MICHIGAN STATE UNIVERSITY

# Project Plan

## Document Management at Google Scale

### The Capstone Experience

Team Technology Services Group

Ali Alaali
Joe Wan
Justin Newman
Luke Kline
Rohit Sen

Department of Computer Science and Engineering
Michigan State University

Fall 2019

*From Students…*
*…to Professionals*

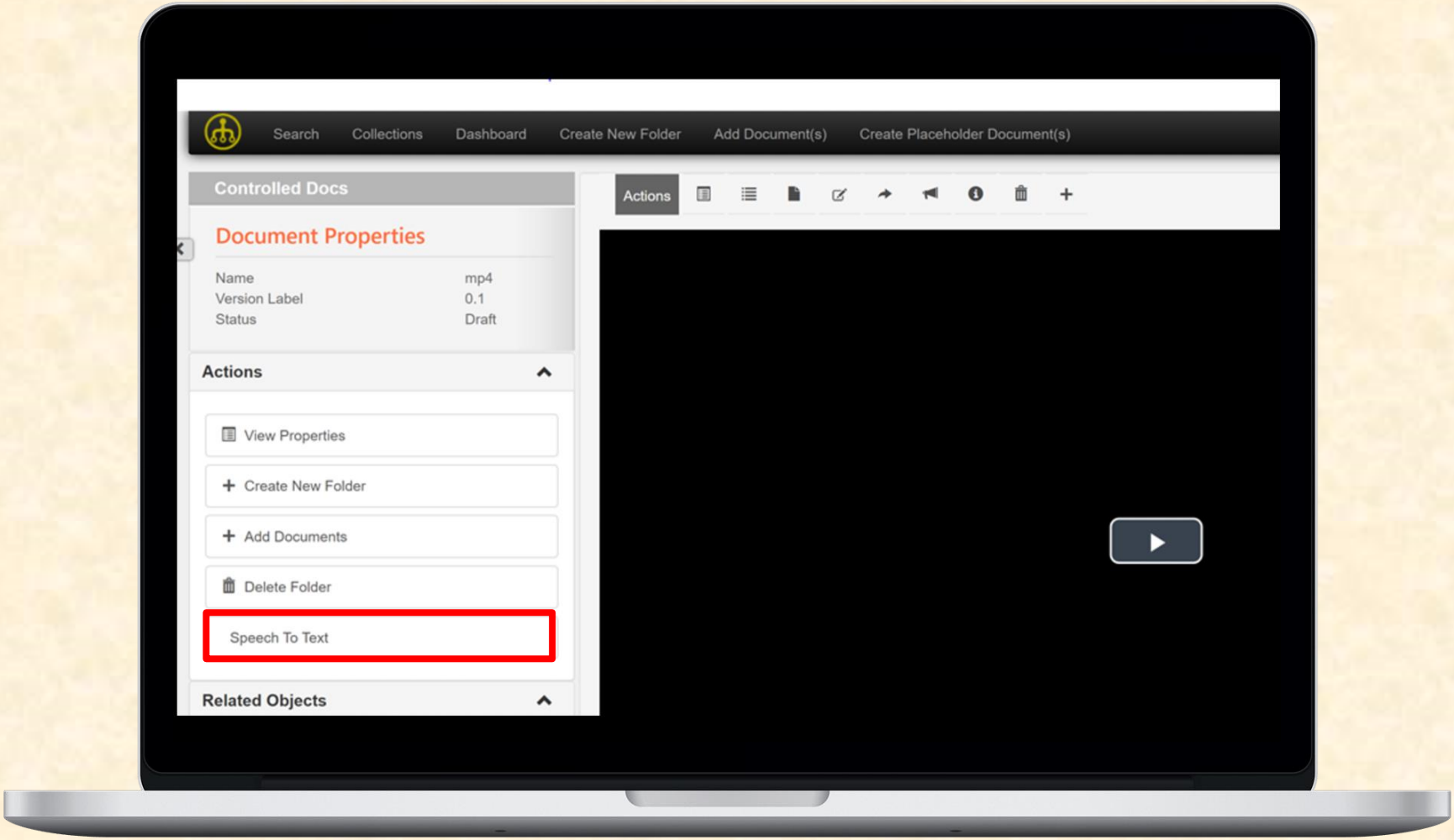
# Functional Specifications

- Due to the rapid growth of data, companies need a reliable solution for data management.
- TSG provides a solution to this problem:
  - OpenContent Management Suite (OCMS)
  - High speed search results
  - Scalable platform
- Our project goal:
  - Research how TSG can utilize Google Cloud Platform (GCP)
  - Surpass the AWS solution of 20,000 documents/s

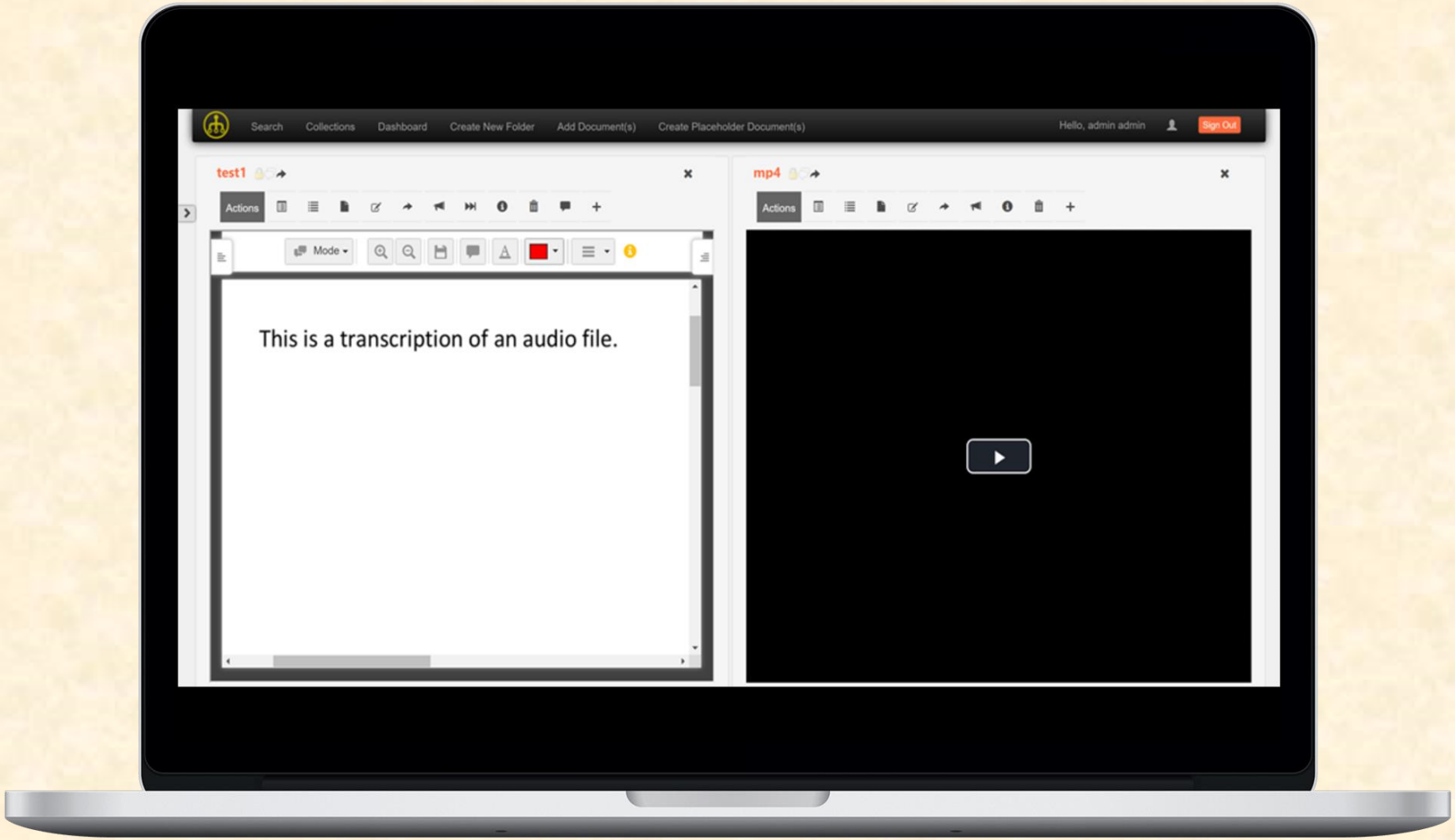
# Design Specifications

- Integrate the existing features of OCMS to be able to communicate with GCP
  - Document searching (OpenContent Search)
  - Document annotation (OpenAnnotate)
- Create a simple and viable UI for:
  - Speech API
  - Vision API

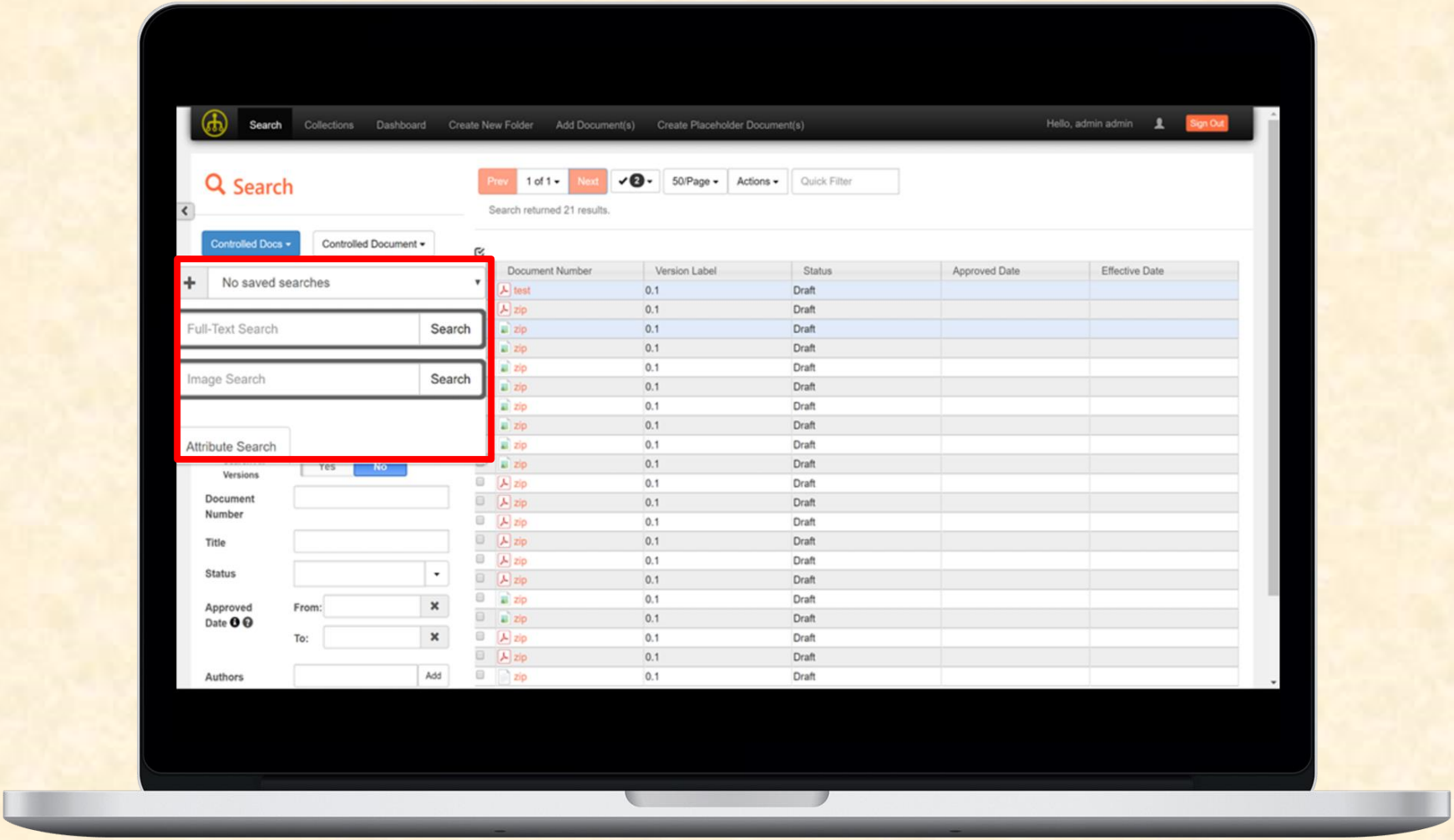
# Screen Mockup: Speech to Text Button



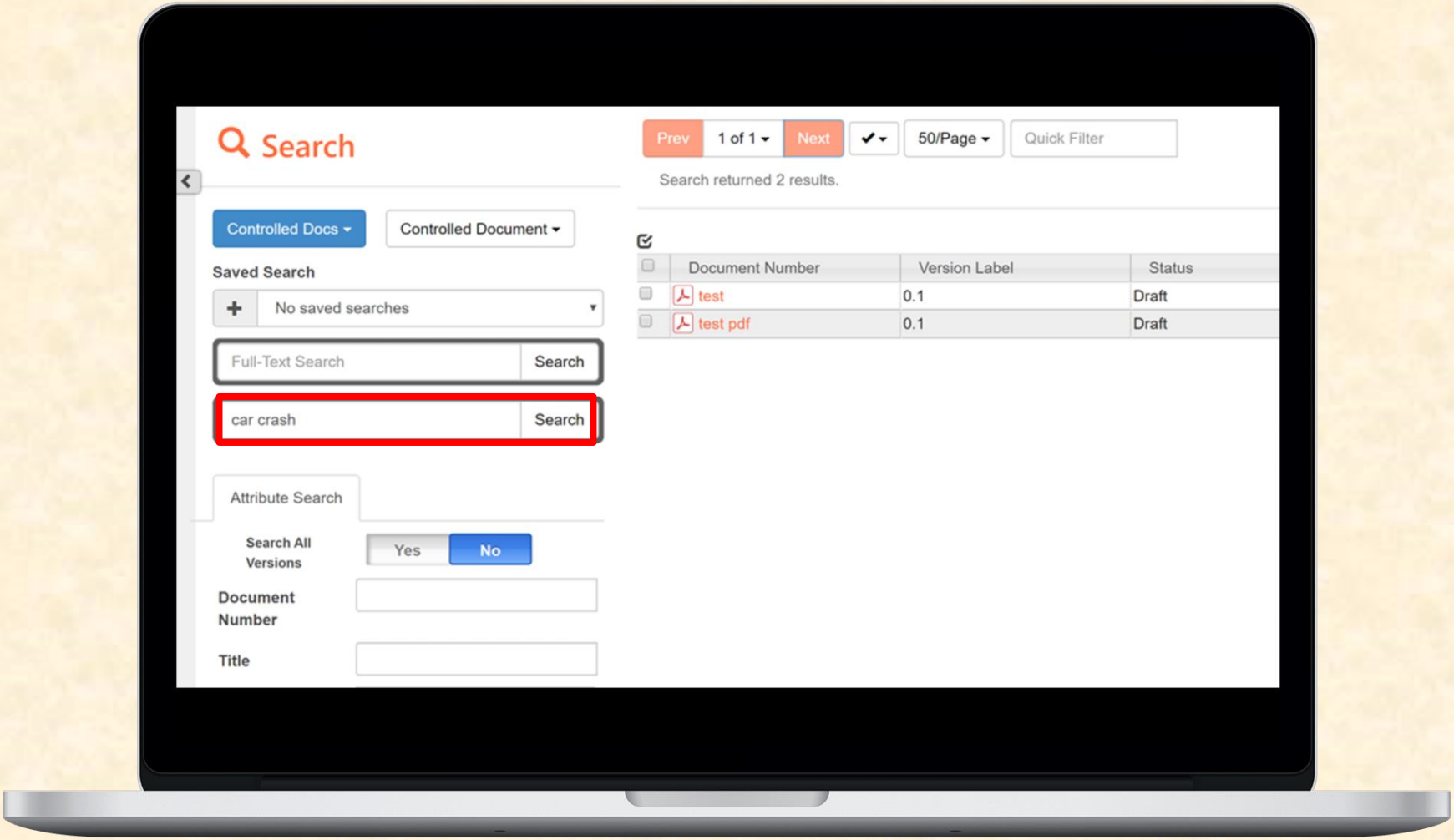
# Screen Mockup: Speech to Text UI



# Screen Mockup: Image Search Box



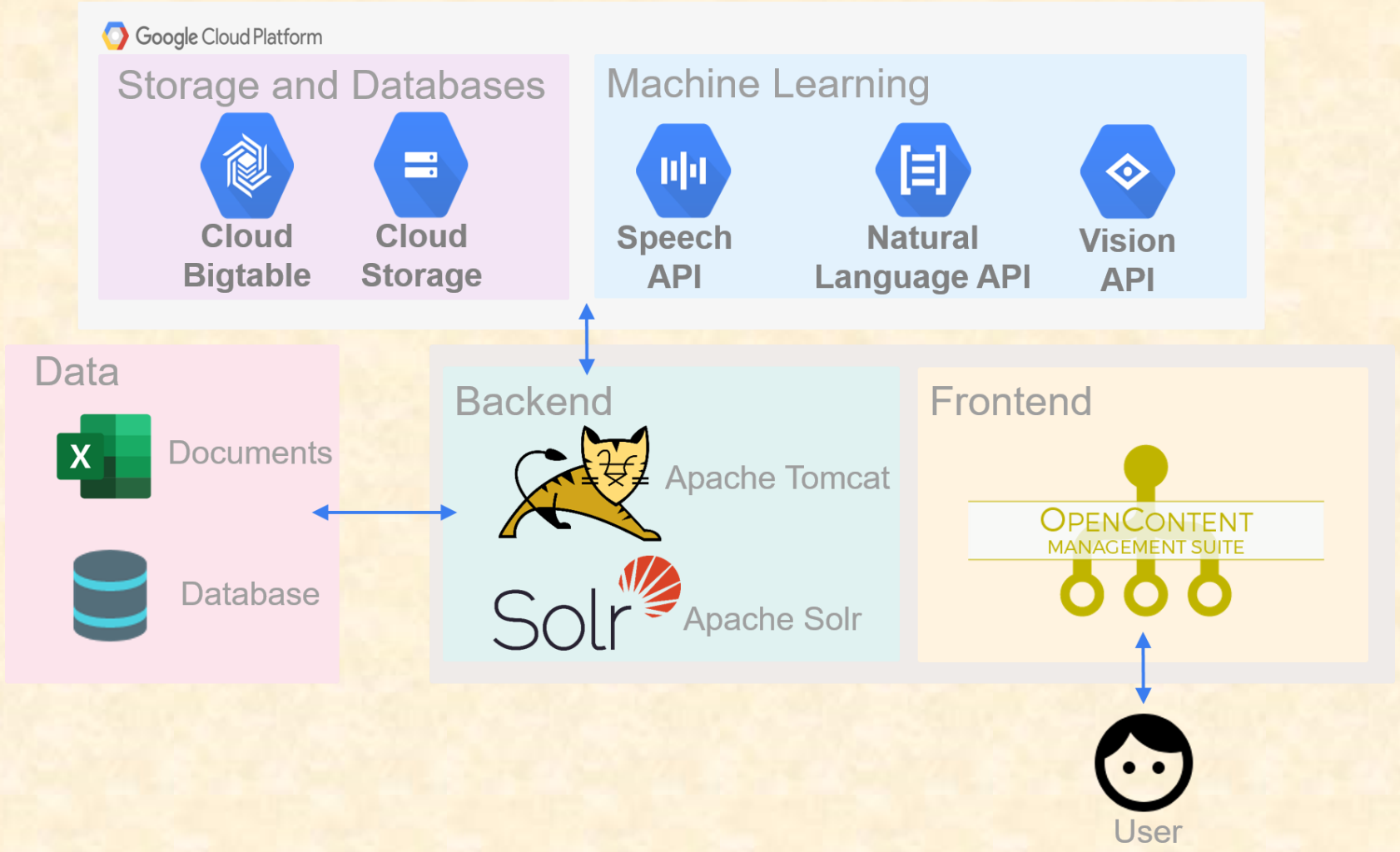
# Screen Mockup: Image Search Results



# Technical Specifications

- Storage Solutions
  - Google Cloud BigTable
    - NoSQL Database
  - Google Cloud Storage
    - Online file storage
- GCP's APIs for enhanced searching and functionality
  - Natural Language API
    - Classify documents
  - Vision API
    - Classify Images
  - Speech API
    - Transcribe Audio Files

# System Architecture



# System Components

- Frontend
  - JavaScript
  - jQuery
  - Bootstrap.js /CSS
  - HTML
- Backend
  - Java
  - Apache Tomcat
  - Apache Solr
  - Google Cloud Platform

# Risks

- **Scalability: Small sample size of testing**
  - Description: A small sample size of testing may result in inaccurate quality assurance

    Mitigation: Actively request access to a proper and larger dataset from the clients or create dummy data to be used for the benchmarking
- **Efficient Google BigTable schema**
  - Description: Optimized schema is essential to achieve high performance from GCP's BigTable
  - Mitigation: Continued research with Google's BigTable documentation and practice designing schemas and test them on our own instances
- **Processing Overhead for GCP's Vision AI**
  - Description: Vision AI processing overhead would decrease document ingestion rate to GCP
  - Mitigation: Processing documents using Vision AI at night or off-peak hours
- **Limited GCP resource**
  - Description: TSG offers a GCP instance for developing that runs during business hours
  - Mitigation: Setup our own GCP instance to be able to test without client's instance running

# Questions?