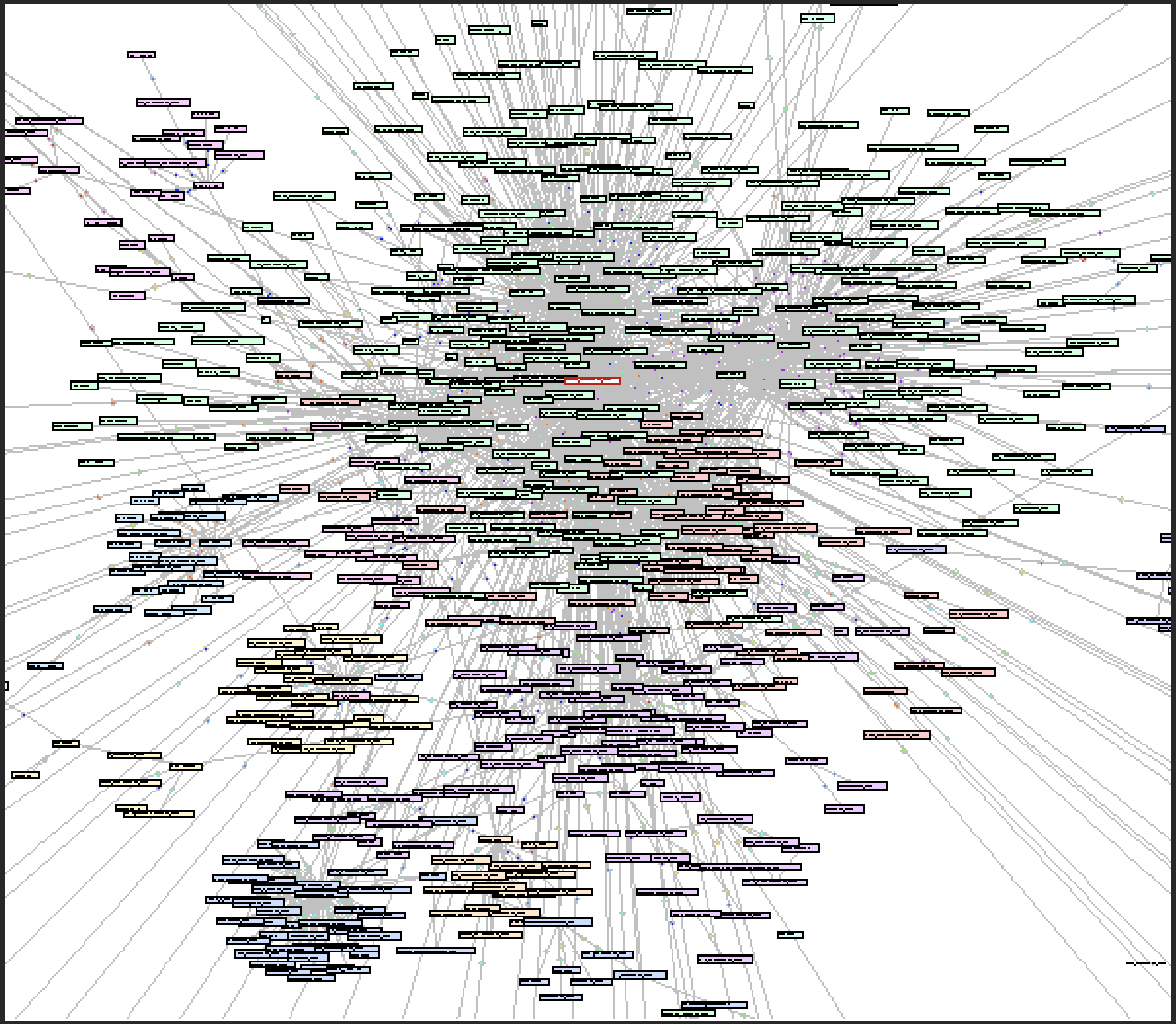
Problem – Enron discovery

150+ people

~500K emails

4 years

Problem – Enron corpus directory structure



Investigators & goal: Extract incriminating evidence

Investigators & goal: Extract incriminating evidence

Document Corpus

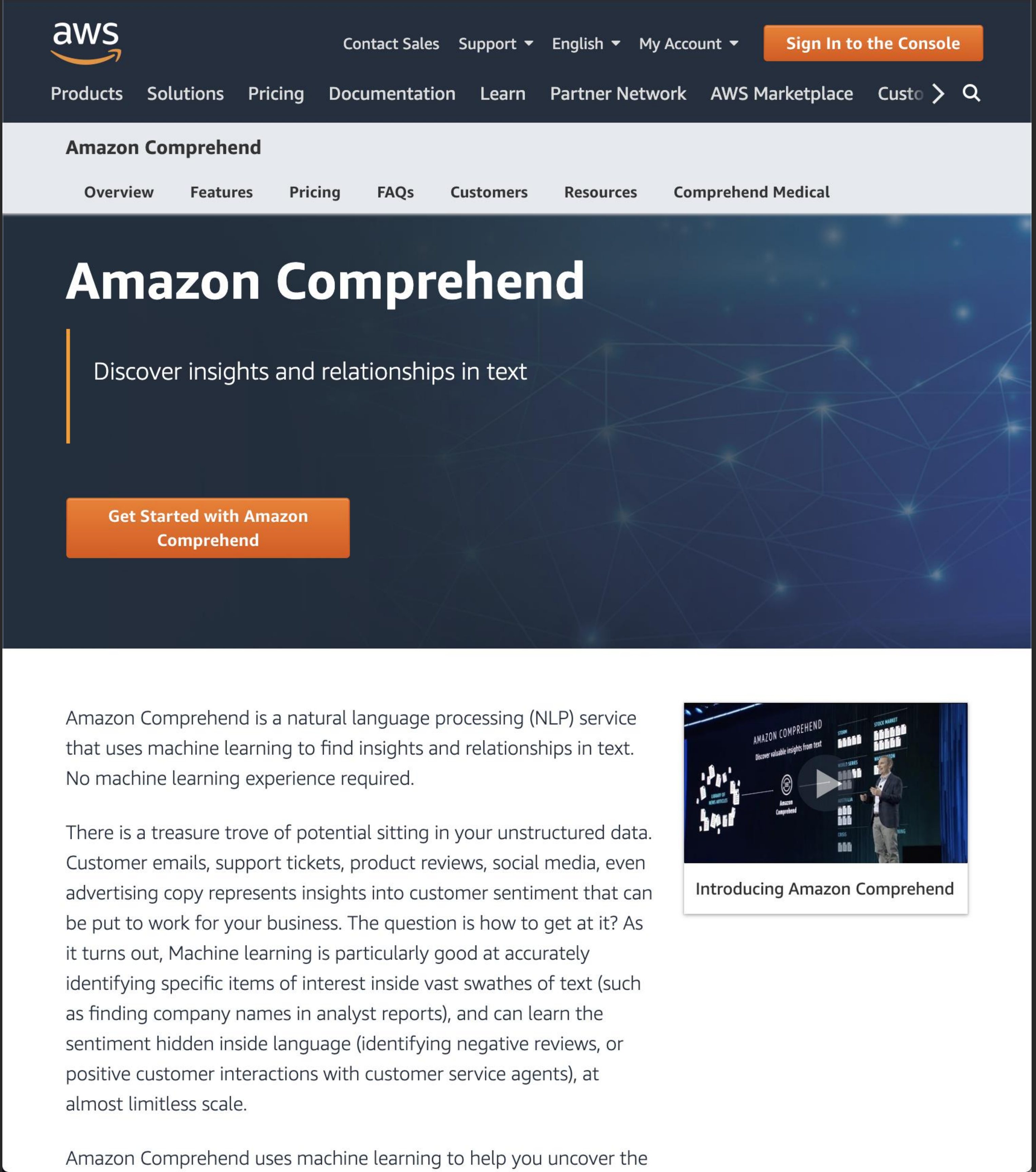
Entities & Key phrases

Investigators & goal: Extract incriminating evidence

Document Corpus

Amazon Comprehend

Entities & Key phrases



Amazon Comprehend

Discover insights and relationships in text

Amazon Comprehend is a natural language processing (NLP) service that uses machine learning to find insights and relationships in text. No machine learning experience required.

Introducing Amazon Comprehend

There is a treasure trove of potential sitting in your unstructured data. Customer emails, support tickets, product reviews, social media, even advertising copy represents insights into customer sentiment that can be put to work for your business. The question is how to get at it? As it turns out, Machine learning is particularly good at accurately identifying specific items of interest inside vast swathes of text (such as finding company names in analyst reports), and can learn the sentiment hidden inside language (identifying negative reviews, or positive customer interactions with customer service agents), at almost limitless scale.

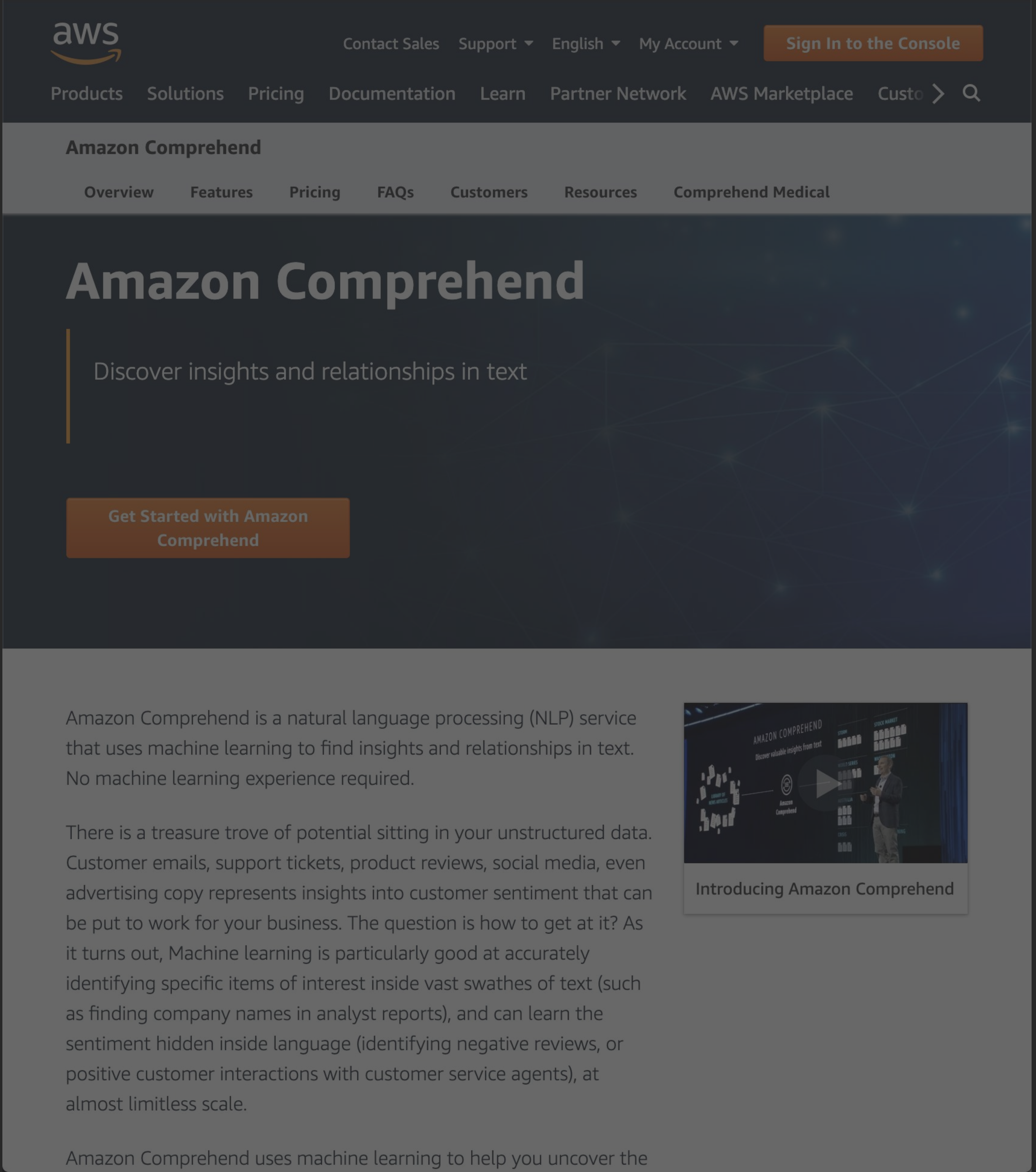
Amazon Comprehend uses machine learning to help you uncover the

Analyze data quickly

Extract entities, phrases, sentiment, syntax, and topics

Dramatically reduce analysis time spent

No ML experience required



Amazon Comprehend

Discover insights and relationships in text

Amazon Comprehend is a natural language processing (NLP) service that uses machine learning to find insights and relationships in text. No machine learning experience required.

There is a treasure trove of potential sitting in your unstructured data. Customer emails, support tickets, product reviews, social media, even advertising copy represents insights into customer sentiment that can be put to work for your business. The question is how to get at it? As it turns out, Machine learning is particularly good at accurately identifying specific items of interest inside vast swathes of text (such as finding company names in analyst reports), and can learn the sentiment hidden inside language (identifying negative reviews, or positive customer interactions with customer service agents), at almost limitless scale.

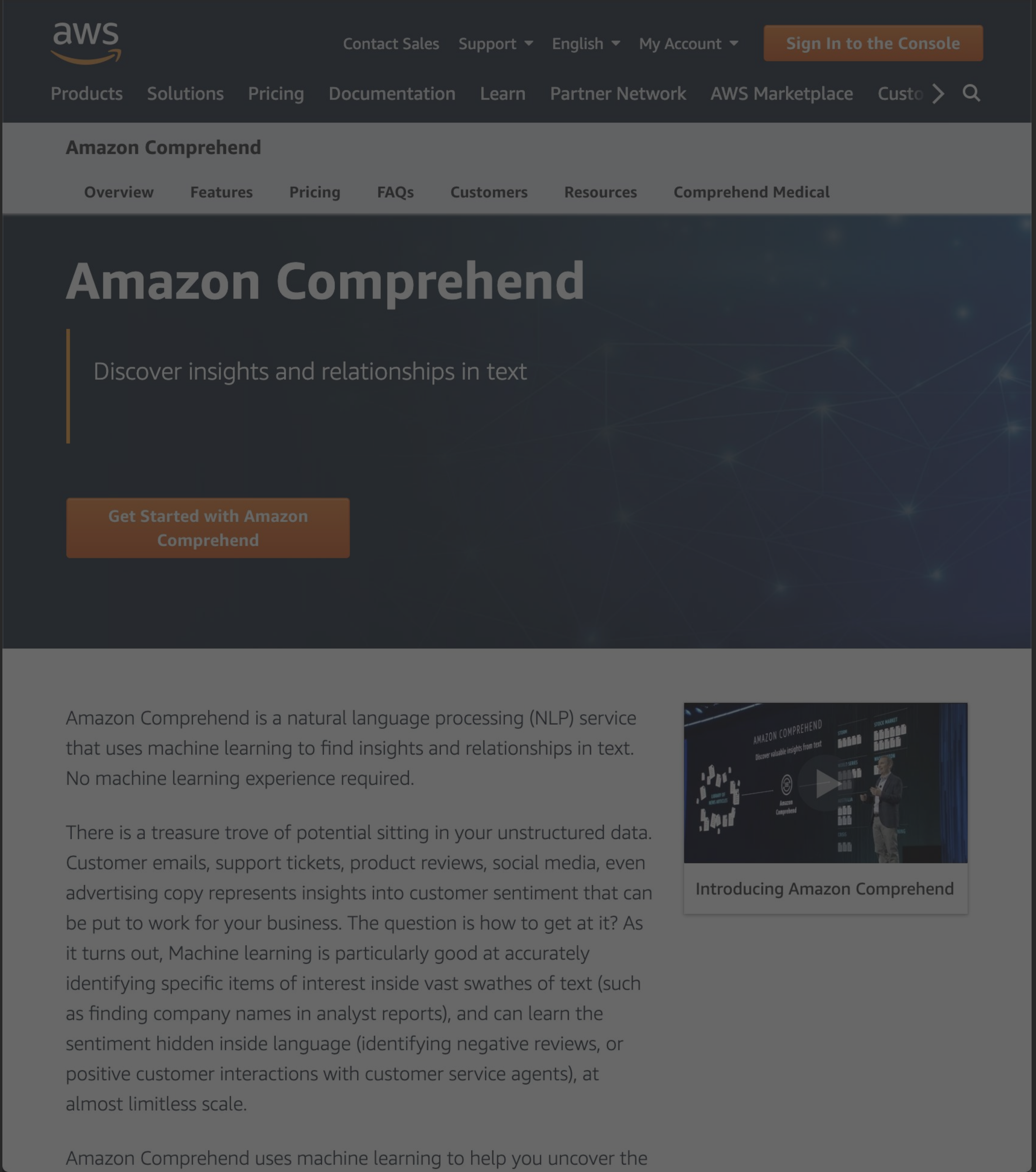
Amazon Comprehend uses machine learning to help you uncover the

Analyze data quickly

Extract entities, phrases, sentiment, syntax, and topics

Dramatically reduce analysis time spent

No ML experience required



Amazon Comprehend

Discover insights and relationships in text

Amazon Comprehend is a natural language processing (NLP) service that uses machine learning to find insights and relationships in text. No machine learning experience required.

There is a treasure trove of potential sitting in your unstructured data. Customer emails, support tickets, product reviews, social media, even advertising copy represents insights into customer sentiment that can be put to work for your business. The question is how to get at it? As it turns out, Machine learning is particularly good at accurately identifying specific items of interest inside vast swathes of text (such as finding company names in analyst reports), and can learn the sentiment hidden inside language (identifying negative reviews, or positive customer interactions with customer service agents), at almost limitless scale.

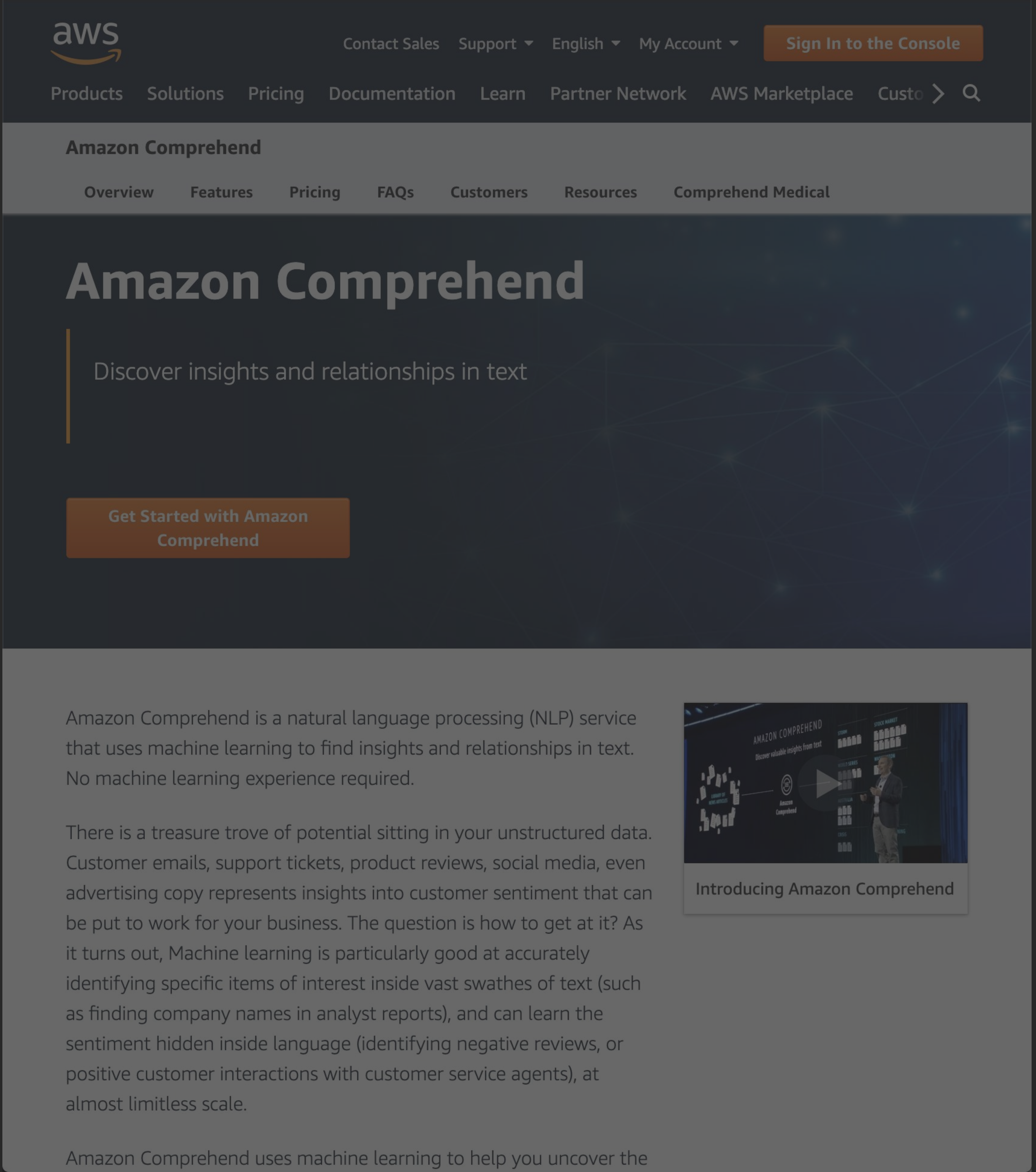
Amazon Comprehend uses machine learning to help you uncover the

Analyze data quickly

Extract entities, phrases, sentiment, syntax, and topics

Dramatically reduce analysis time spent

No ML experience required



Amazon Comprehend

Discover insights and relationships in text

Amazon Comprehend is a natural language processing (NLP) service that uses machine learning to find insights and relationships in text. No machine learning experience required.

There is a treasure trove of potential sitting in your unstructured data. Customer emails, support tickets, product reviews, social media, even advertising copy represents insights into customer sentiment that can be put to work for your business. The question is how to get at it? As it turns out, Machine learning is particularly good at accurately identifying specific items of interest inside vast swathes of text (such as finding company names in analyst reports), and can learn the sentiment hidden inside language (identifying negative reviews, or positive customer interactions with customer service agents), at almost limitless scale.

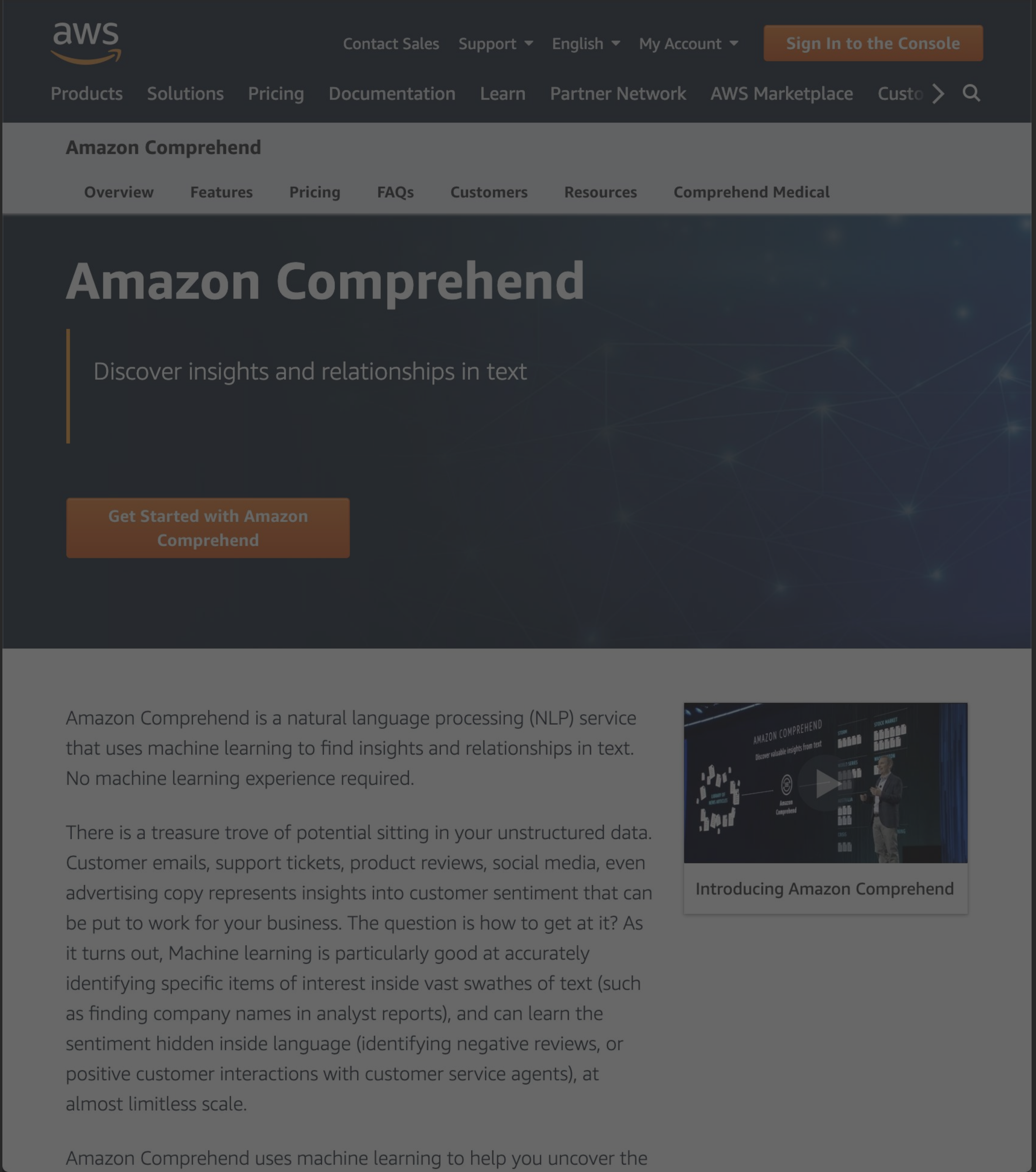
Amazon Comprehend uses machine learning to help you uncover the

Analyze data quickly

Extract entities, phrases, sentiment, syntax, and topics

Dramatically reduce analysis time spent

No ML experience required



Amazon Comprehend

Discover insights and relationships in text

Amazon Comprehend is a natural language processing (NLP) service that uses machine learning to find insights and relationships in text. No machine learning experience required.

There is a treasure trove of potential sitting in your unstructured data. Customer emails, support tickets, product reviews, social media, even advertising copy represents insights into customer sentiment that can be put to work for your business. The question is how to get at it? As it turns out, Machine learning is particularly good at accurately identifying specific items of interest inside vast swathes of text (such as finding company names in analyst reports), and can learn the sentiment hidden inside language (identifying negative reviews, or positive customer interactions with customer service agents), at almost limitless scale.

Amazon Comprehend uses machine learning to help you uncover the

Introducing Amazon Comprehend

Analyze data quickly

Extract entities, phrases, sentiment, syntax, and topics

Dramatically reduce analysis time spent

No ML experience required

Additional goal

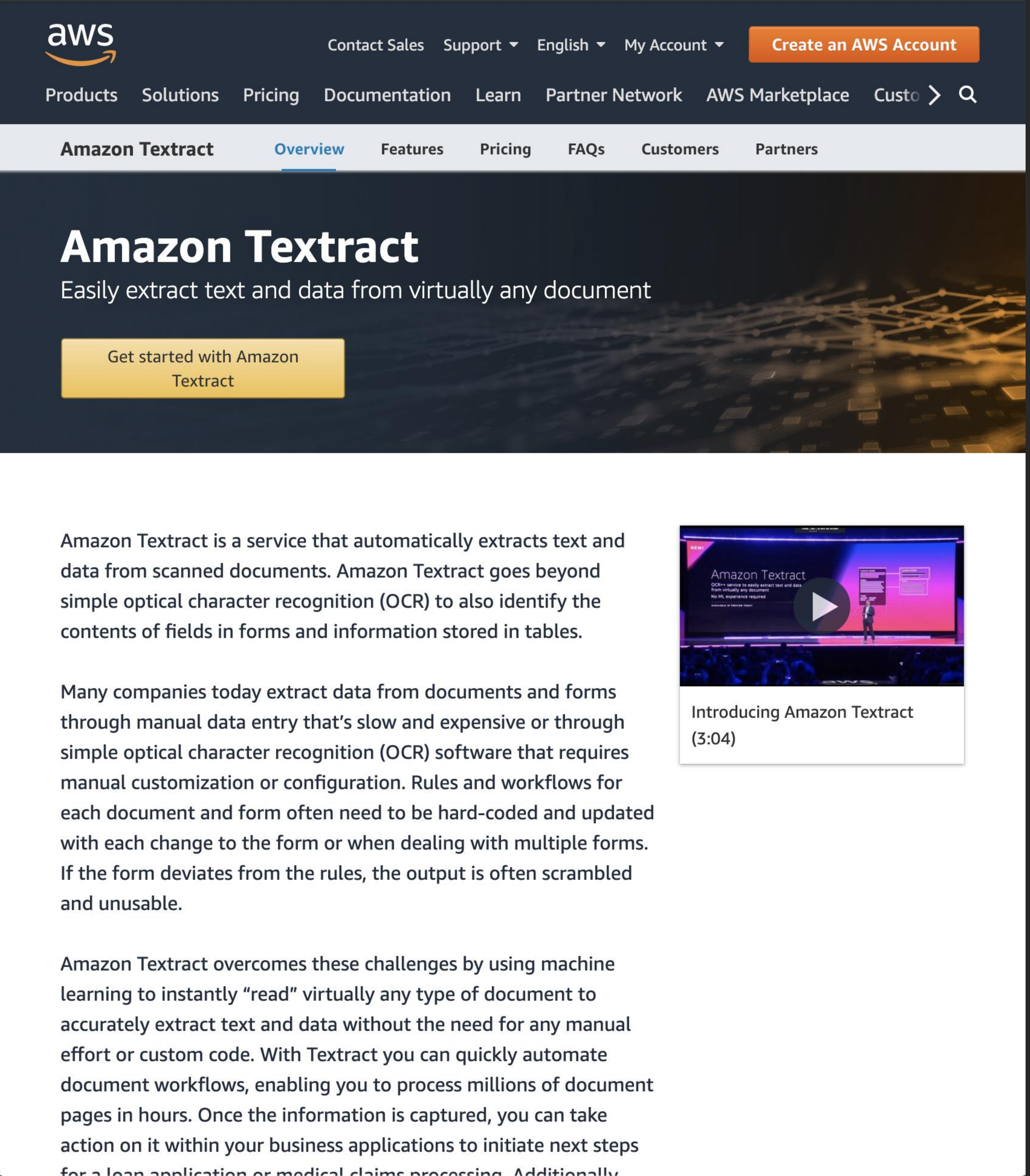
Document corpus

Document scans

Amazon Comprehend

Amazon Textract

Entities & key phrases



Amazon Textract

Easily extract text and data from virtually any document

Amazon Textract is a service that automatically extracts text and data from scanned documents. Amazon Textract goes beyond simple optical character recognition (OCR) to also identify the contents of fields in forms and information stored in tables.

Introducing Amazon Textract (3:04)

Many companies today extract data from documents and forms through manual data entry that's slow and expensive or through simple optical character recognition (OCR) software that requires manual customization or configuration. Rules and workflows for each document and form often need to be hard-coded and updated with each change to the form or when dealing with multiple forms. If the form deviates from the rules, the output is often scrambled and unusable.

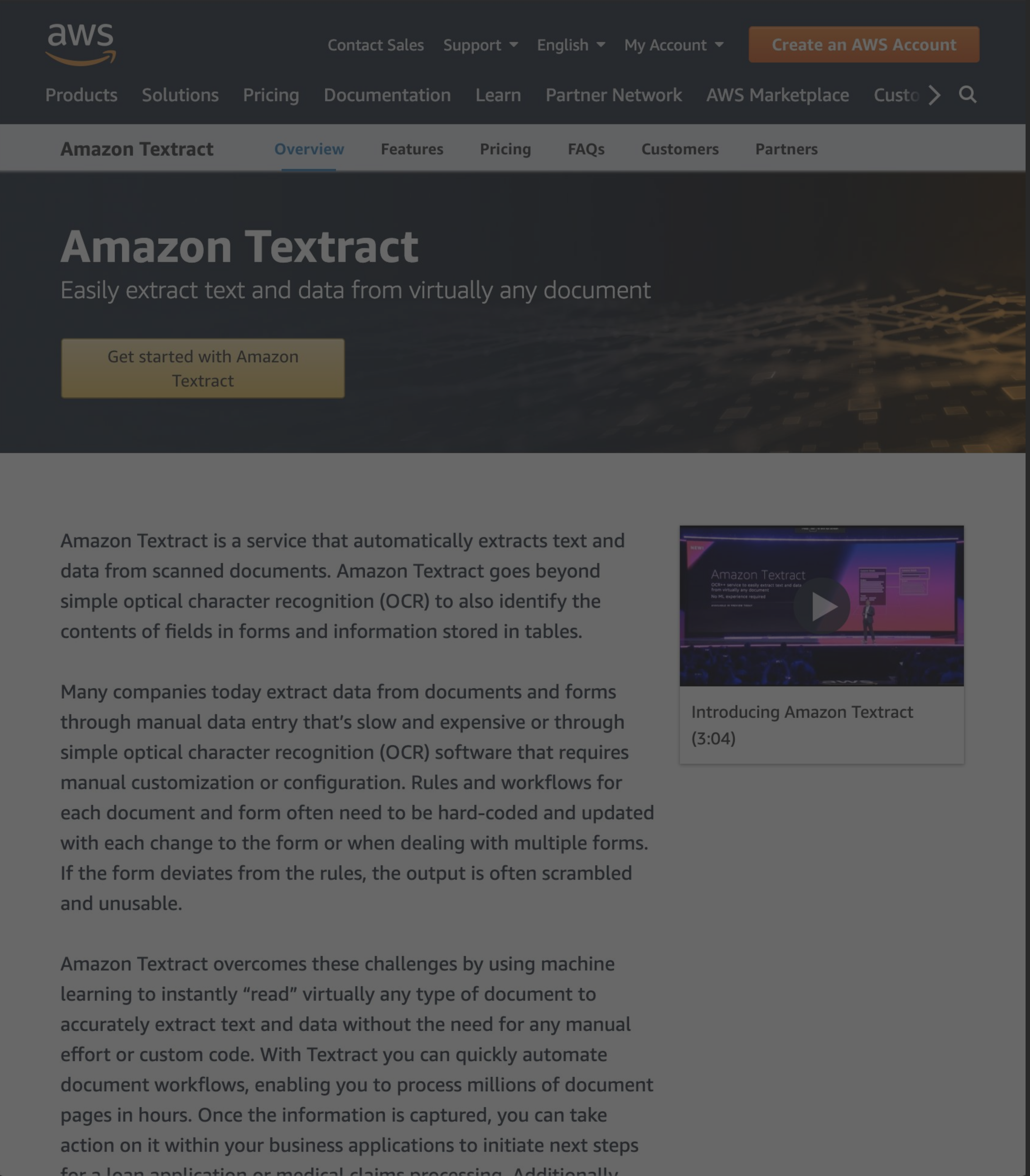
Amazon Textract overcomes these challenges by using machine learning to instantly "read" virtually any type of document to accurately extract text and data without the need for any manual effort or custom code. With Textract you can quickly automate document workflows, enabling you to process millions of document pages in hours. Once the information is captured, you can take action on it within your business applications to initiate next steps for a loan application or medical claims processing. Additionally

Extract data quickly and accurately

Eliminate manual effort

Lower document processing costs

No OCR experience required



Amazon Textract

Easily extract text and data from virtually any document

Amazon Textract is a service that automatically extracts text and data from scanned documents. Amazon Textract goes beyond simple optical character recognition (OCR) to also identify the contents of fields in forms and information stored in tables.

Many companies today extract data from documents and forms through manual data entry that's slow and expensive or through simple optical character recognition (OCR) software that requires manual customization or configuration. Rules and workflows for each document and form often need to be hard-coded and updated with each change to the form or when dealing with multiple forms. If the form deviates from the rules, the output is often scrambled and unusable.

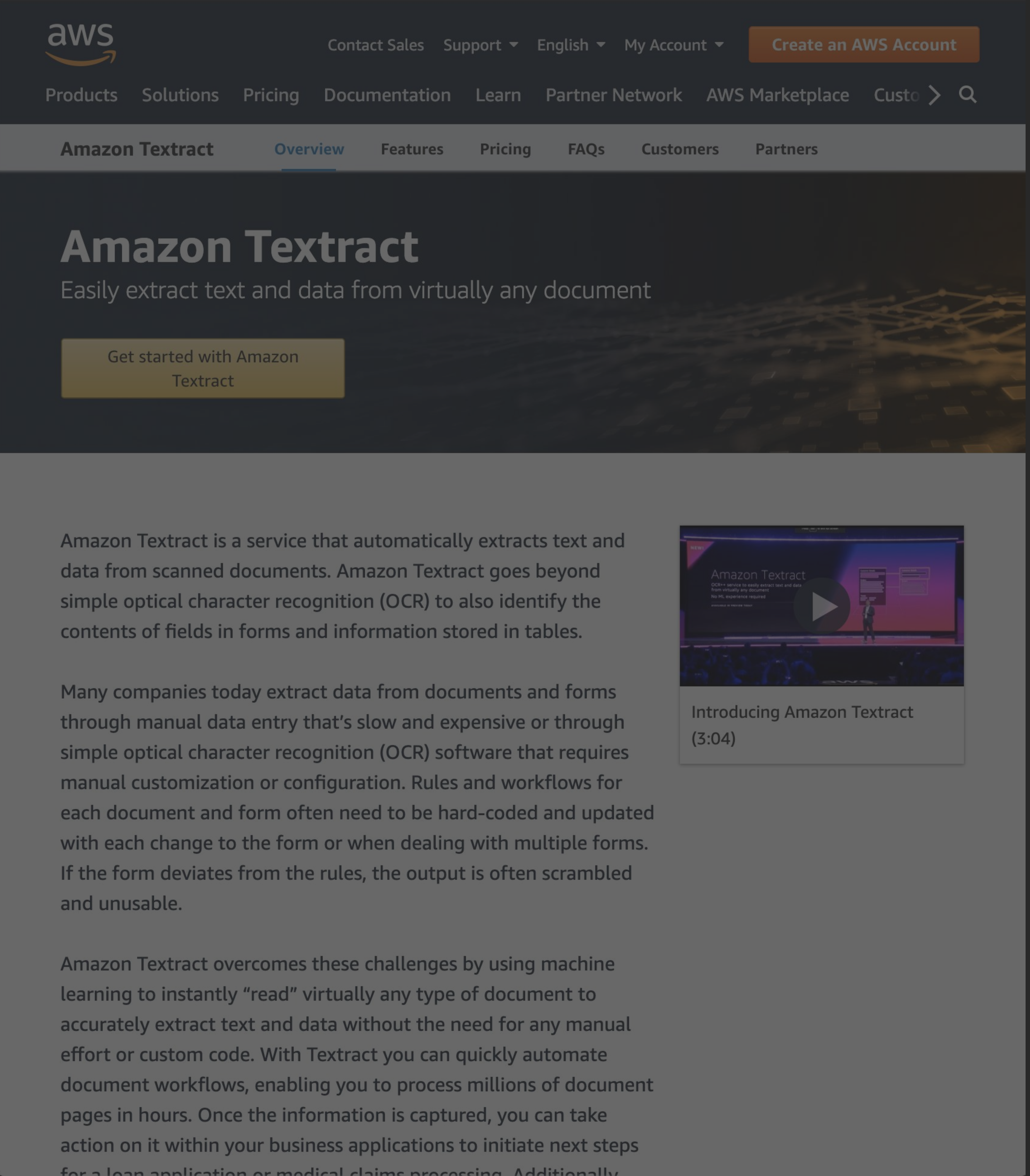
Amazon Textract overcomes these challenges by using machine learning to instantly "read" virtually any type of document to accurately extract text and data without the need for any manual effort or custom code. With Textract you can quickly automate document workflows, enabling you to process millions of document pages in hours. Once the information is captured, you can take action on it within your business applications to initiate next steps for a loan application or medical claims processing. Additionally

Extract data quickly and accurately

Eliminate manual effort

Lower document processing costs

No OCR experience required



Amazon Textract

Easily extract text and data from virtually any document

Amazon Textract is a service that automatically extracts text and data from scanned documents. Amazon Textract goes beyond simple optical character recognition (OCR) to also identify the contents of fields in forms and information stored in tables.

Many companies today extract data from documents and forms through manual data entry that's slow and expensive or through simple optical character recognition (OCR) software that requires manual customization or configuration. Rules and workflows for each document and form often need to be hard-coded and updated with each change to the form or when dealing with multiple forms. If the form deviates from the rules, the output is often scrambled and unusable.

Amazon Textract overcomes these challenges by using machine learning to instantly "read" virtually any type of document to accurately extract text and data without the need for any manual effort or custom code. With Textract you can quickly automate document workflows, enabling you to process millions of document pages in hours. Once the information is captured, you can take action on it within your business applications to initiate next steps for a loan application or medical claims processing. Additionally

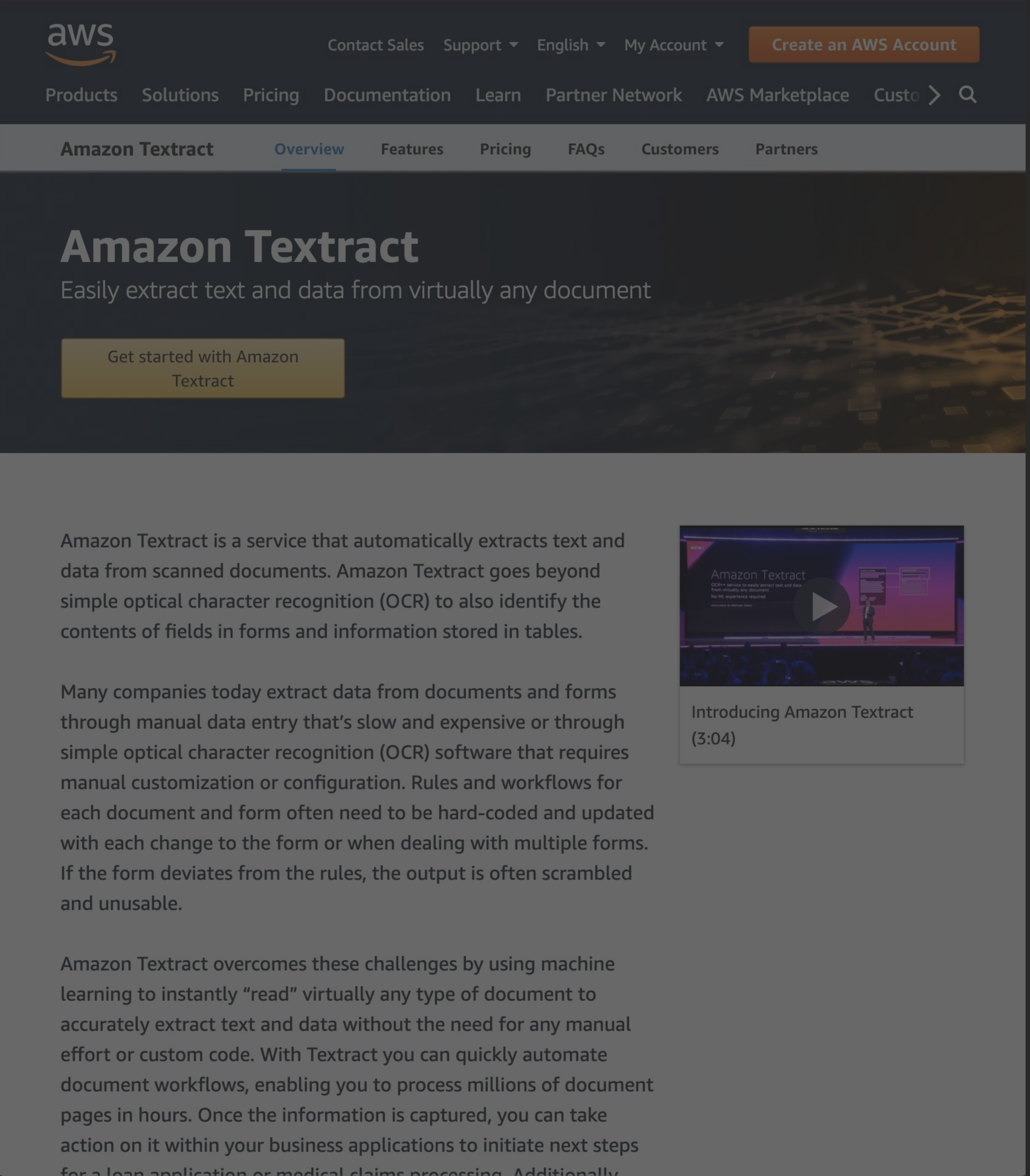
Introducing Amazon Textract (3:04)

Extract data quickly and accurately

Eliminate manual effort

Lower document processing costs

No OCR experience required



Amazon Textract

Easily extract text and data from virtually any document

Get started with Amazon Textract

Amazon Textract is a service that automatically extracts text and data from scanned documents. Amazon Textract goes beyond simple optical character recognition (OCR) to also identify the contents of fields in forms and information stored in tables.

Many companies today extract data from documents and forms through manual data entry that's slow and expensive or through simple optical character recognition (OCR) software that requires manual customization or configuration. Rules and workflows for each document and form often need to be hard-coded and updated with each change to the form or when dealing with multiple forms. If the form deviates from the rules, the output is often scrambled and unusable.

Amazon Textract overcomes these challenges by using machine learning to instantly "read" virtually any type of document to accurately extract text and data without the need for any manual effort or custom code. With Textract you can quickly automate document workflows, enabling you to process millions of document pages in hours. Once the information is captured, you can take action on it within your business applications to initiate next steps for a loan application or medical claims processing. Additionally

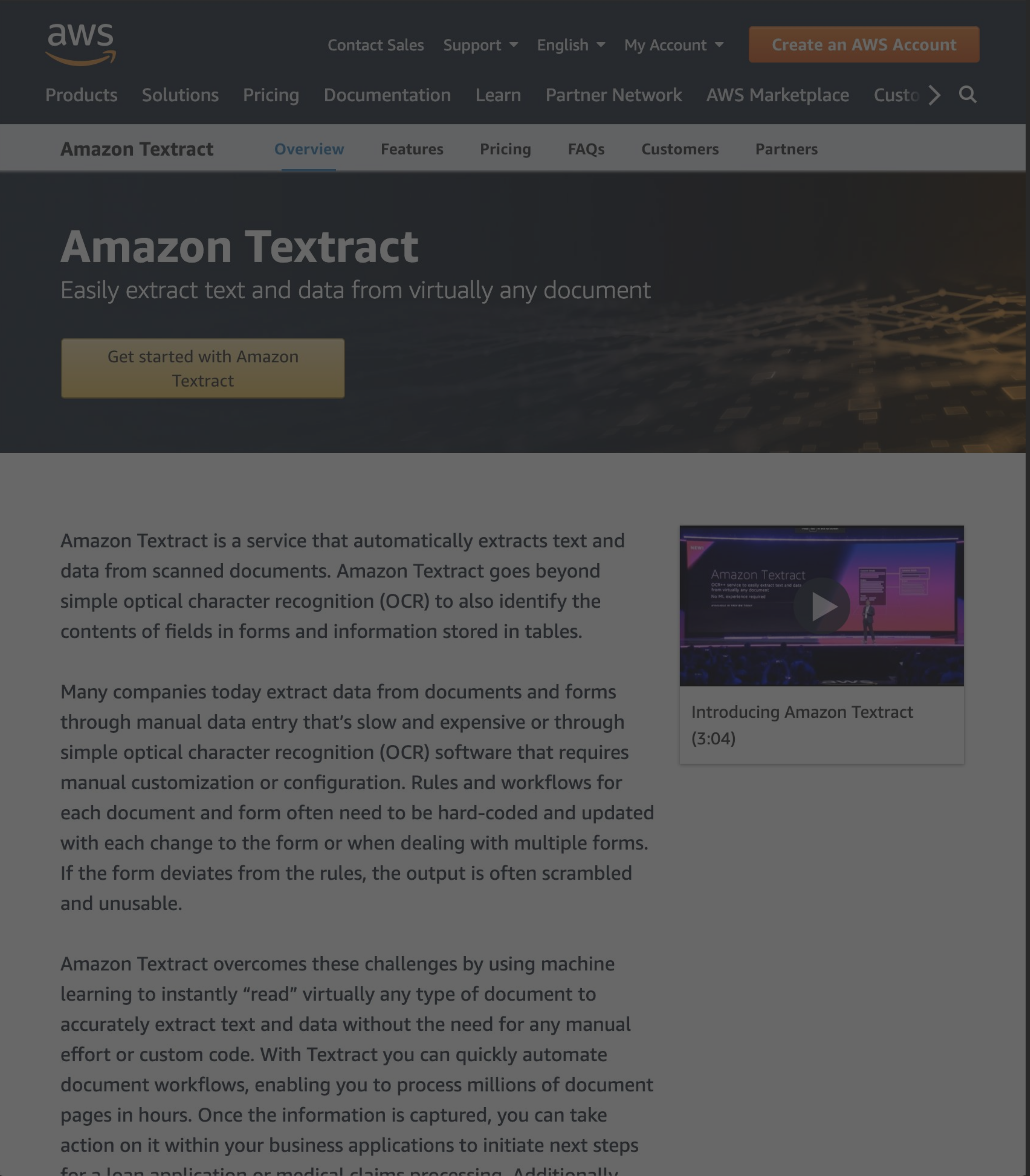
Introducing Amazon Textract (3:04)

Extract data quickly and accurately

Eliminate manual effort

Lower document processing costs

No OCR experience required



Amazon Textract

Easily extract text and data from virtually any document

Amazon Textract is a service that automatically extracts text and data from scanned documents. Amazon Textract goes beyond simple optical character recognition (OCR) to also identify the contents of fields in forms and information stored in tables.

Many companies today extract data from documents and forms through manual data entry that's slow and expensive or through simple optical character recognition (OCR) software that requires manual customization or configuration. Rules and workflows for each document and form often need to be hard-coded and updated with each change to the form or when dealing with multiple forms. If the form deviates from the rules, the output is often scrambled and unusable.

Amazon Textract overcomes these challenges by using machine learning to instantly "read" virtually any type of document to accurately extract text and data without the need for any manual effort or custom code. With Textract you can quickly automate document workflows, enabling you to process millions of document pages in hours. Once the information is captured, you can take action on it within your business applications to initiate next steps for a loan application or medical claims processing. Additionally

Introducing Amazon Textract (3:04)

Extract data quickly and accurately

Eliminate manual effort

Lower document processing costs

No OCR experience required

Solution: Cloud services + machine learning

Amazon
Comprehend

Amazon
Textract

Solution: Cloud services + machine learning

Amazon
S3

Amazon
Comprehend

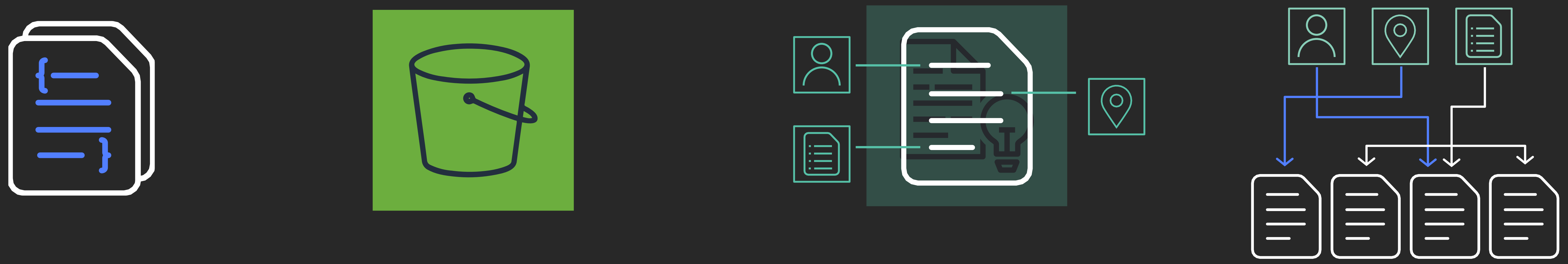
Amazon
Textract

Amazon S3

Document corpus

Amazon S3

Amazon Comprehend



Document corpus

Amazon S3

Amazon Comprehend

Entities & key phrases

Amazon services

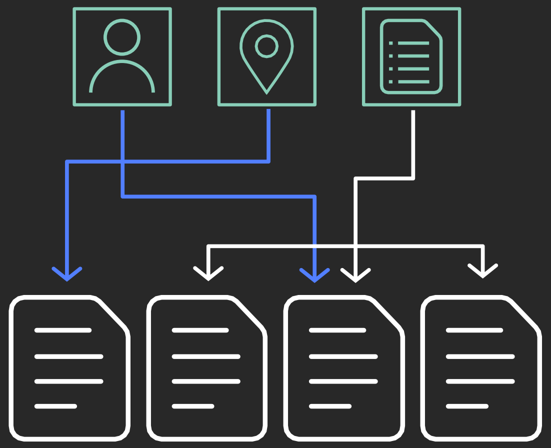
Document corpus

Document scans

Amazon Comprehend

Amazon S3

Amazon Textract



Entities & key phrases

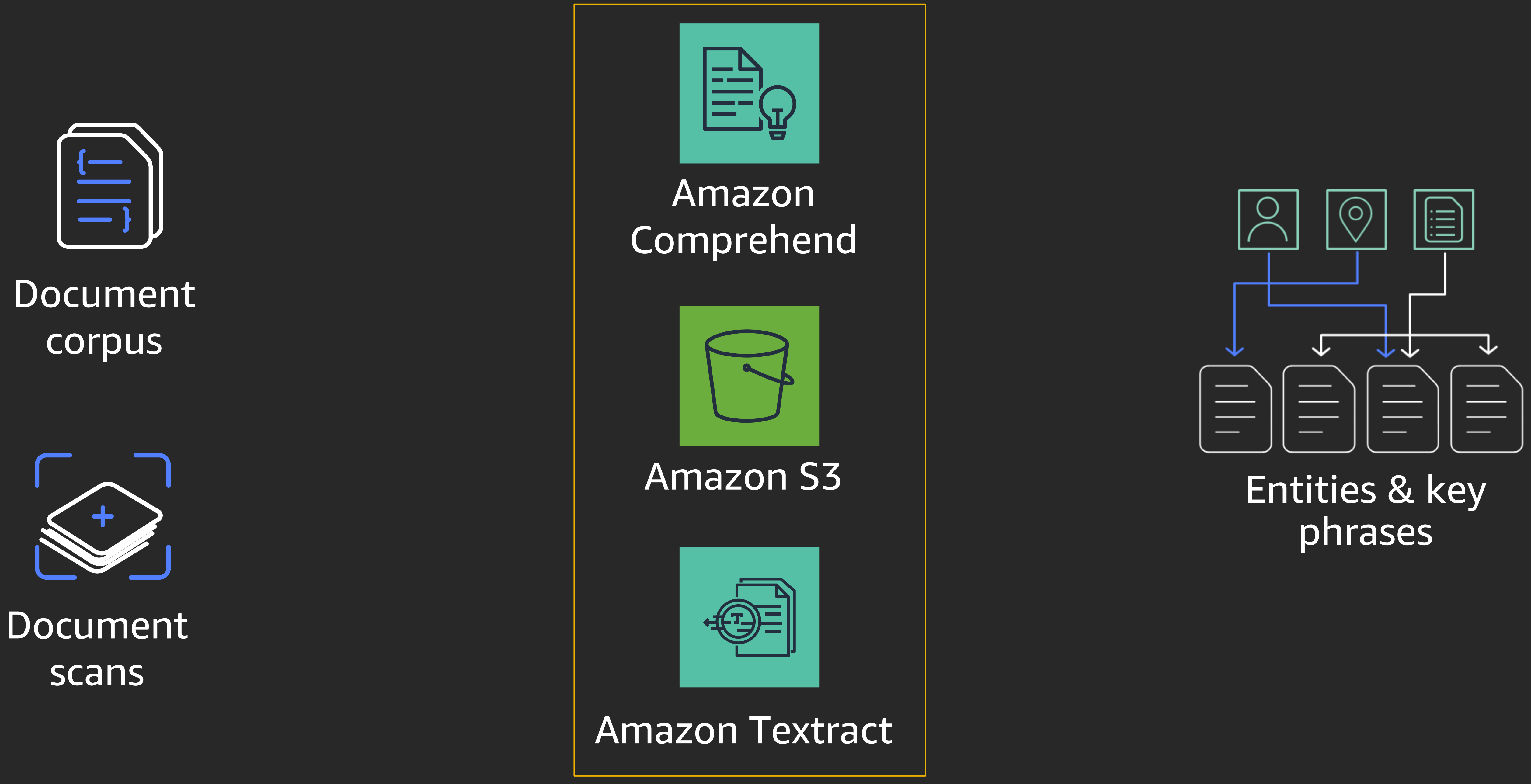
Amazon Textract

Scans of documents

Amazon Textract

Document text

Amazon services



Document corpus

Document scans

Amazon Comprehend

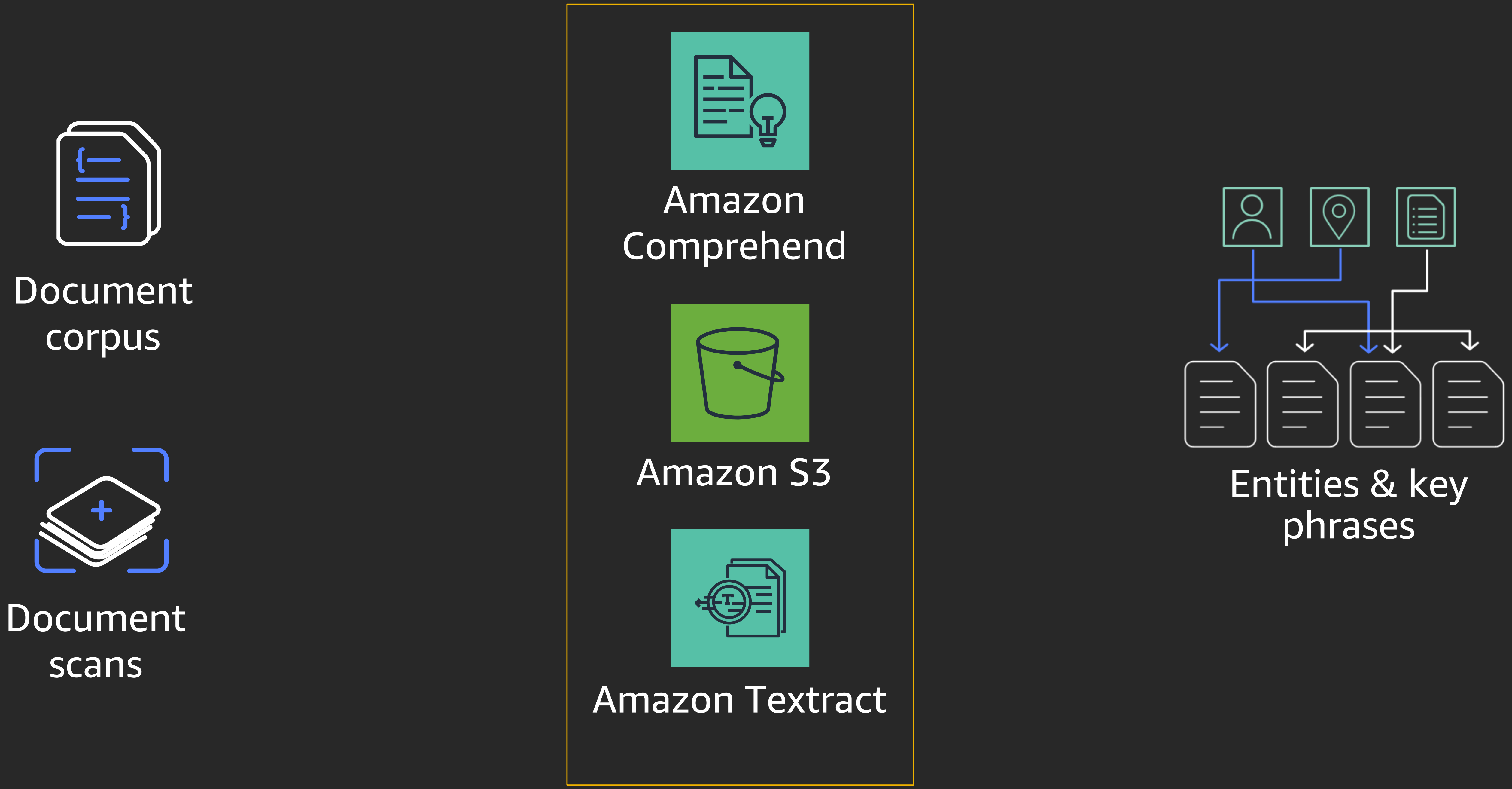
Amazon S3

Amazon Textract

Entities & key phrases

Stitching it together with MongoDB Atlas

Amazon services



Document corpus

Document scans

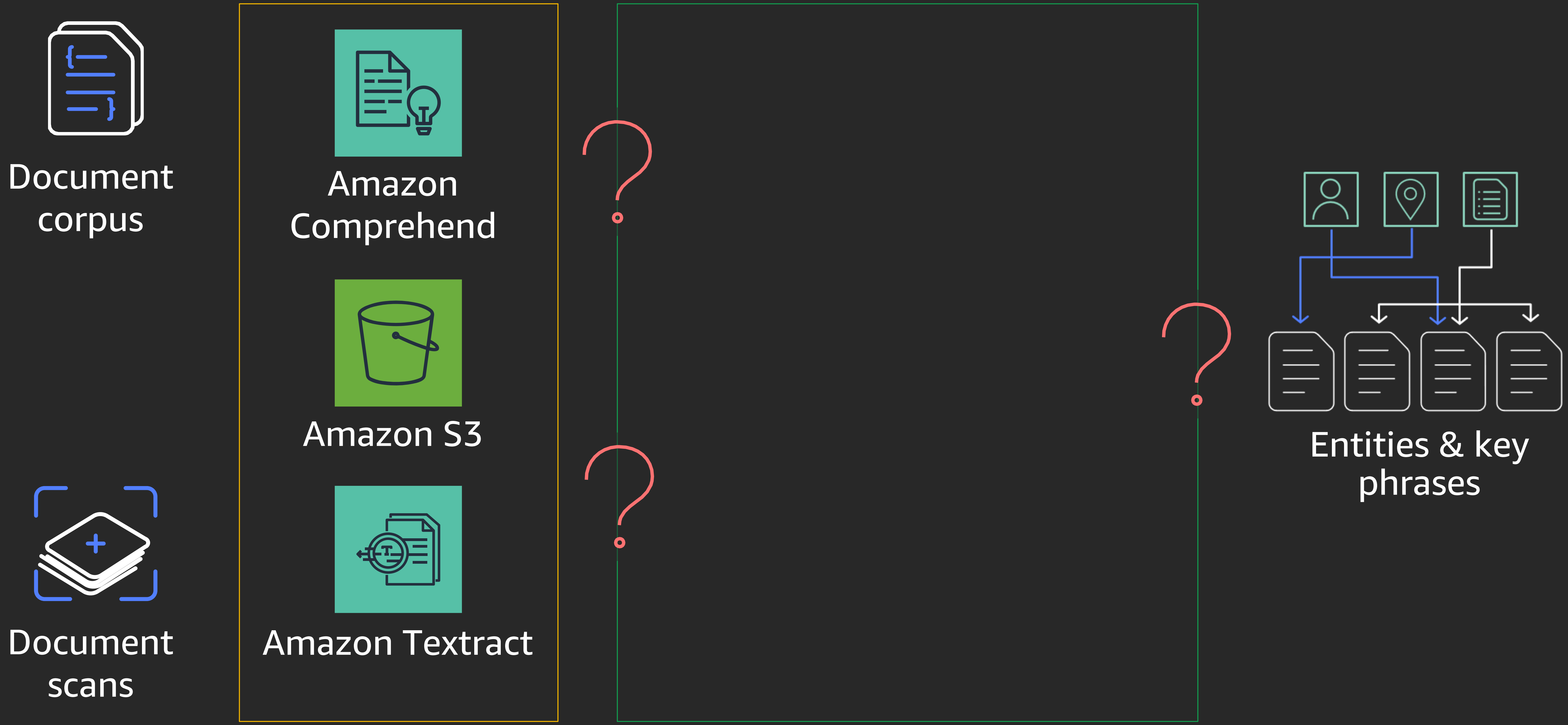
Amazon Comprehend

Amazon S3

Amazon Textract

Entities & key phrases

Stitching it together



Document corpus

Document scans

Amazon Comprehend

Amazon S3

Amazon Textract

Entities & key phrases

Comprehend metadata in Atlas

```
{
  "_id": { "$oid": "5d8918fc5f5d5d78b9985909" },
  "author": "arnold-j",
  "content": "Message-ID: <19171686.1075857585034.JavaMail.evans@thyme>\r\nDate: Fri, 8
Dec 2000 05:05:00 -0800 (PST)\r\nFrom: slafontaine@globalp.com\r\nTo:
john.arnold@enron.com\r\nSubject: re:summer inverses\r\nMime-Version: 1.0\r\nContent-Type:
text/plain; charset=us-ascii\r\nContent-Transfer-Encoding: 7bit\r\nX-From:
slafontaine@globalp.com\r\nX-To: John.Arnold@enron.com\r\nX-cc: \r\nX-bcc: \r\nX-Folder:
\\John_Arnold_Dec2000\\Notes Folders\\Notes inbox\r\nX-Origin: Arnold-J\r\nX-FileName:
Jarnold.nsf\r\n\r\n\ni suck-hope youve made more money in natgas last 3 weeks than i have.
mkt shud\nbe getting bearish feb forward-cuz we already have the weather upon us-
fuel\nswitching and the rest shud invert the whole curve not just dec cash to jan
\nand\nfeb forward???? have a good weekend john\n",
  "date": { "$date": { "$numberLong": "976280700000" } }, "name": "re:summer inverses",
  "analysis": {
    "comprehend": {
      "entities":   [ (...) ],
      "keyPhrases": [ (...) ] }
  }
}
```

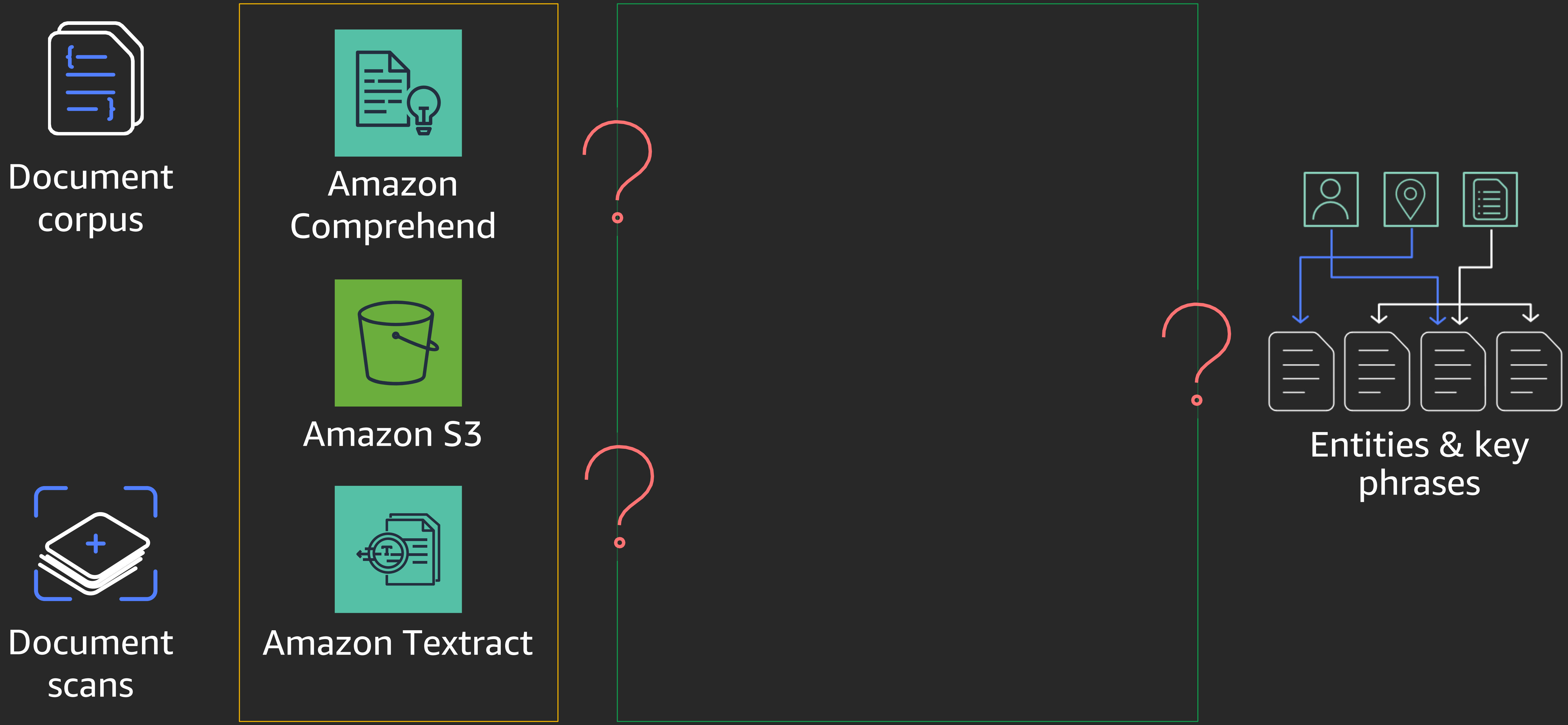
Comprehend metadata in Atlas

```
{ (...)
  "analysis": { "comprehend": {
    "entities": [
      { "beginOffset": { "$numberInt": "46" }, "endOffset": { "$numberInt": "58" },
        "score": { "$numberDouble": "0.8958866596221924" },
        "text": "last 3 weeks",
        "type": "DATE" },
      { "beginOffset": { "$numberInt": "229" }, "endOffset": { "$numberInt": "232" },
        "score": { "$numberDouble": "0.63105309009552" },
        "text": "jan",
        "type": "PERSON" },
      { "beginOffset": { "$numberInt": "274" }, "endOffset": { "$numberInt": "278" },
        "score": { "$numberDouble": "0.967898428440094" },
        "text": "john",
        "type": "PERSON" }
  ],
  "keyPhrases": [ (...) ]
} } }
```

Comprehend metadata in Atlas

```
{ (...)
 "analysis": { "comprehend": {
  "entities":   [ (...) ],
  "keyPhrases": [
   { "beginOffset": { "$numberInt": "9" }, "endOffset": { "$numberInt": "19" },
    "score": { "$numberDouble": "0.6691960692405701" },
    "text": "hope youve" },
   { "beginOffset": { "$numberInt": "25" }, "endOffset": { "$numberInt": "35" },
    "score": { "$numberDouble": "0.995799720287323" },
    "text": "more money" },
   { "beginOffset": { "$numberInt": "39" }, "endOffset": { "$numberInt": "45" },
    "score": { "$numberDouble": "0.8161398768424988" },
    "text": "natgas" },
   { "beginOffset": { "$numberInt": "92" }, "endOffset": { "$numberInt": "115" },
    "score": { "$numberDouble": "0.8442162871360779" },
    "text": "bearish feb forward-cuz" }, (...)
  ]
} } }
```

Stitching it together



Document corpus

Document scans

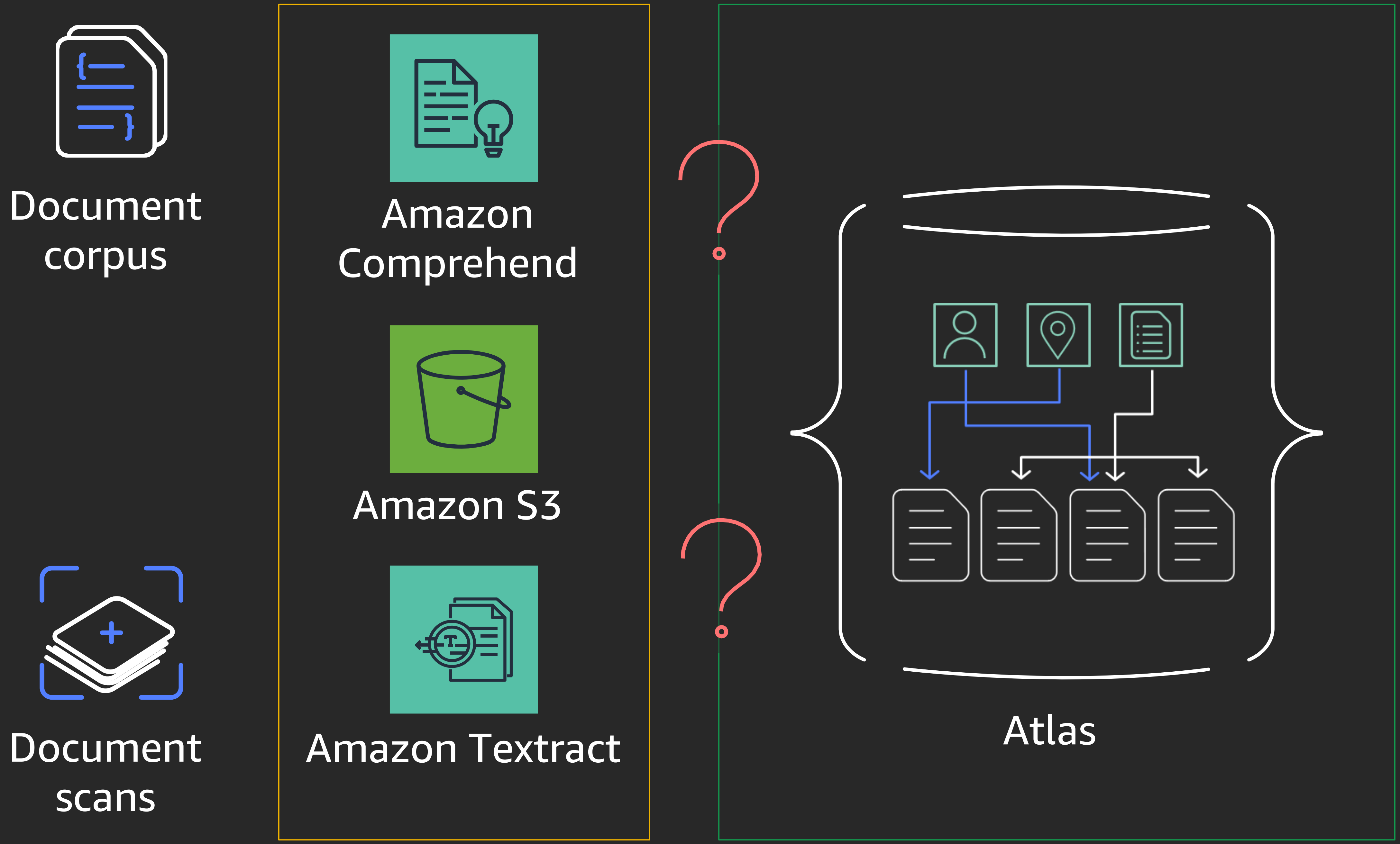
Amazon Comprehend

Amazon S3

Amazon Textract

Entities & key phrases

Stitching it together



Document corpus

Document scans

Amazon Comprehend

Amazon S3

Amazon Textract

Atlas

MongoDB Atlas – Global cloud database

Self-service & elastic

Global

Enterprise-grade security & SLAs

Comprehensive monitoring

Managed backup

Stitch: Serverless platform services

MongoDB Atlas – Global cloud database

Self-service & elastic

Global

Enterprise-grade security & SLAs

Comprehensive monitoring

Managed backup

Stitch: Serverless platform services

MongoDB Atlas – Global cloud database

Self-service & elastic

Global

Enterprise-grade security & SLAs

Comprehensive monitoring

Managed backup

Stitch: Serverless platform services

MongoDB Atlas – Global cloud database

Self-service & elastic

Global

Enterprise-grade security & SLAs

Comprehensive monitoring

Managed backup

Stitch: Serverless platform services

MongoDB Atlas – Global cloud database

Self-service & elastic

Global

Enterprise-grade security & SLAs

Comprehensive monitoring

Managed backup

Stitch: Serverless platform services

MongoDB Atlas – Global cloud database

Self-service & elastic

Global

Enterprise-grade security & SLAs

Comprehensive monitoring

Managed backup

Stitch: Serverless platform services

MongoDB Atlas – Global cloud database

Self-service & elastic

Global

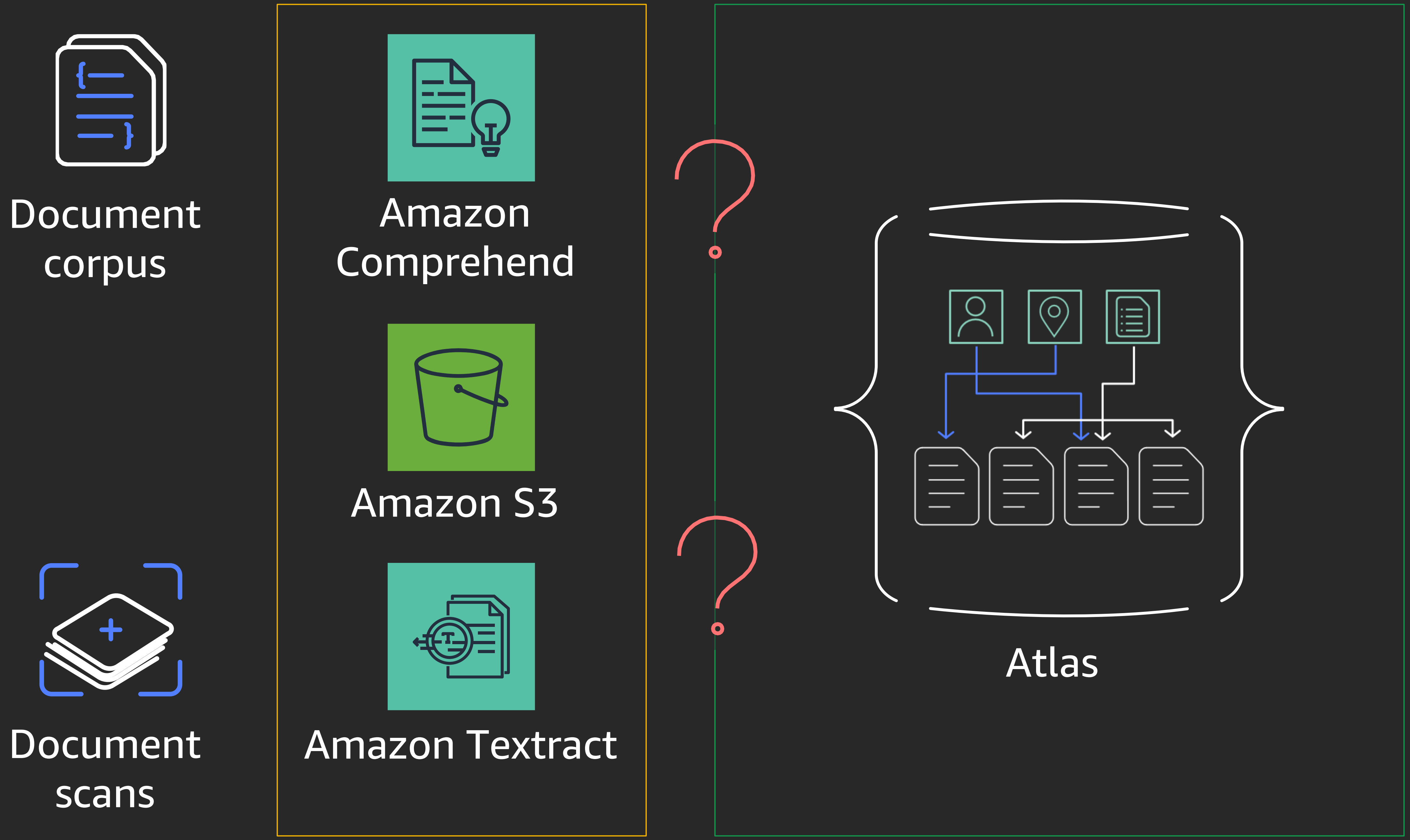
Enterprise-grade security & SLAs

Comprehensive monitoring

Managed backup

Stitch: Serverless platform services

Stitching it together



Document corpus

Document scans

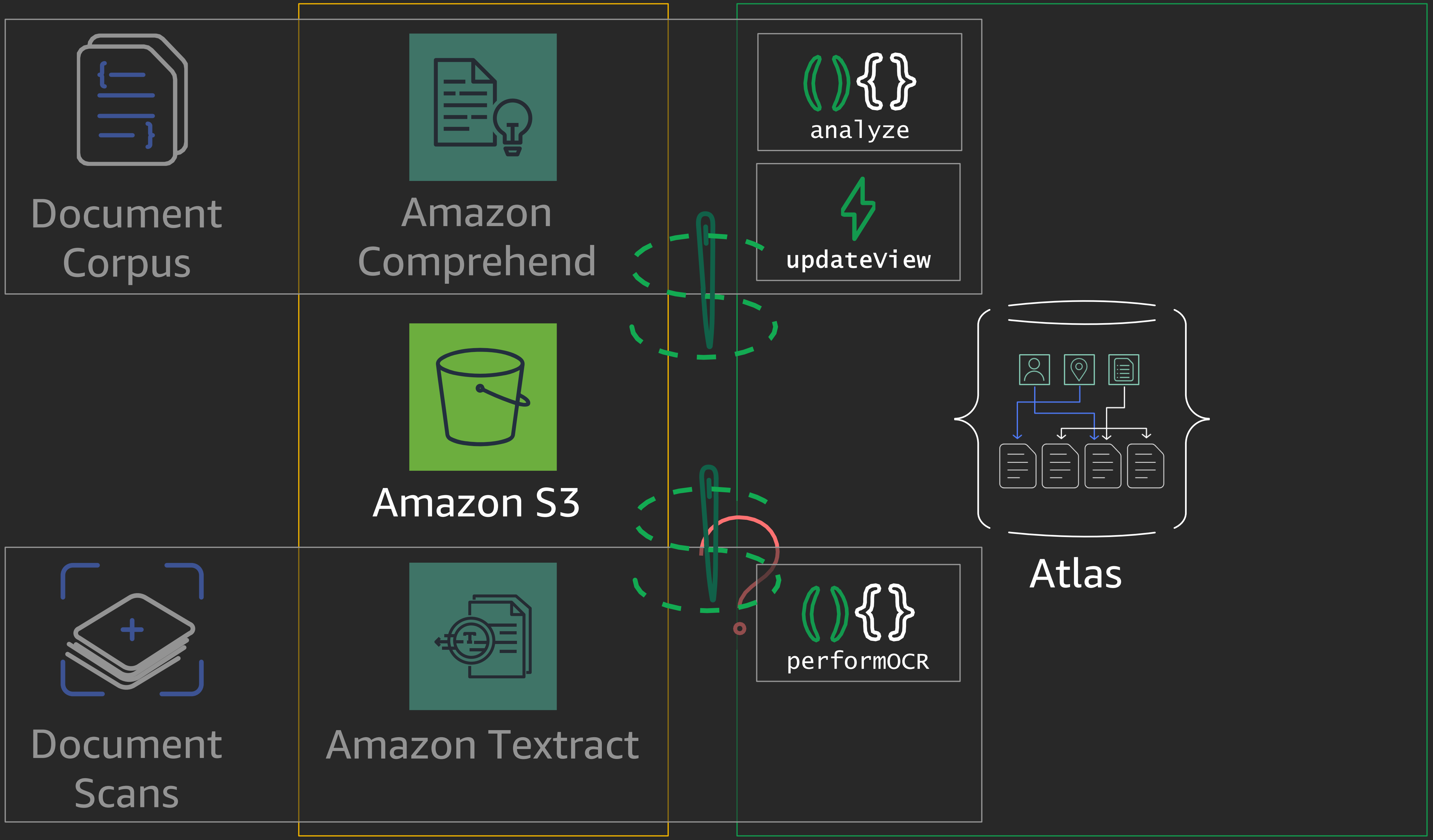
Amazon Comprehend

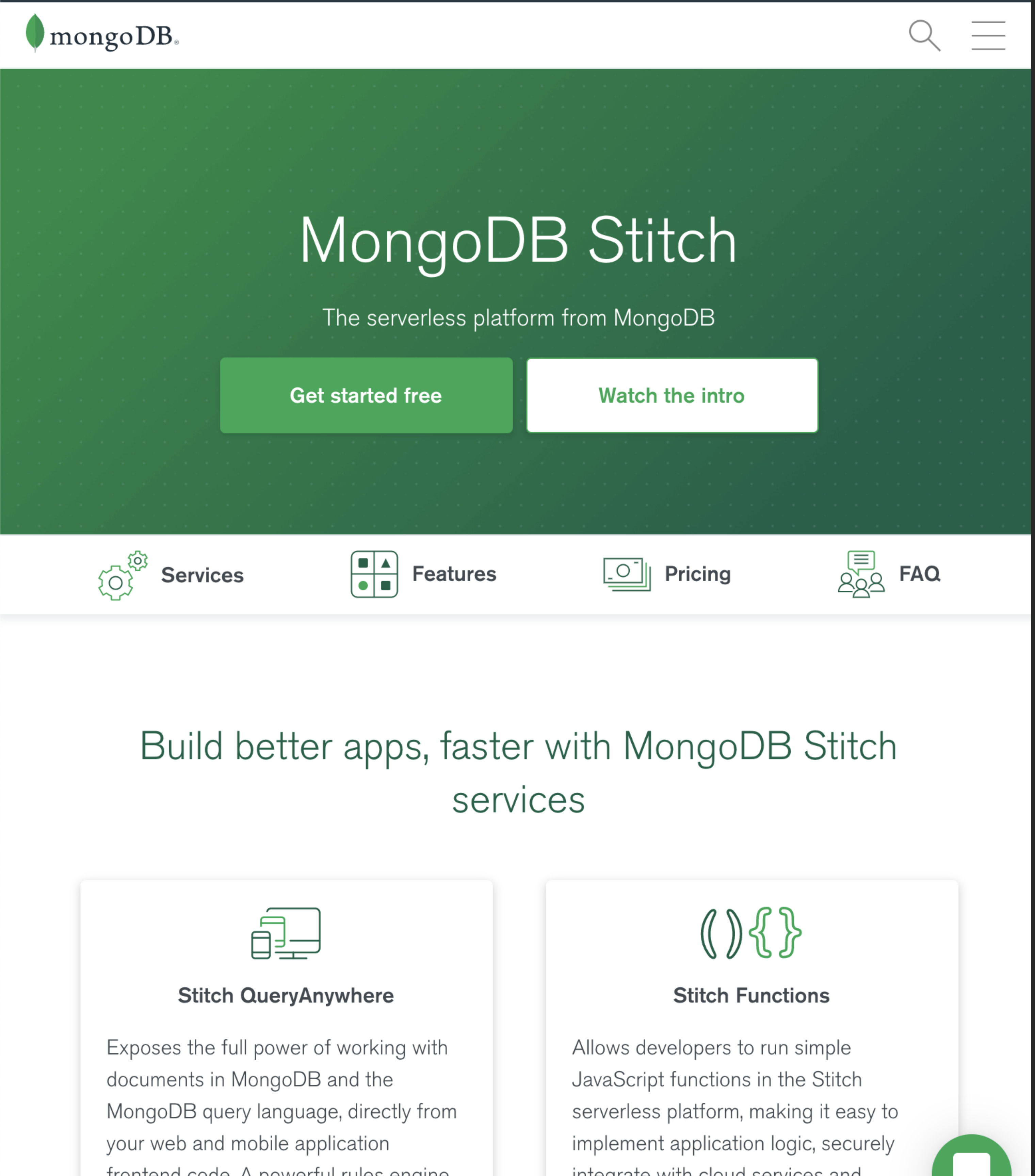
Amazon S3

Amazon Textract

Atlas

Stitching it together





MongoDB Stitch

The serverless platform from MongoDB

Build better apps, faster with MongoDB Stitch services

Stitch QueryAnywhere

Exposes the full power of working with documents in MongoDB and the MongoDB query language, directly from your web and mobile application frontend code. A powerful rules engine

Stitch Functions

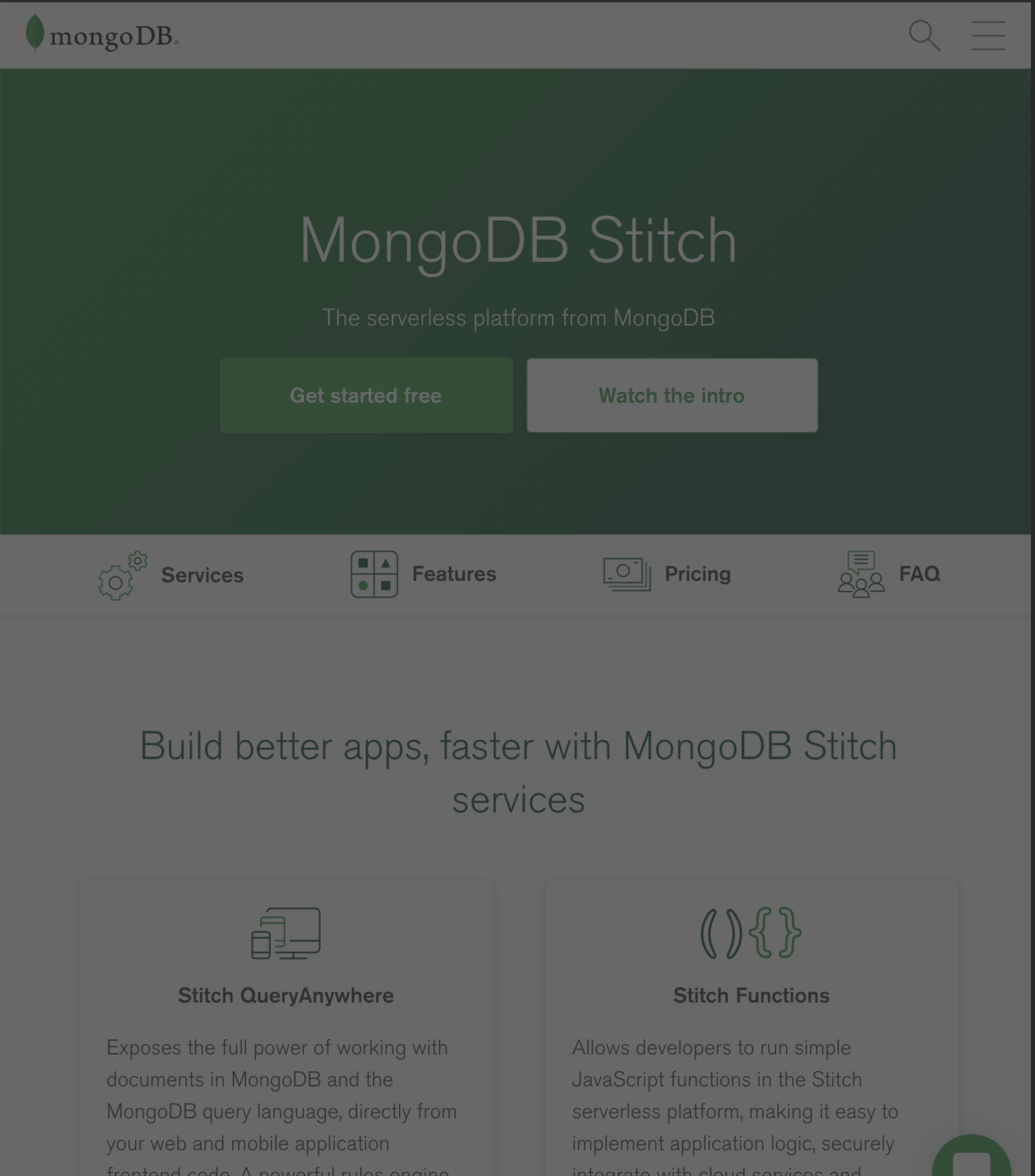
Allows developers to run simple JavaScript functions in the Stitch serverless platform, making it easy to implement application logic, securely integrate with cloud services and

Functions

Hosting

Triggers

Query Anywhere



MongoDB Stitch

The serverless platform from MongoDB

Build better apps, faster with MongoDB Stitch services

Stitch QueryAnywhere

Exposes the full power of working with documents in MongoDB and the MongoDB query language, directly from your web and mobile application frontend code. A powerful rules engine

Stitch Functions

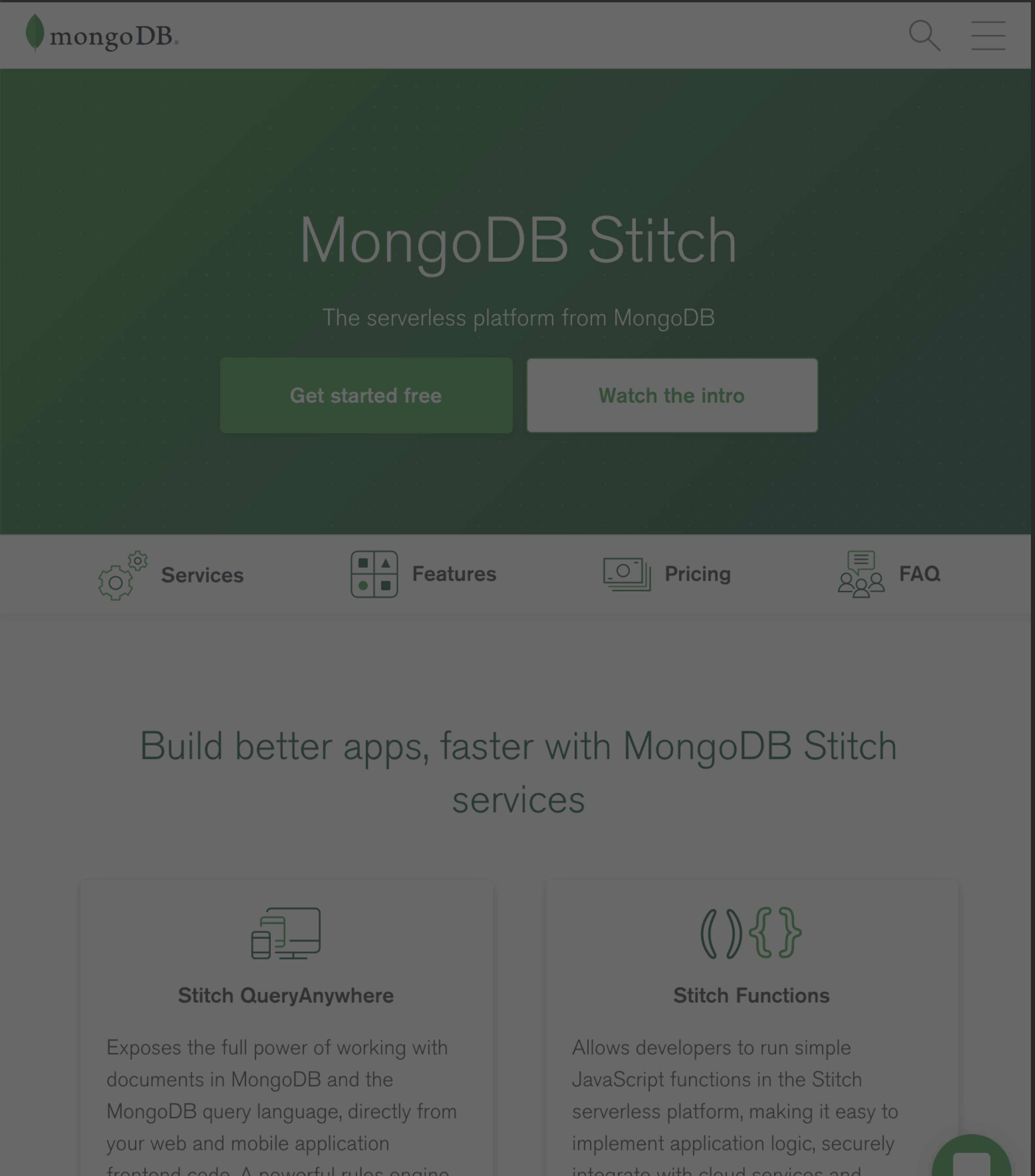
Allows developers to run simple JavaScript functions in the Stitch serverless platform, making it easy to implement application logic, securely integrate with cloud services and

Functions

Hosting

Triggers

Query anywhere



MongoDB Stitch

The serverless platform from MongoDB

Get started free

Watch the intro

Build better apps, faster with MongoDB Stitch services

Stitch QueryAnywhere

Exposes the full power of working with documents in MongoDB and the MongoDB query language, directly from your web and mobile application frontend code. A powerful rules engine

Stitch Functions

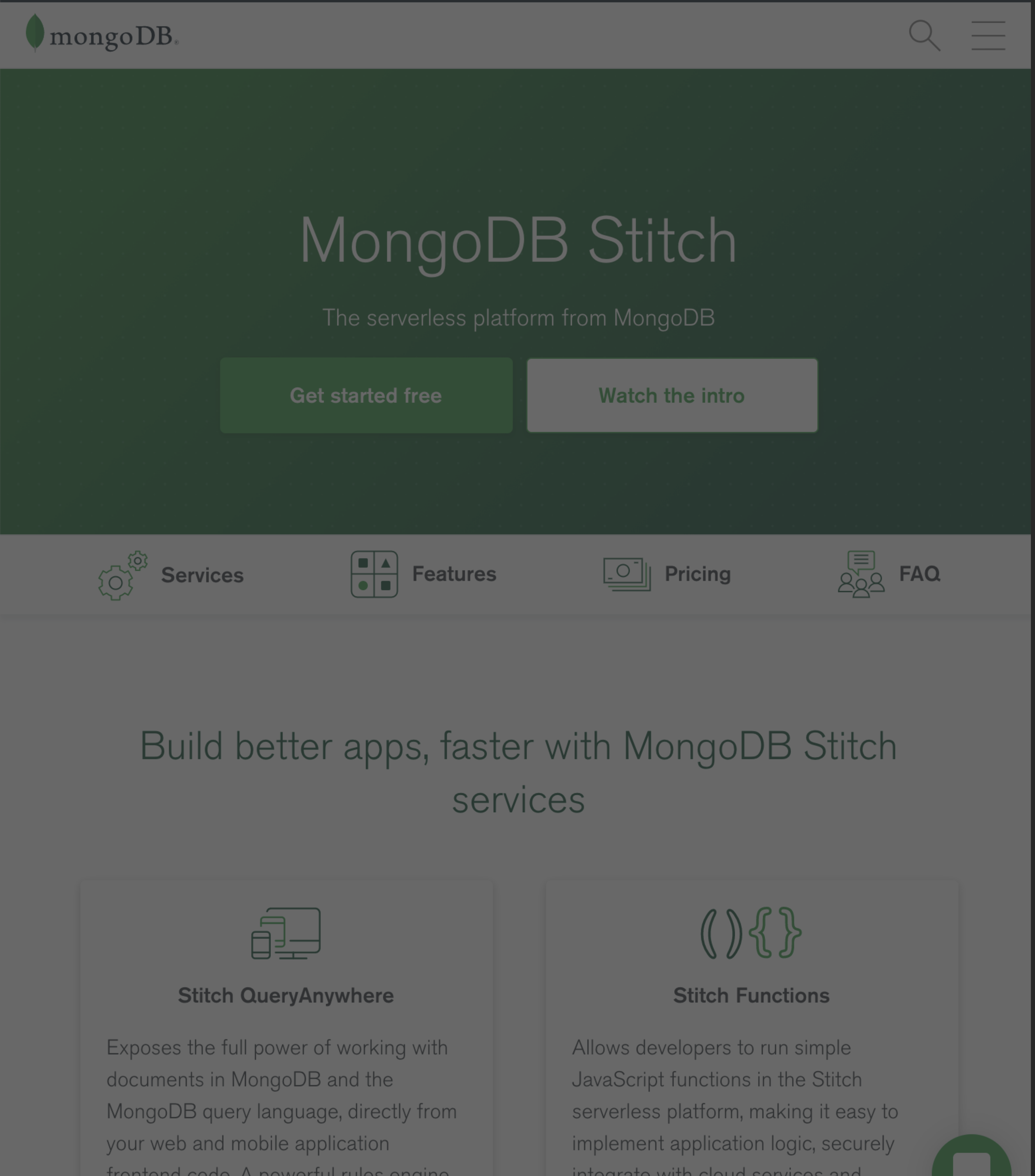
Allows developers to run simple JavaScript functions in the Stitch serverless platform, making it easy to implement application logic, securely integrate with cloud services and

Functions

Hosting

Triggers

Query anywhere



MongoDB Stitch

The serverless platform from MongoDB

Build better apps, faster with MongoDB Stitch services

Stitch QueryAnywhere

Exposes the full power of working with documents in MongoDB and the MongoDB query language, directly from your web and mobile application frontend code. A powerful rules engine

Stitch Functions

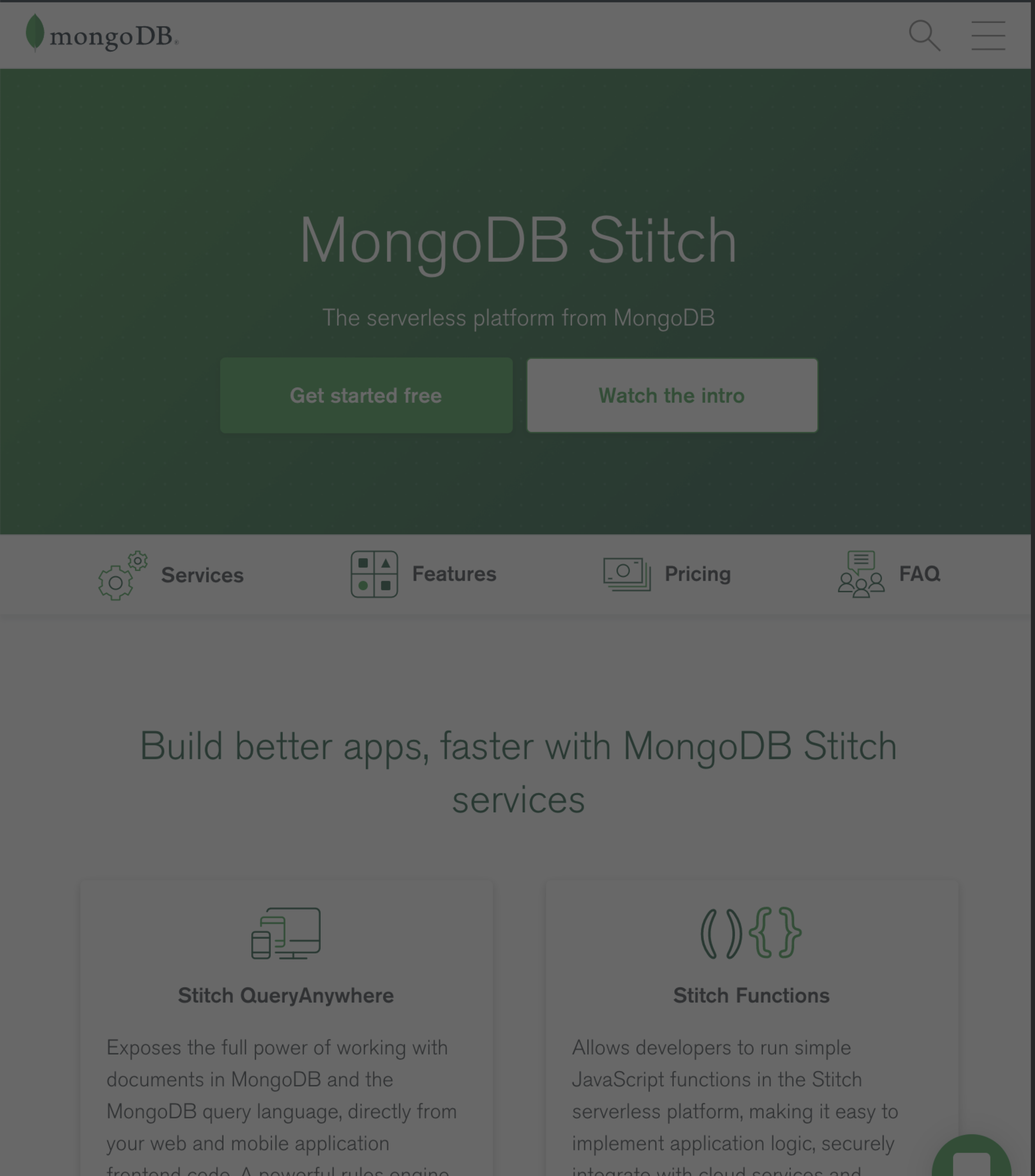
Allows developers to run simple JavaScript functions in the Stitch serverless platform, making it easy to implement application logic, securely integrate with cloud services and

Functions

Hosting

Triggers

Query anywhere



MongoDB Stitch

The serverless platform from MongoDB

Build better apps, faster with MongoDB Stitch services

Stitch QueryAnywhere

Exposes the full power of working with documents in MongoDB and the MongoDB query language, directly from your web and mobile application frontend code. A powerful rules engine

Stitch Functions

Allows developers to run simple JavaScript functions in the Stitch serverless platform, making it easy to implement application logic, securely integrate with cloud services and

Functions

Hosting

Triggers

Query anywhere

MongoDB Stitch

Functions

Hosting

Triggers

Query anywhere

AWS SDK for embedding into Stitch

CLI available

Git integrations

MongoDB Stitch

Functions

Hosting

Triggers

Query anywhere

AWS SDK for embedding into Stitch

CLI available

Git integrations

MongoDB Stitch

Functions

Hosting

Triggers

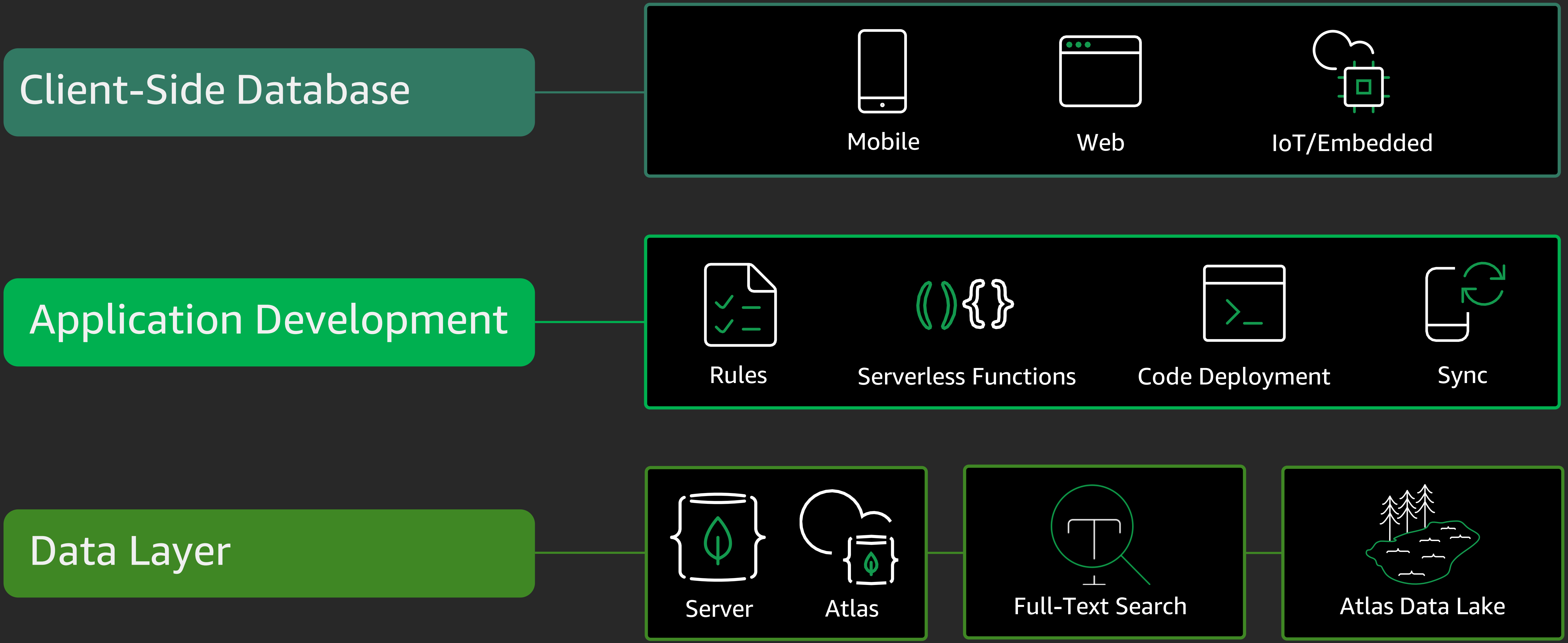
Query anywhere

AWS SDK for embedding into Stitch

CLI available

Git integrations

MongoDB is a data platform



Client-Side Database

Mobile

Web

IoT/Embedded

Application Development

Rules

Serverless Functions

Code Deployment

Sync

Data Layer

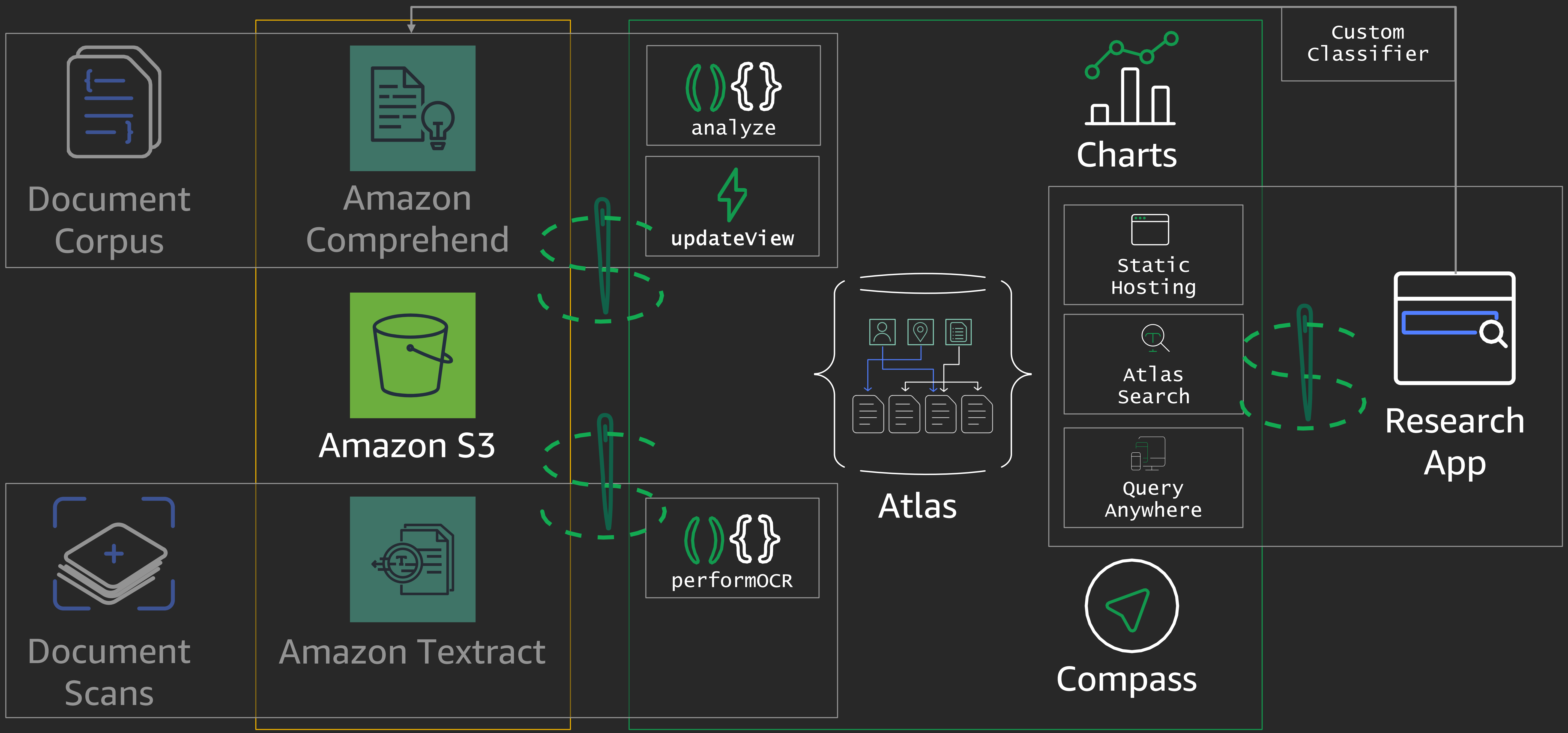
Server

Atlas

Full-Text Search

Atlas Data Lake

Stitching it together

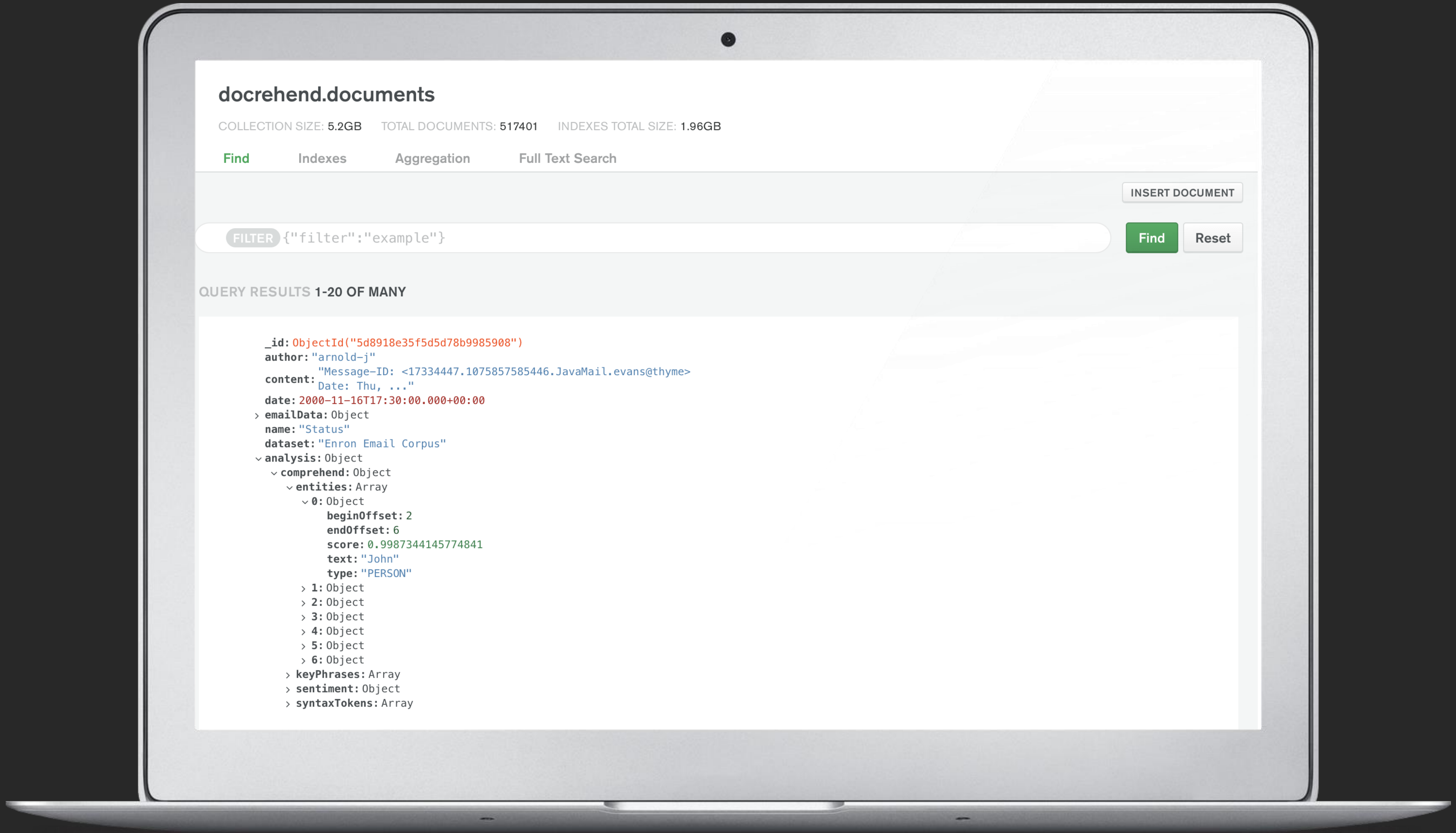


Data is power

AWS re:Invent

aws

Putting data into a searchable format is powerful



docrehend.documents

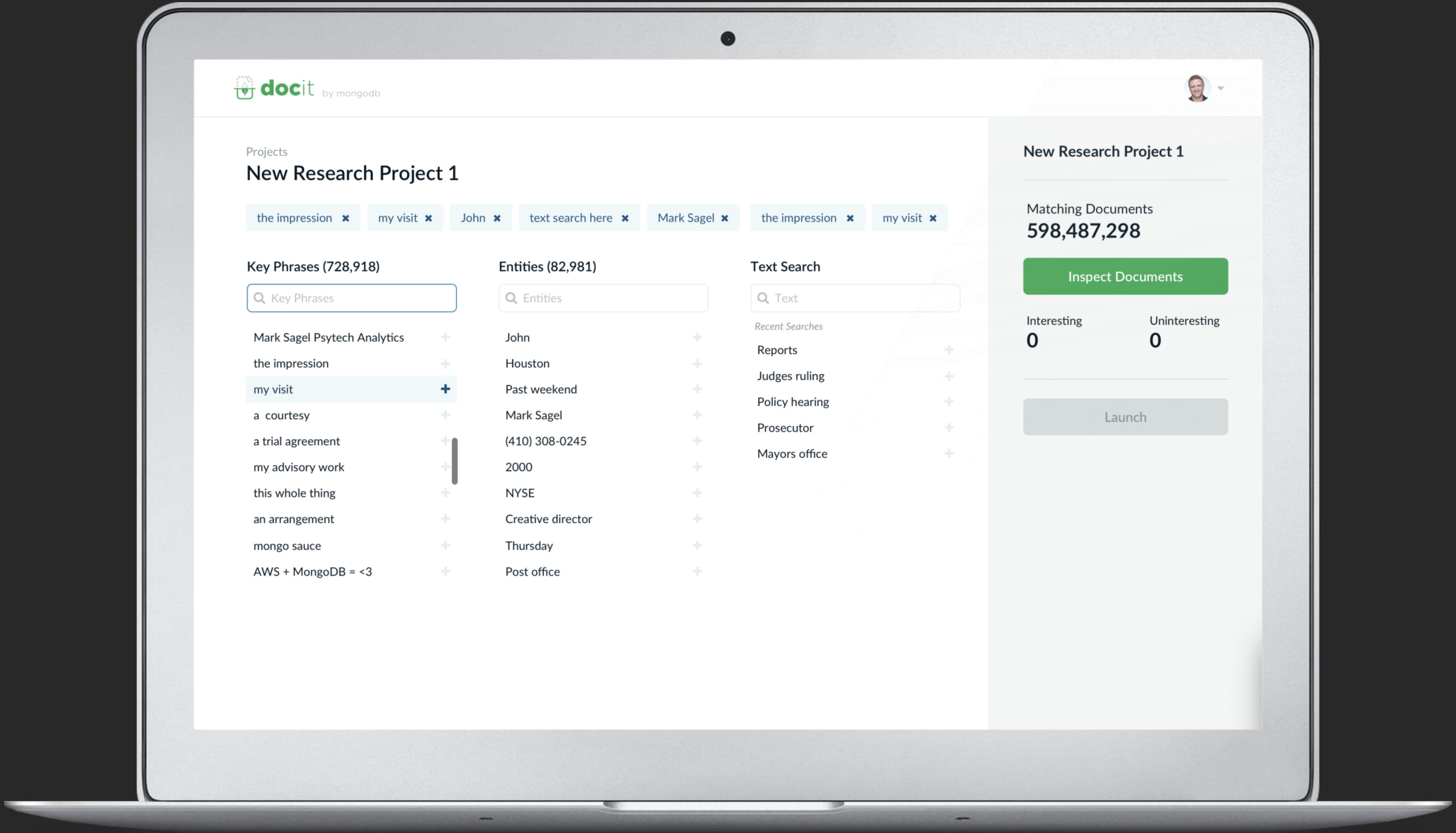
COLLECTION SIZE: 5.2GB TOTAL DOCUMENTS: 517401 INDEXES TOTAL SIZE: 1.96GB

FILTER {"filter":"example"}

QUERY RESULTS 1-20 OF MANY

```
  _id: ObjectId("5d8918e35f5d5d78b9985908")
  author: "arnold-j"
  content: "Message-ID: <17334447.1075857585446.JavaMail.evans@thyme>
           Date: Thu, ..."
  date: 2000-11-16T17:30:00.000+00:00
› emailData: Object
  name: "Status"
  dataset: "Enron Email Corpus"
⌄ analysis: Object
  ⌄ comprehend: Object
    ⌄ entities: Array
      ⌄ 0: Object
          beginOffset: 2
          endOffset: 6
          score: 0.9987344145774841
          text: "John"
          type: "PERSON"
      › 1: Object
      › 2: Object
      › 3: Object
      › 4: Object
      › 5: Object
      › 6: Object
    › keyPhrases: Array
    › sentiment: Object
    › syntaxTokens: Array
```

Reviewing metadata and searching that? Untouchable.



Projects

New Research Project 1

the impression ✖ | my visit ✖ | John ✖ | text search here ✖ | Mark Sagel ✖ | the impression ✖ | my visit ✖

Key Phrases (728,918)

Key Phrases

- Mark Sagel Psytech Analytics +
- the impression +
- my visit +
- a courtesy +
- a trial agreement +
- my advisory work +
- this whole thing +
- an arrangement +
- mongo sauce +
- AWS + MongoDB = <3 +

Entities (82,981)

Entities

- John +
- Houston +
- Past weekend +
- Mark Sagel +
- (410) 308-0245 +
- 2000 +
- NYSE +
- Creative director +
- Thursday +
- Post office +

Text Search

Text

Recent Searches

- Reports +
- Judges ruling +
- Policy hearing +
- Prosecutor +
- Mayors office +

New Research Project 1

Matching Documents
598,487,298

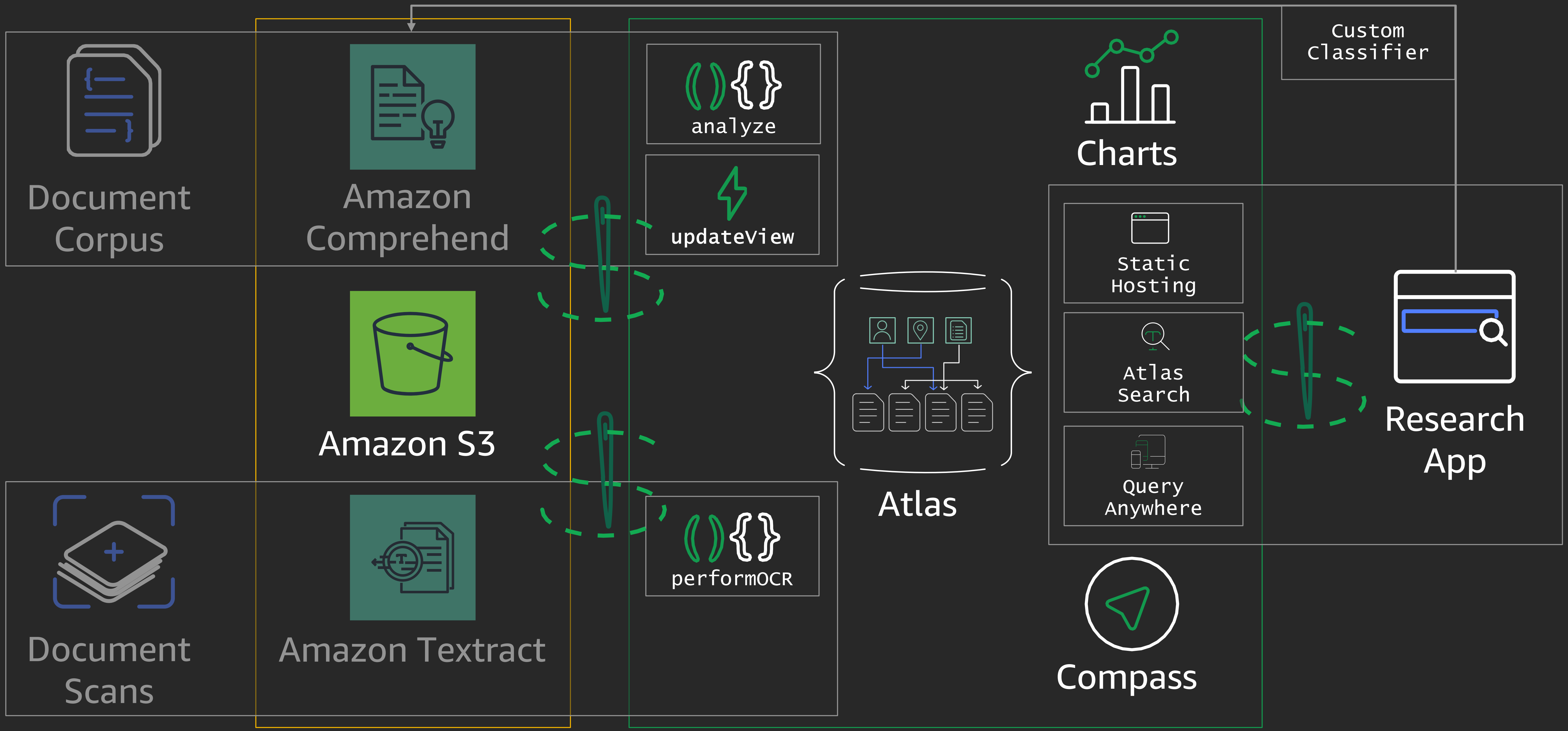
Inspect Documents

Interesting 0 | Uninteresting 0

Launch

Wrapping up

Application Architecture



Next steps

Mobile vs.
desktop

Charting
data

Analyzing
search queries

Adding
additional APIs

Email
integration

Benefits of this stack

Ease of storing and querying data

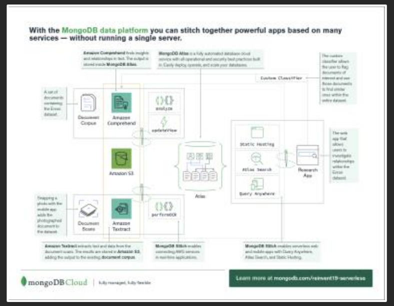
Powerful machine learning with adaptive tech to intelligently learn about data

Flexibility with state & calling out to other services

OCR taken to the next level, identifying contents in fields, forms, tables, and photos

MongoDB with Amazon Comprehend and Amazon Textract is incredibly powerful

Try it yourself



With the MongoDB data platform you can stitch together powerful apps based on many services — without running a single server.

Amazon Comprehend finds insights and relationships in text. The output is stored inside MongoDB Atlas.

MongoDB Atlas is a fully automated database cloud service with all operational and security best practices built in. Easily deploy, operate, and scale your databases.

The custom classifier allows the user to flag documents of interest and use those documents to find similar ones within the entire dataset.

A set of documents containing the Enron dataset.

The web app that allows users to investigate relationships within the Enron dataset.

Snapping a photo with the mobile app adds the photographed document to the dataset.

Amazon Textract extracts text and data from the document scans. The results are stored in Amazon S3, adding the output to the existing document corpus.

MongoDB Stitch enables connecting AWS services in real-time applications.

MongoDB Stitch enables serverless web and mobile apps with Query Anywhere, Atlas Search, and Static Hosting.

mongoDB Cloud | fully managed, fully flexible

Learn more at mongodb.com/reinvent19-serverless

Take your handout

Clone the repo

Claim your Atlas credits

mongodb.com/reinvent19-serverless

Thank you!

Diana Esteves

diana@mongodb.com

Ralph Capasso

ralph@mongodb.com

Please complete the session survey in the mobile app.

AWS re:Invent

aws