Webinar

RangeDocs

Searchable Science for Rangeland Management

August 21, 2024

Funding provided by a USDA Natural Resources Conservation Service Conservation Innovation Grant NR193A750008G003 & NR233A750008C005.
This work is supported by the Renewable Resources Extension Act Program [award no. 2021-46401-34740] from the USDA National Institute of Food and Agriculture.

Jason Karl, University of Idaho

Overview of RangeDocs

Matt King, University of Arizona

RangeDocs WebApp

Kristina Riemer & J.D. Gibbs, University of Arizona

Machine Learning & AI

Question & Answer Session

docs.rangelandsgateway.org

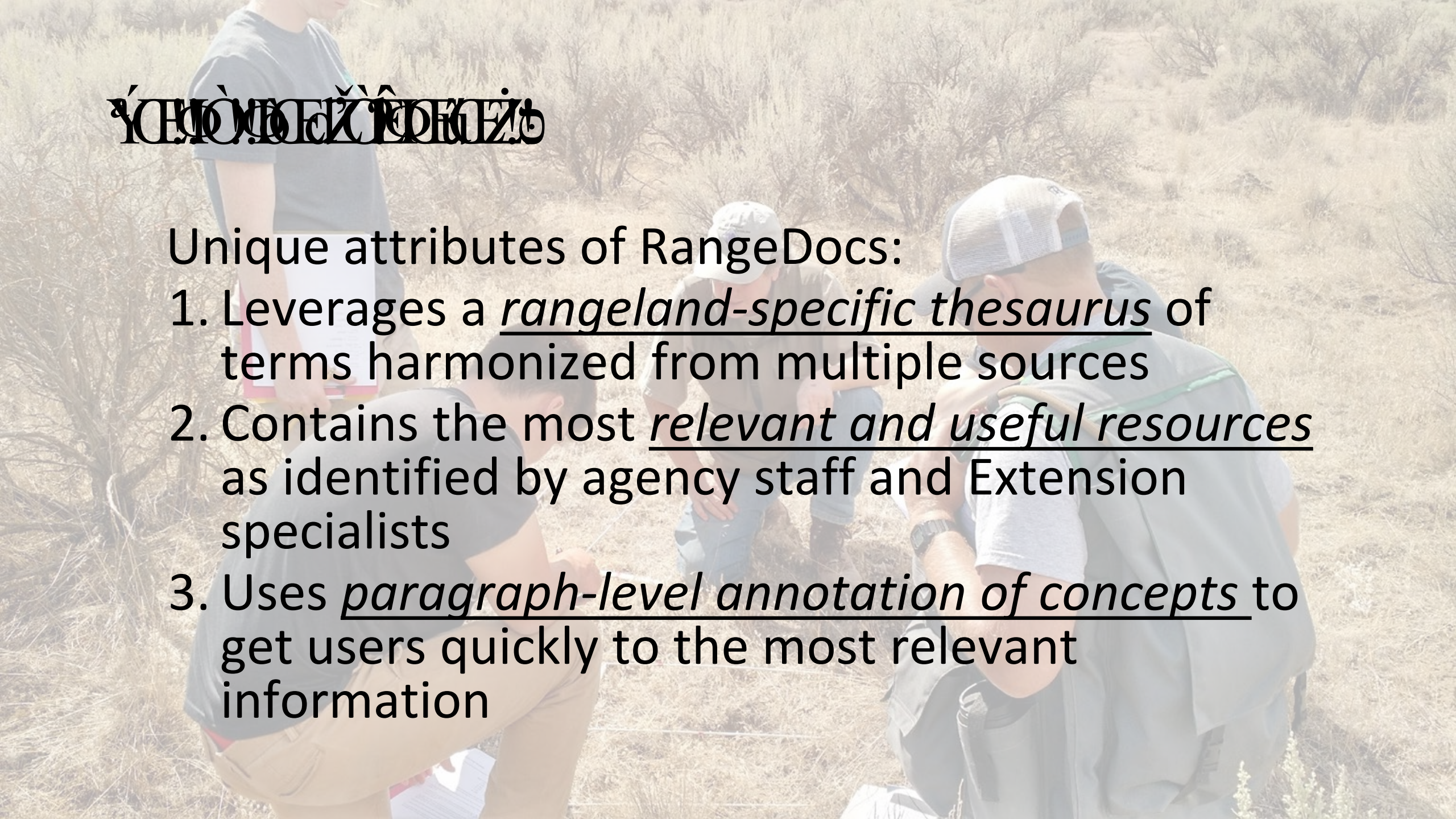
ÒCEd’ŽÒ?ÍÊÐ·úÊþÒdÊÒþÐdÍ¬ÐÒCEŻÊÝŒÂ¯Ò

THE CHALLENGE

Technical references and scientific papers are often long and dense. With limited time and energy to read literature, finding specific information can feel impossible.

THE SOLUTION

Range Docs is an online tool designed to search technical references and other literature at the page level. Rangeland experts have read the documents and labelled common rangeland terms where they appear in the documents. Science is now searchable.



Unique attributes of RangeDocs:

1. Leverages a *rangeland-specific thesaurus* of terms harmonized from multiple sources
2. Contains the most *relevant and useful resources* as identified by agency staff and Extension specialists
3. Uses *paragraph-level annotation of concepts* to get users quickly to the most relevant information

Core Project Components

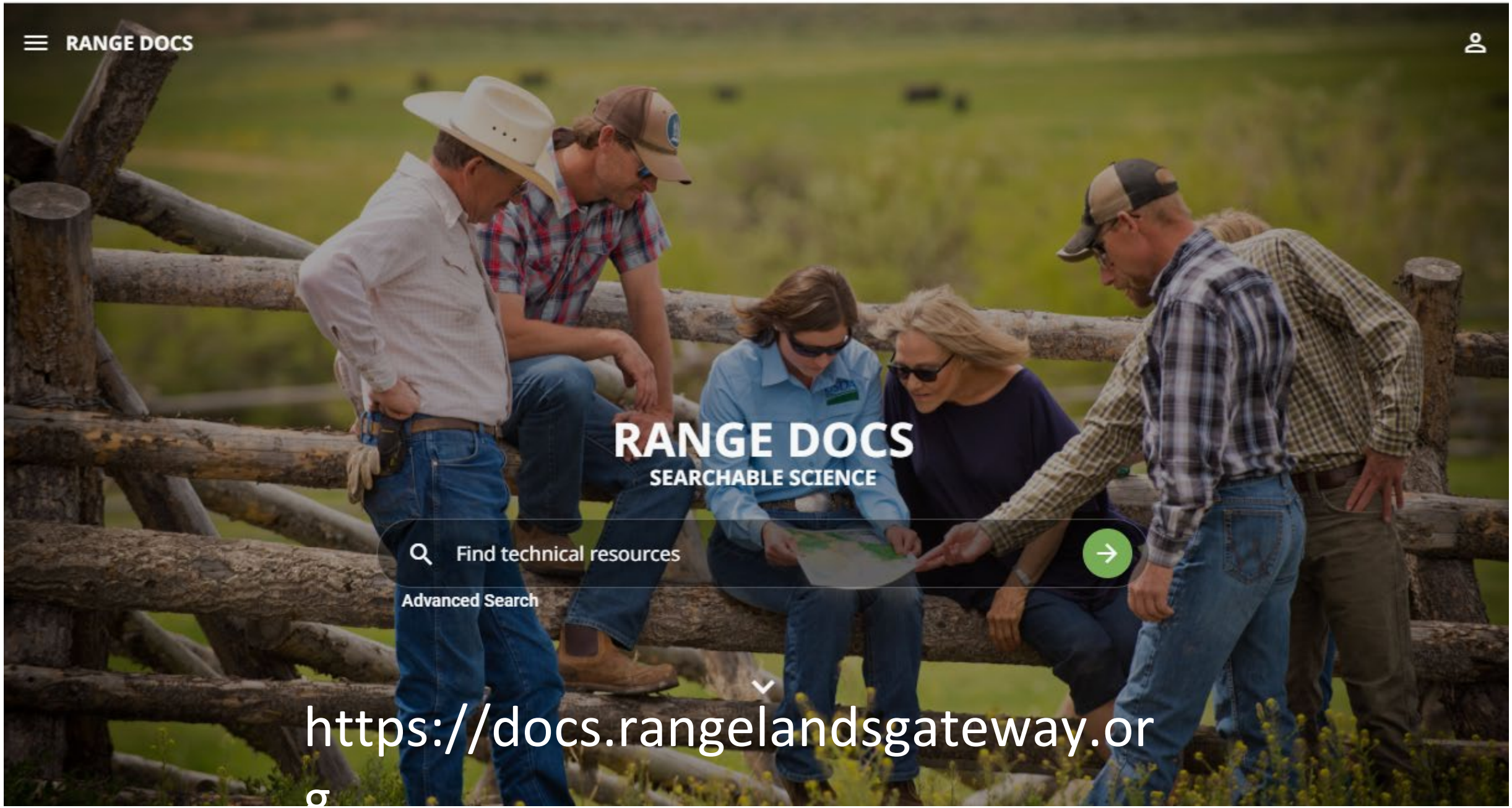
Grazingland Thesaurus	•Coordinate/update existing glossaries as a basis for tagging and finding grazingland reference materials
Content Selection	•Work with the Rangelands Partnership, NRCS and others to identify content
Content Indexing	•Tag documents and document sections with terms from the thesaurus for improved search results
Application Development	•Develop web and mobile interfaces for the system within the Global Rangelands platform
Outreach and Training	•Work with the Randelands Partnership and other partners to engage users in helping develop and implement the system

Core terms harmonized from multiple sources

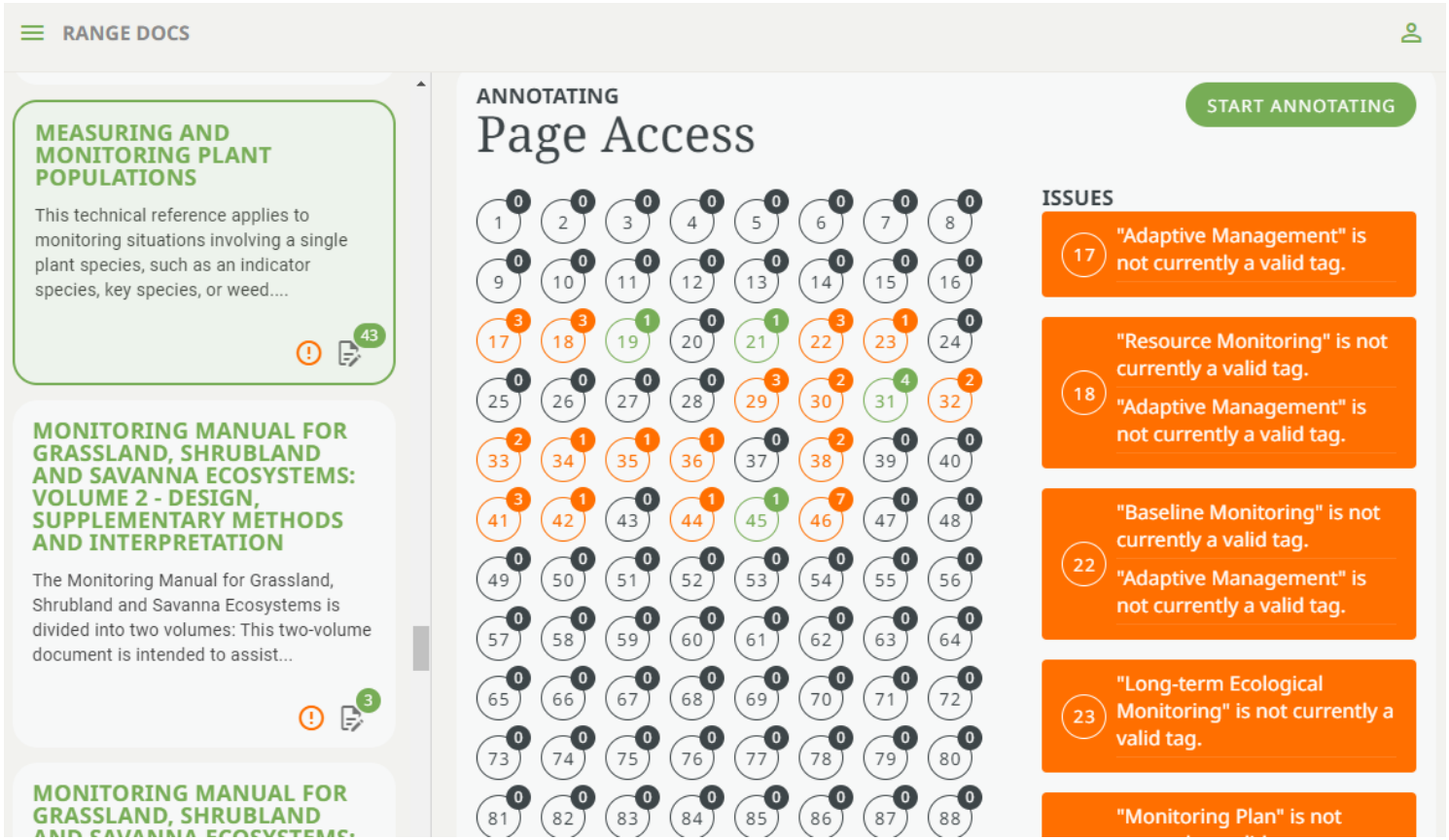
- Relationships of terms
- Standardized definitions
- Coordinated with AGROVOC and NALT

Maintained in VocBench database for easy updating

Launch point for potential revision to SRM Glossary (**1998**)

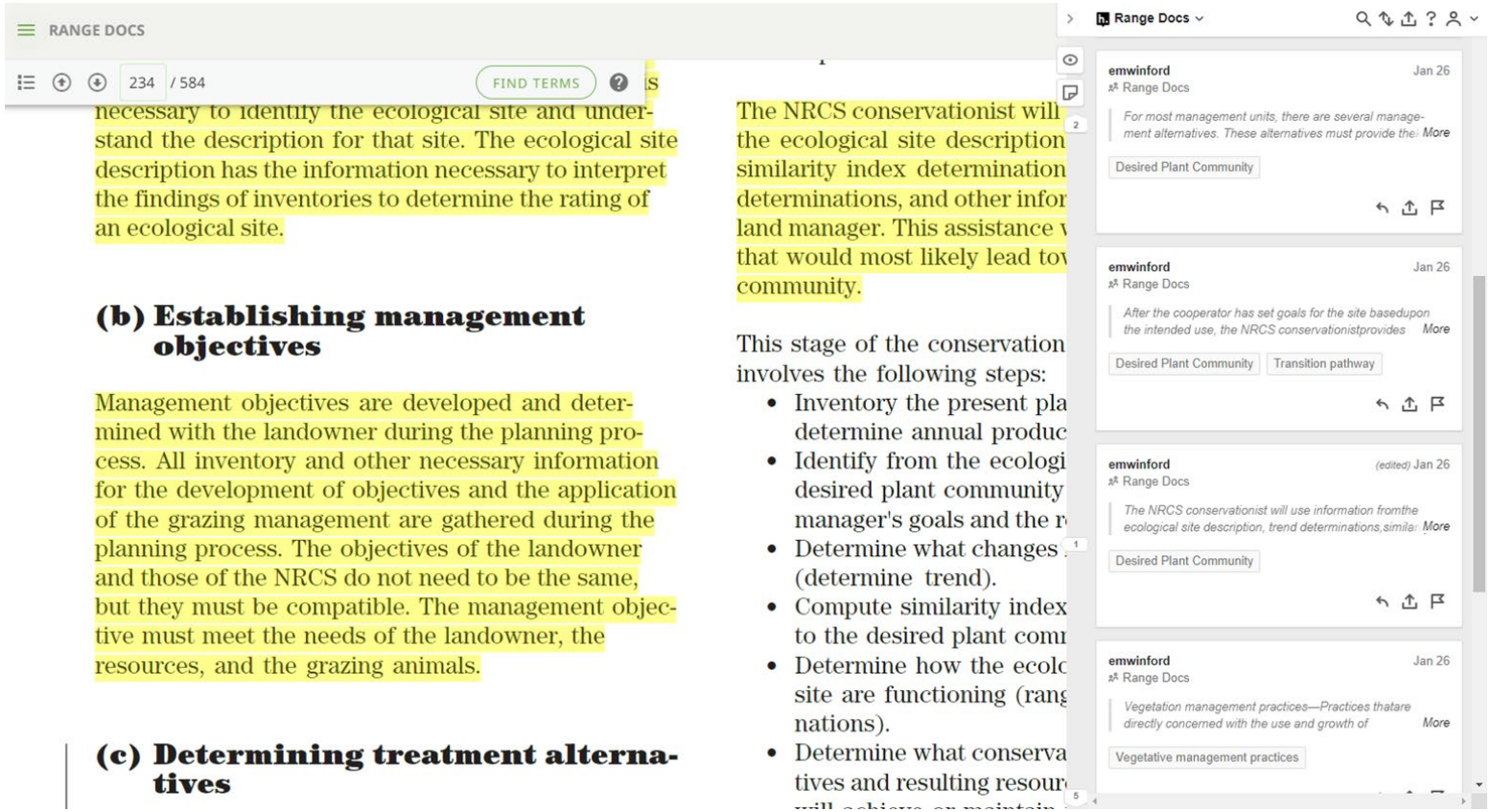

docs.rangelandsgateway.org
RANGE DOCS
RANGE DOCS
SEARCHABLE SCIENCE
Find technical resources
Advanced Search
https://docs.rangelandsgateway.org

òŒd’ŽÒ?ÍÊŒ?ÍÊ—ŽÒd!!
! d’d’Í !!Œ!Ú d’



- Documents are annotated by rangeland professionals using concepts from the thesaurus

òŒd’ŽÒ?íÊŒ?íÊ—ŽÒd!!
! d’d’Í !!Œ!Ú d’



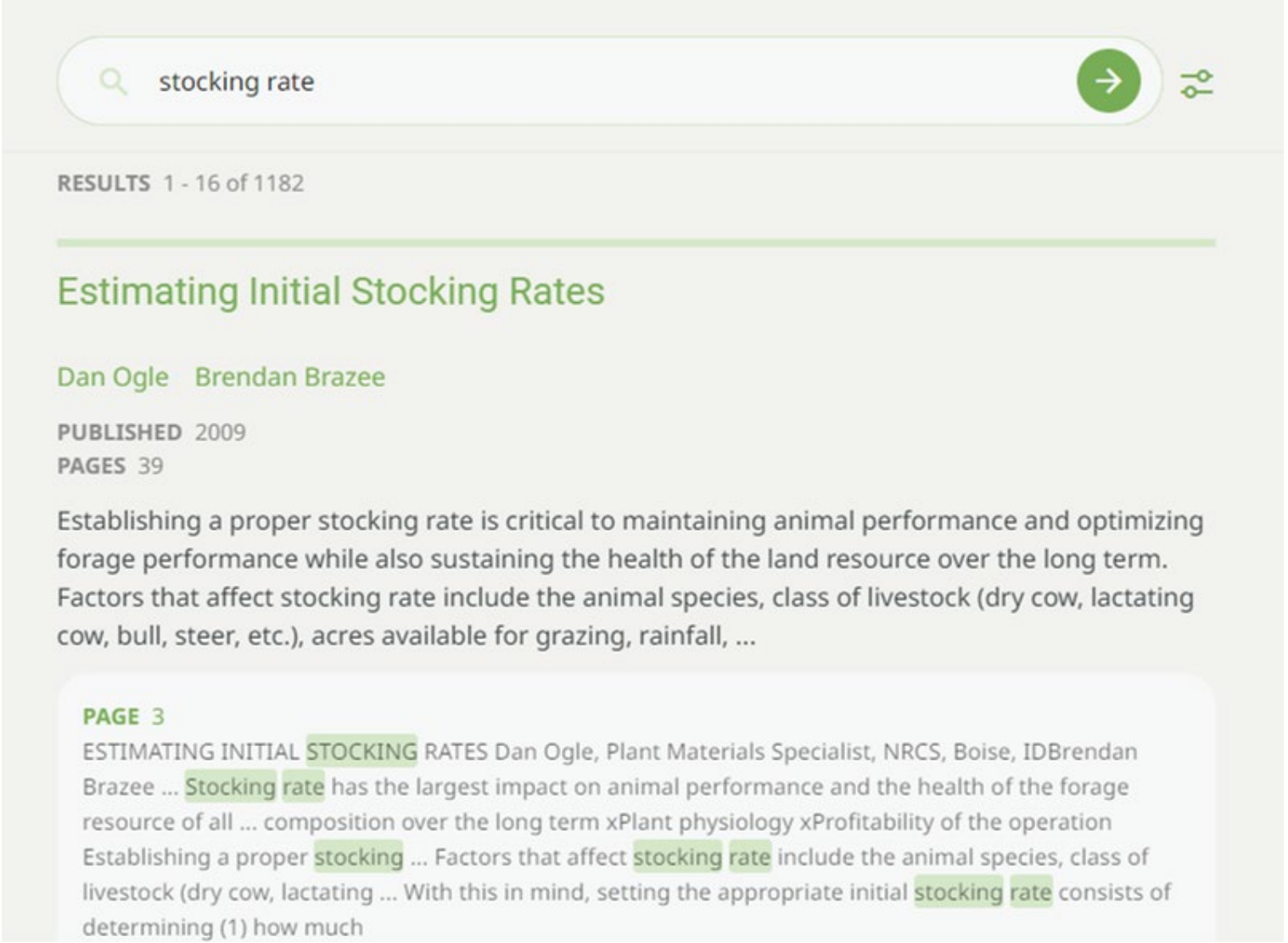
- Concepts are annotated at the page and paragraph level
- Allows search to direct user to specific parts of a document

ðŒEd’ŽÒ?ÍÊÞúÒŒŽÊÝðÒÞ→¯!ΦΦ

Search results direct users to the page and paragraph level

Annotated documents are prioritized (show up first)

Accessed documents are cached for offline access



ßüïŒ!ÒΦ!!í
ò Œl’ŽÒ? Í ÊΦ

- Local Extension documents
- Conservation Practice Standards & 2022 NRPH
- User Guide
- Support Videos

RangeDocs
Searchable Science for Rangeland Management

Project Team

Jason Karl
Amber Dalke
Sean DiStefano
Barb Hutchinson
Jeremy Kenyon
Matt King
Karen Launchbaugh
Sheila Merrigan
Jeanne Pfander
Matt Rahr
George Ruyle
Eric Winford
Retta Bruegger
Mark Thorne
J.D. Gibbs
Kristina Riemer

University of Idaho
The University of Arizona
University of Hawai'i Mānoa
Colorado State University Extension

Funding provided by a USDA Natural Resources Conservation Service Conservation Innovation Grant NR193A750008G003 & NR233A750008C005.

This work is supported by the Renewable Resources Extension Act Program [award no. 2021-46401-34740] from the USDA National Institute of Food and Agriculture.

Additional assistance provided by members of the Rangelands Partnership and the Altar Valley Conservation Alliance.

docs.rangelandsgateway.org

Jason Karl, University of Idaho

Overview of RangeDocs

Matt King, University of Arizona

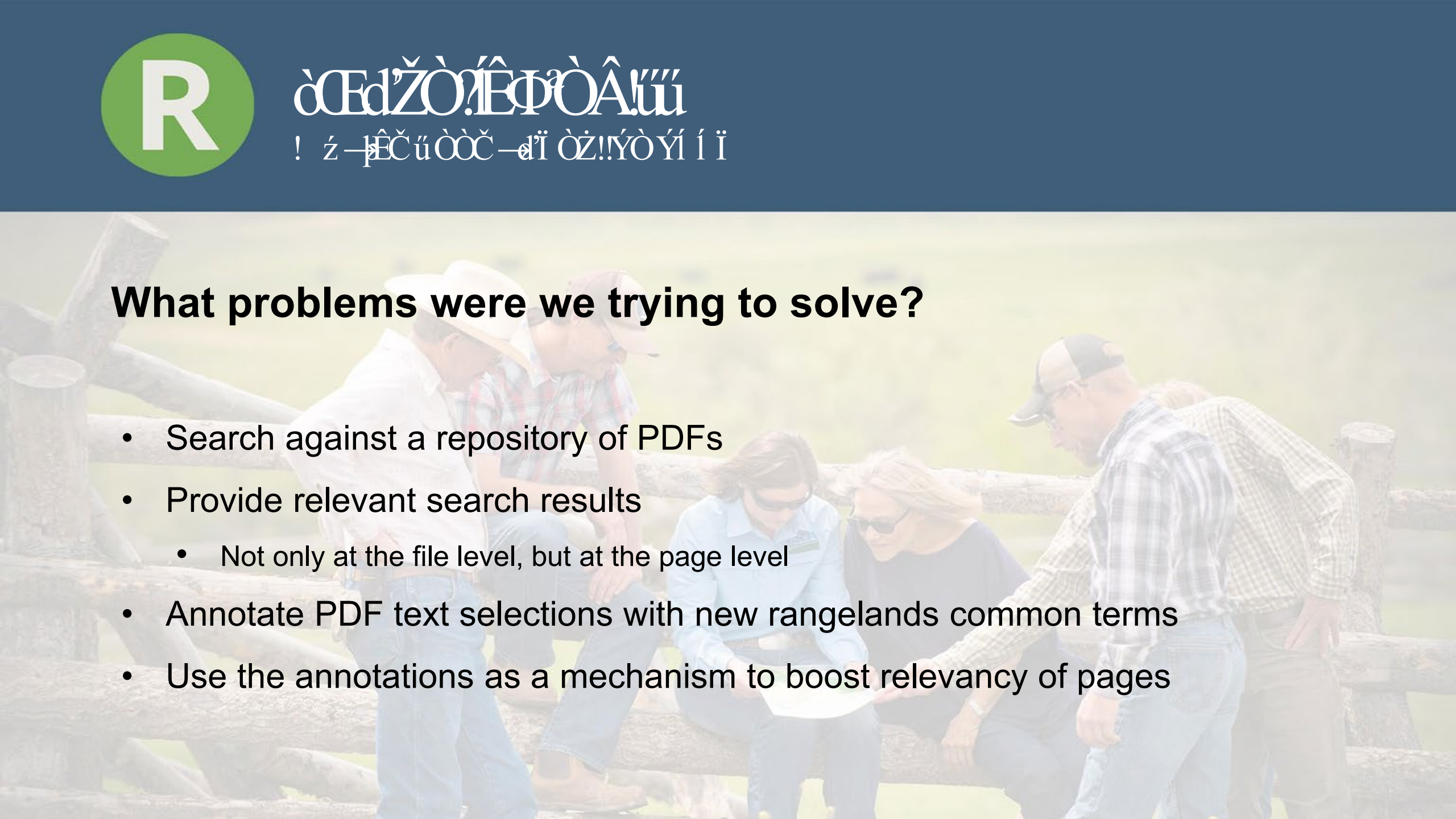
RangeDocs WebApp

Kristina Riemer & J.D. Gibbs, University of Arizona

Machine Learning & AI

Question & Answer Session

docs.rangelandsgateway.org



What problems were we trying to solve?

- Search against a repository of PDFs
- Provide relevant search results
  - Not only at the file level, but at the page level
- Annotate PDF text selections with new rangelands common terms
- Use the annotations as a mechanism to boost relevancy of pages

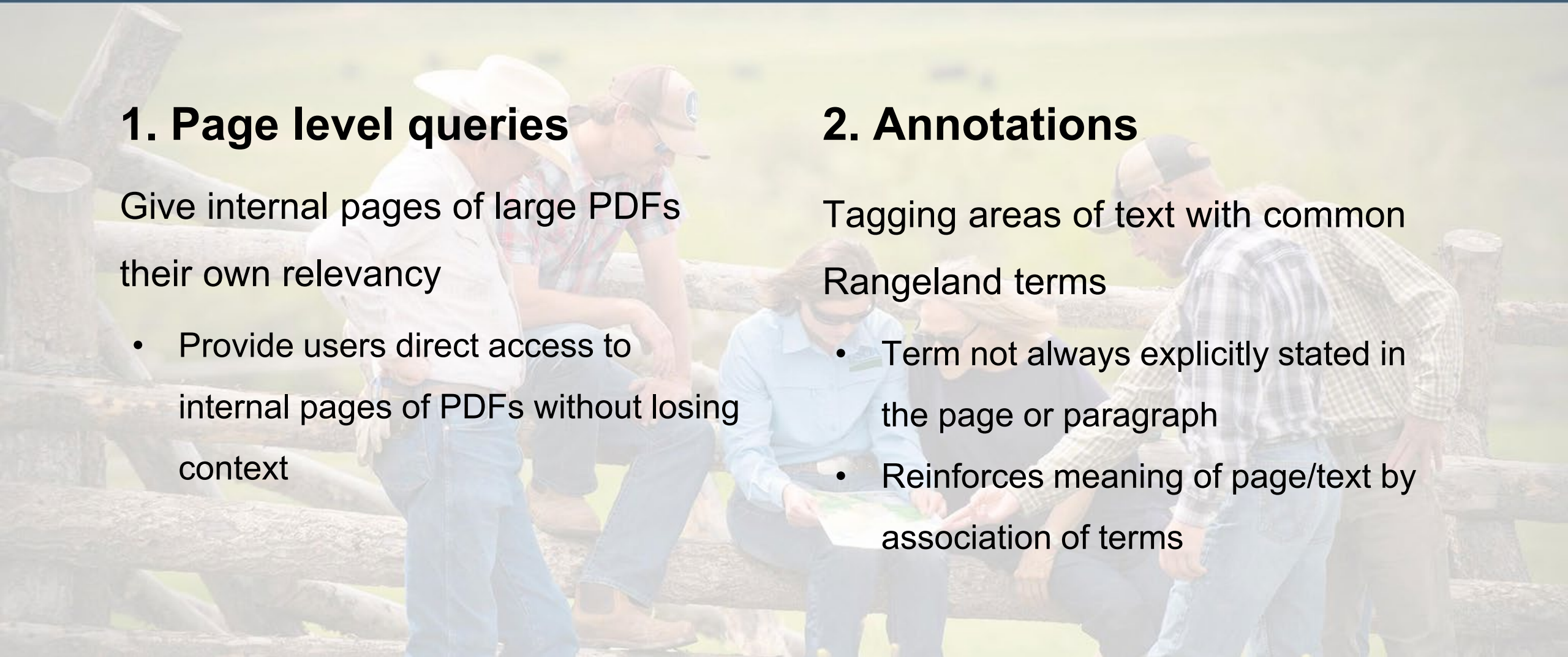
Off the shelf products and search engines only get us so far

We didn’t want to reinvent the wheel…

… but rather make a new wheel out of parts of other wheels.

RangeDocs WebApp

A quick peek under the hood



1. Page level queries

Give internal pages of large PDFs their own relevancy

- Provide users direct access to internal pages of PDFs without losing context

2. Annotations

Tagging areas of text with common Rangeland terms

- Term not always explicitly stated in the page or paragraph
- Reinforces meaning of page/text by association of terms

RangeDocs WebApp

A quick peek under the hood

Page level queries

PDF uploaded in RangeDocs web app

Text layer extracted from each page in PDF

Text is indexed in search database with metadata to retain page number and relationship to uploaded file, along with empty annotation properties.

- This gives each page its own relevancy score when being searched
- The score is boosted by overlapping relevancy with the original file's metadata (e.g. title, authors, abstract, tags, etc)

RangeDocs WebApp

A quick peek under the hood

Annotations: Powered by Hypothes.is

Online open source web annotation tool

- Third party code embedded in RangeDocs to annotate any text selection
  - Text selections can be uniquely identified by their placement on a webpage, PDF fingerprints, and web page URL
- Hypothes.is provides private groups that allow only elevated users in RangeDocs to submit valid annotations

RangeDocs WebApp

A quick peek under the hood

Annotations: Hypothes.is (cont'd)

Annotator selects paragraph in PDF in RangeDocs

Text selection tagged with term from glossary

Annotations are saved in Hypothesi.is' system. RangeDocs uses Hypothes.is API to harvest / synch annotations to the page annotation properties.

- Additional boost algorithms are given to pages where search has overlapping relevancy with page annotations

RangeDocs WebApp

A quick peek under the hood

Collections: new virtual documents

- Leveraging the output of Page Level Queries, PDFs uploaded into RangesDocs are stored as individual documents
- Collections allow any user to stitch together different pages from across all resources in RangeDocs
    - Colate like topics into one place
- Private vs Curated
    - View virtually in the browser or download as a new PDFs

RangeDocs
Searchable Science for Rangeland Management

Project Team

Jason Karl
Amber Dalke
Sean DiStefano
Barb Hutchinson
Jeremy Kenyon
Matt King
Karen Launchbaugh
Sheila Merrigan
Jeanne Pfander
Matt Rahr
George Ruyle
Eric Winford
Retta Bruegger
Mark Thorne
J.D. Gibbs
Kristina Riemer

University of Idaho
The University of Arizona
University of Hawai'i Mānoa
Colorado State University Extension

Funding provided by a USDA Natural Resources Conservation Service Conservation Innovation Grant NR193A750008G003 & NR233A750008C005.

This work is supported by the Renewable Resources Extension Act Program [award no. 2021-46401-34740] from the USDA National Institute of Food and Agriculture.

Additional assistance provided by members of the Rangelands Partnership and the Altar Valley Conservation Alliance.

docs.rangelandsgateway.org

Jason Karl, University of Idaho

Overview of RangeDocs

Matt King, University of Arizona

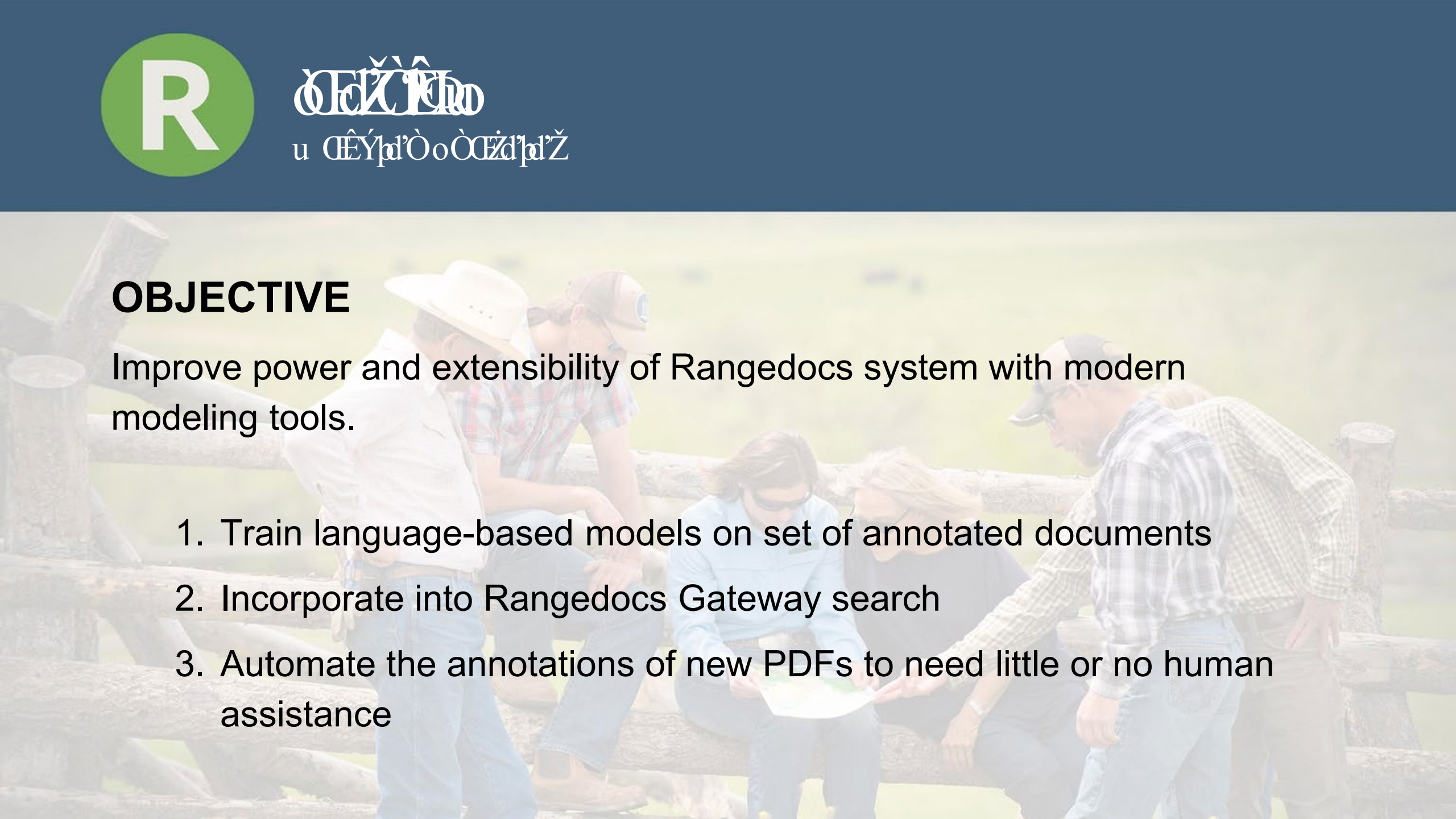
RangeDocs WebApp

Kristina Riemer & J.D. Gibbs, University of Arizona

Machine Learning & AI

Question & Answer Session

docs.rangelandsgateway.org



OBJECTIVE

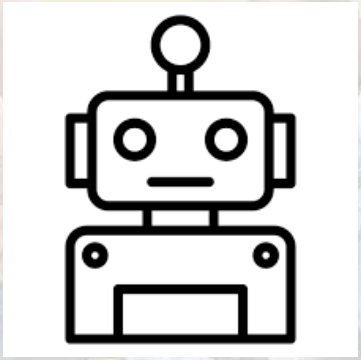
Improve power and extensibility of Rangedocs system with modern modeling tools.

1. Train language-based models on set of annotated documents
2. Incorporate into Rangedocs Gateway search
3. Automate the annotations of new PDFs to need little or no human assistance

Created *machine learning models* to classify new documents using annotated documents for training

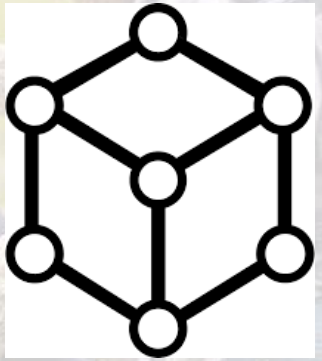
Text & tags

+

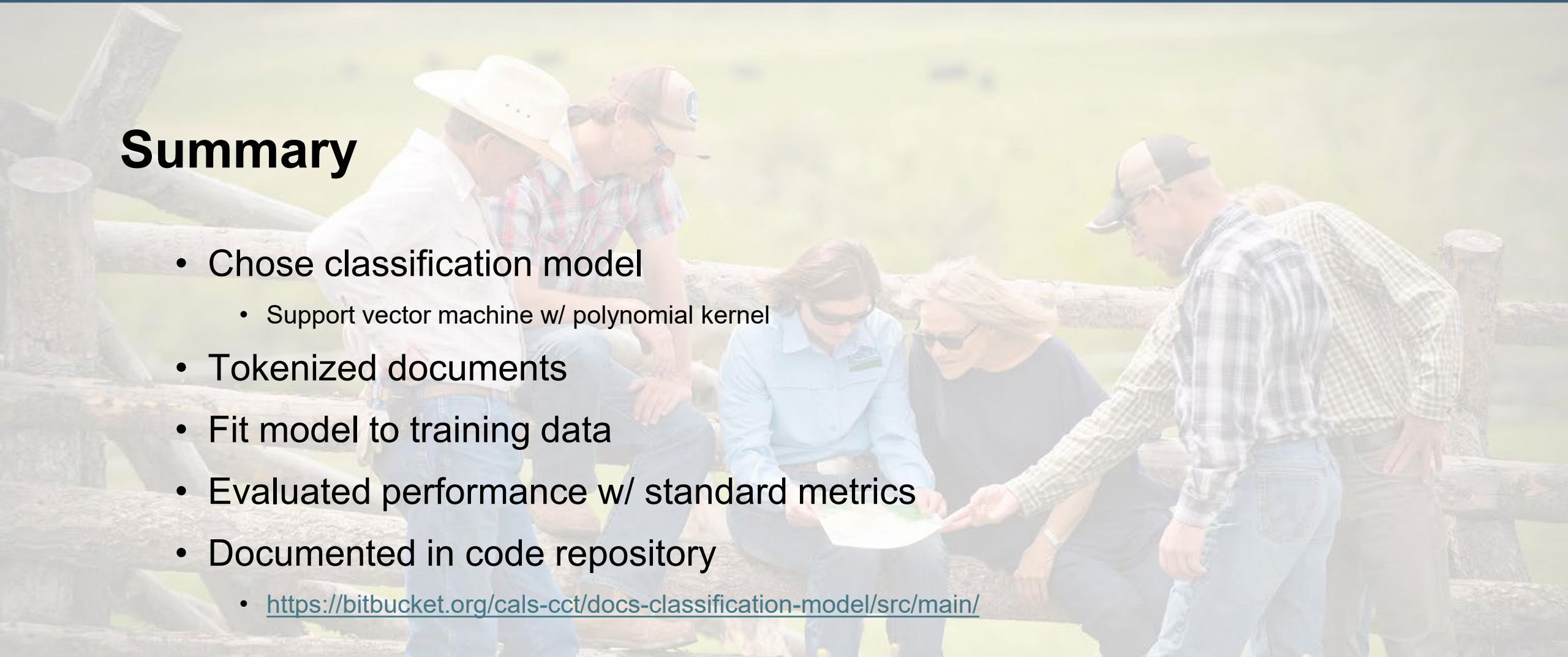


Algorithm

=



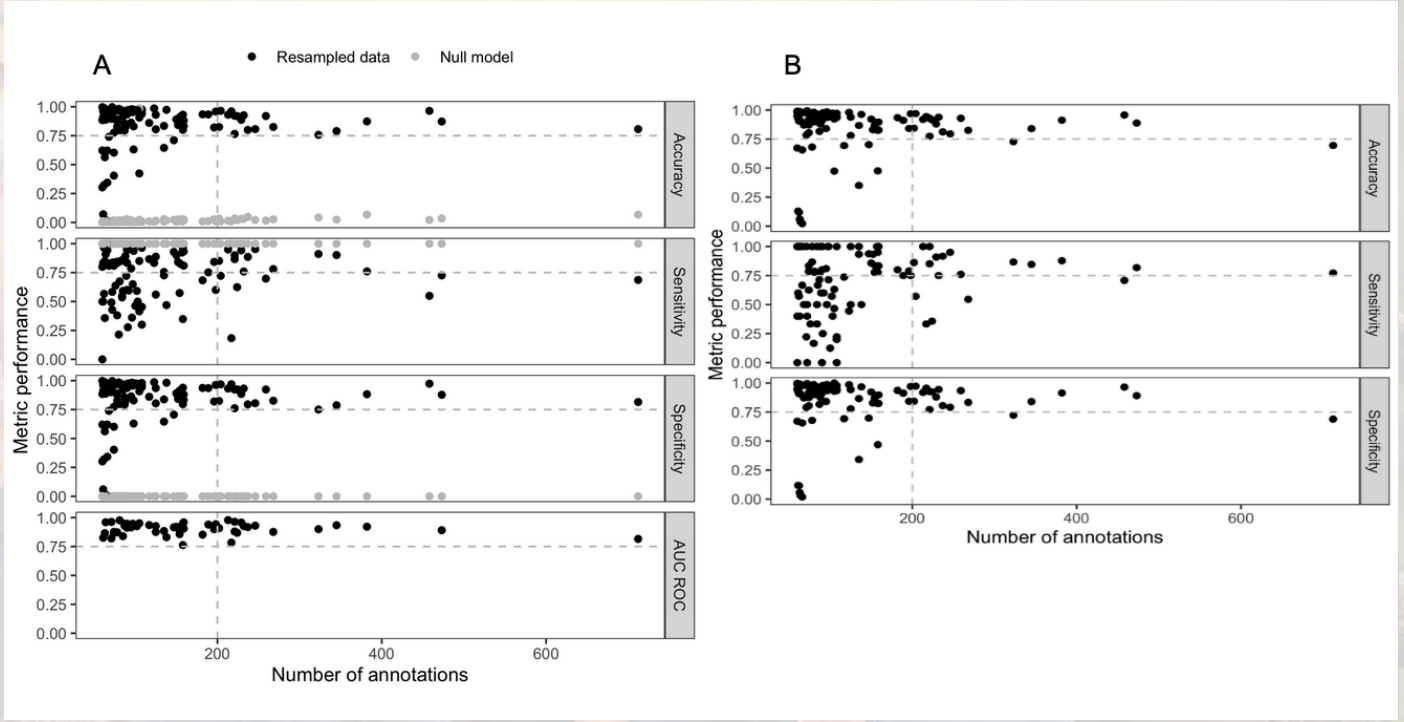
Model for each tag

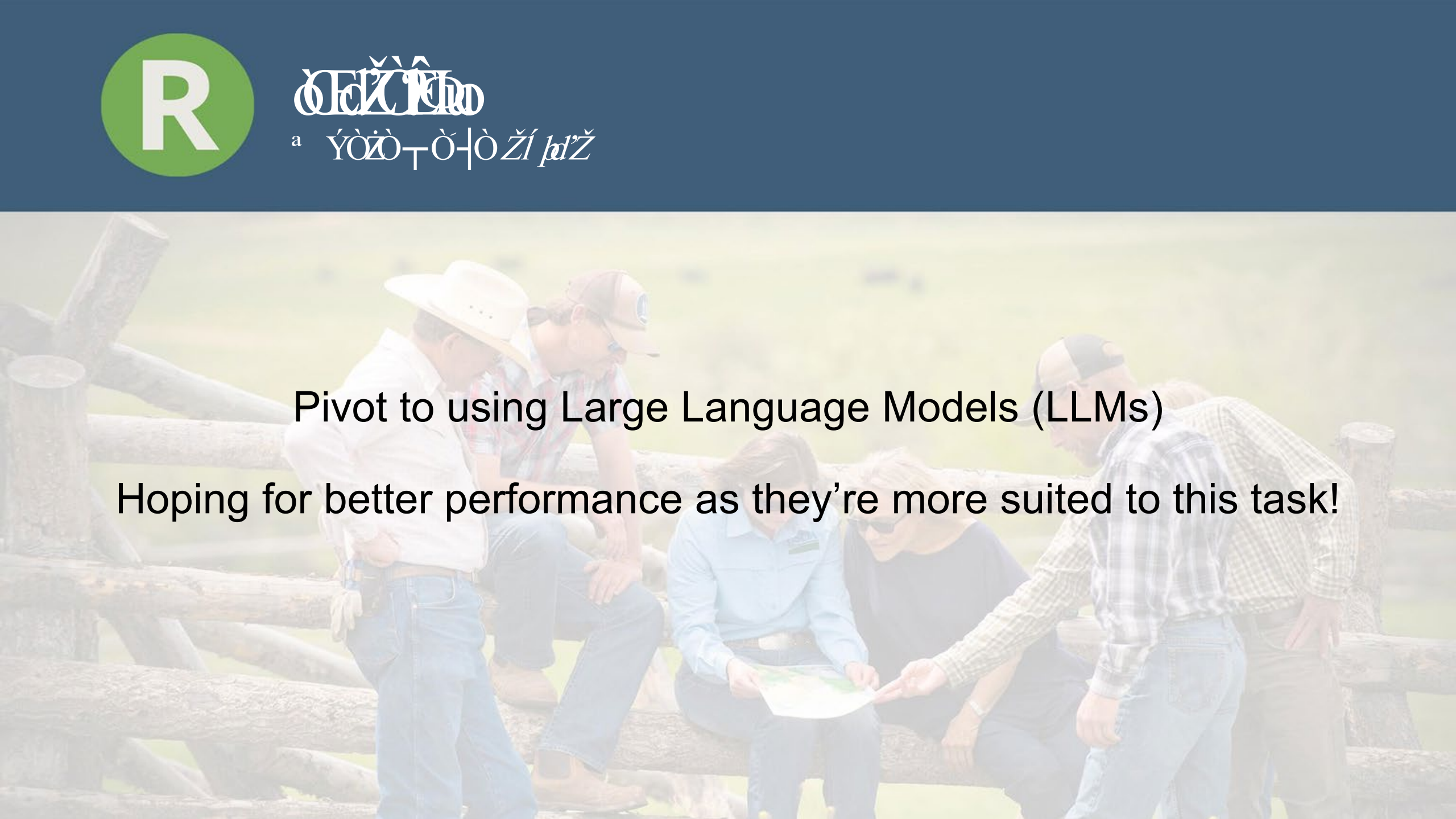


Summary

- Chose classification model
  - Support vector machine w/ polynomial kernel
- Tokenized documents
- Fit model to training data
- Evaluated performance w/ standard metrics
- Documented in code repository
  - https://bitbucket.org/cals-cct/docs-classification-model/src/main/

- Results
  - Can only reliably predict correct thesaurus terms if tagged at least 200 times in text
- Most terms were annotated much fewer times, so ML models weren't as useful
- Then LLMs came on the scene!





Pivot to using Large Language Models (LLMs)

Hoping for better performance as they’re more suited to this task!

RangeDocsML

What are LLMs?

- Large Language Models (LLMs) are advanced AI models designed to understand, generate, and interact with human language in a sophisticated manner
  - Trained on vast amounts of text data from diverse sources to develop a nuanced understanding of language
  - Can perform a wide range of language-related tasks, including text generation, translation, summarization, and question-answering
- Examples:
  - ChatGPT (OpenAI)
  - Gemini (Google)
  - Claude (Anthropic)
  - LLama (Meta)

OpenAI

Gemini

Claude

BY ANTHROP\C

Meta

Tokenize thesaurus terms

Tokenize range documents

Use pretrained models and fit to our use case

Evaluate model performance

Tokenize thesaurus terms

Store and retrieve data as vectors in a vector database

- Provides efficient search capabilities for fast retrieval of relevant documents

Tokenize range documents

Split up text data into smaller pieces (called chunking) + use sentence transformers

- Both make retrieval of document data faster and easier

Use pretrained models and fit to our use case

Chose dense passage retrieval models

- Enables asking a question using terms of interest and getting relevant document text back
- Compare multiple models
- Tune hyperparameters

RangeDocs LLM Overview

Evaluate model performance

Determine how well model does using standard metrics

ŕŽĊpd'
űŽ ŽŻÒΦΦ

- Tokenize thesaurus terms ✓
- Tokenize range documents ✓
- Use pretrained models and fit to our use case
- Evaluate model performance

òŒd’ŽÒ?ÍÊΦuo৩ ¢ŻŒþd’þd’Ž

OpenSource Models (HuggingFace)

- Testing models fine-tuned for retrieval to see which best fits our use case
- Evaluate model performance with metrics and adjust with Hyperparameter Tuning for optimal results

Vector Database

- Search Optimization: Provides efficient search capabilities for fast retrieval of relevant documents
- chromadb or pinecone.io

Fine Tuning

- Training with Jetstream2 GPU Allocation
- Flexible codebase for multiple collaborators, shared learning and quick coordinations
  - Using Tools such as: Python, Anaconda, JupyterLabs/Notebooks, Bitbucket/GitHub

User Testing

- A/B Testing within WebApp

òŒd”ŽÒ?ÍÊ◊uœ·U—→Ż!!ÝÒŻòÒŒËþd”Ž

- **Infrastructure**
  - Portable learning environments for hands-on computational instruction: Using container- and cloud-based technology to teach data science – ResearchGate.net
- **Task-Specific Models**
  - HuggingFace.co provides Open-Source ML infrastructure, models pre-trained on tasks, fine-tuning pipelines for transfer learning.
  - Models like `msmarco` (Sbert docs) are fine-tuned specifically for retrieval tasks, improving accuracy.
- **Chunking Strategies**
  - Breaking down large pieces of text into smaller segments, optimizing the relevance results from a vector database – Pinecone.io
- **Dense Passage Retrieval (DPR)**
  - High Accuracy model excelling at retrieving passages by creating dense representations of queries and passages.
  - Dual Encoder Architecture: Uses separate encoders for queries and passages, allowing efficient indexing and retrieval, re-ranking and fine-tuning to boost retrieval performance.
  - DPR can utilize models that that make it easier to retrieve relevant documents based on semantic similarity, such as those from **Sentence Transformers**, but it employs its own architecture for retrieval tasks
- **Retrieval-Augmented Generation (“RAG”)**
  - Combine the powers of pretrained DPR and sequence-to-sequence models
  - Retrieves documents, passes to a fine-tuned seq2seq model, then generates outputs, allowing both retrieval and generation to adapt to downstream tasks.

RangeDocs
Searchable Science for Rangeland Management

Project Team

Jason Karl
Amber Dalke
Sean DiStefano
Barb Hutchinson
Jeremy Kenyon
Matt King
Karen Launchbaugh
Sheila Merrigan
Jeanne Pfander
Matt Rahr
George Ruyle
Eric Winford
Retta Bruegger
Mark Thorne
J.D. Gibbs
Kristina Riemer

University of Idaho
The University of Arizona
University of Hawai'i Mānoa
Colorado State University Extension

Funding provided by a USDA Natural Resources Conservation Service Conservation Innovation Grant NR193A750008G003 & NR233A750008C005.

This work is supported by the Renewable Resources Extension Act Program [award no. 2021-46401-34740] from the USDA National Institute of Food and Agriculture.

Additional assistance provided by members of the Rangelands Partnership and the Altar Valley Conservation Alliance.

docs.rangelandsgateway.org

RangeDocs

Searchable Science for Rangeland Management

docs.rangelandsgateway.org

THE RANGELANDS PARTNERSHIP

University of Idaho

THE UNIVERSITY OF ARIZONA

UNIVERSITY of HAWAI'I MANOA

COLORADO STATE UNIVERSITY EXTENSION

Funding provided by a USDA Natural Resources Conservation Service Conservation Innovation Grant NR193A750008G003 & NR233A750008C005.
This work is supported by the Renewable Resources Extension Act Program [award no. 2021-46401-34740] from the USDA National Institute of Food and Agriculture.