GHD Digital

How AI is helping improve data transparency and regulatory compliance for environmental sites

Jonathan Eller, Ph.D.
Data Scientist – USA DIz - Advanced Analytics & AI

How can data improve understanding?

Shared goal: To clean sites well and efficiently

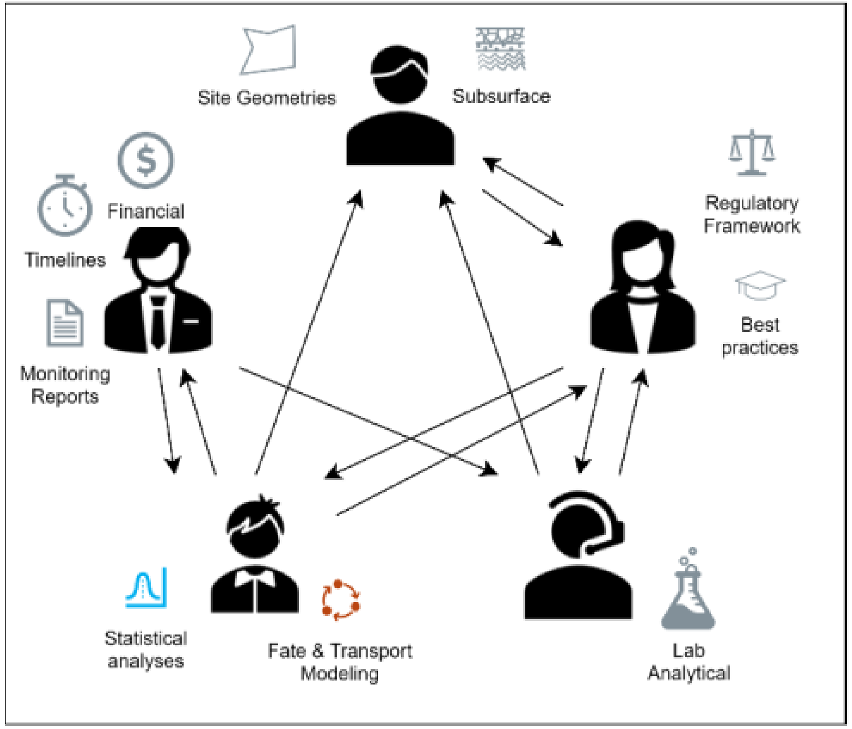
> Then why is cleaning sites difficult?

A major barrier to site cleanups is a **lack of understanding** between:

– An organization and its own data

– RP <-> Regulator, Regulator <-> Public, RP <-> Public

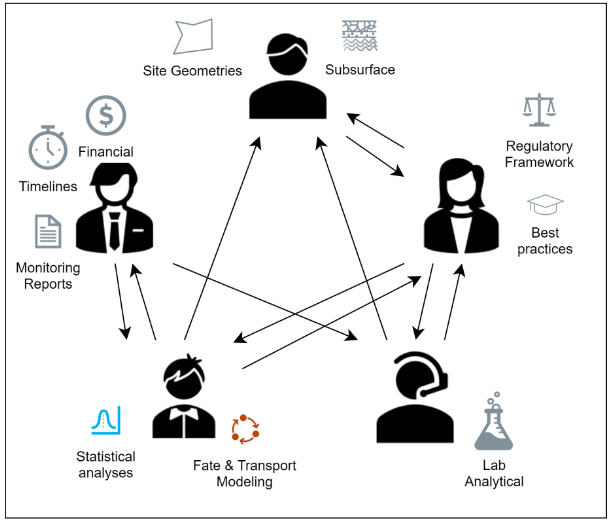
We believe harmonization of data can help bridge these gaps by building transparency and trust

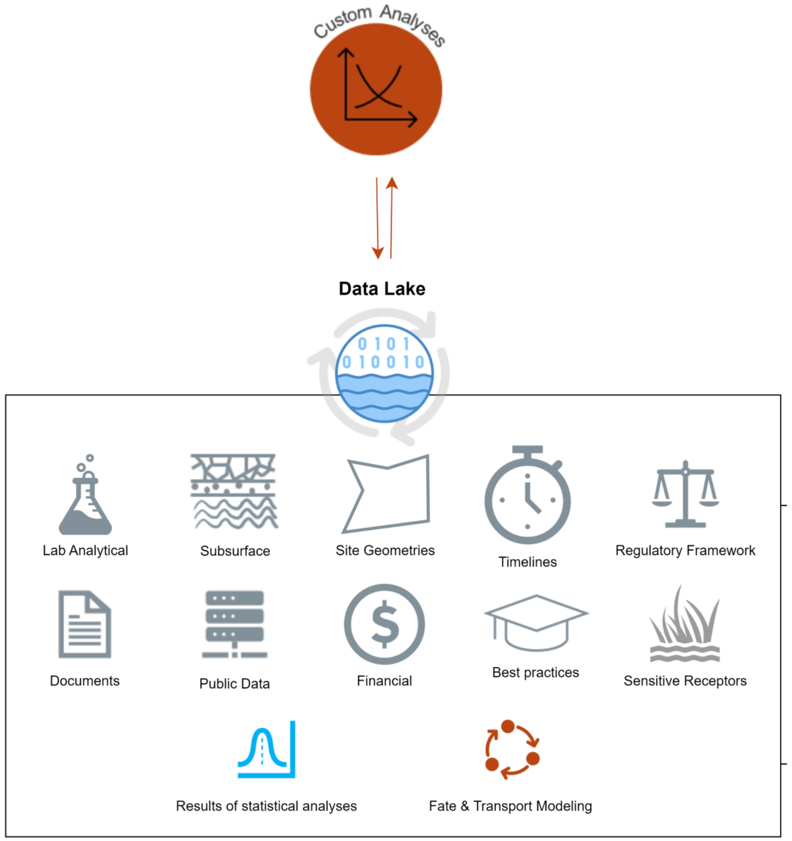


Chaotic data overload
exists within and between organizations

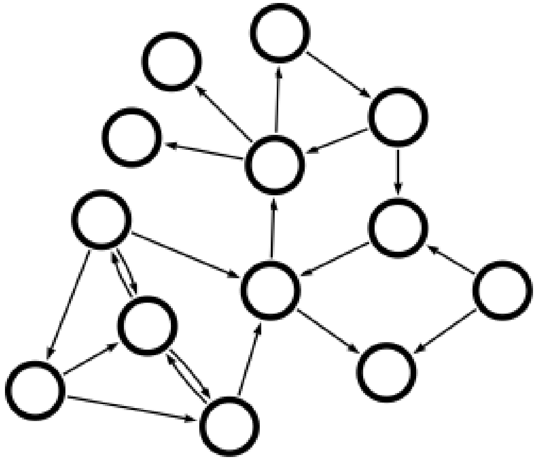
From chaos to connection

Organizational Chaos





Knowledge Graph



- Current state
- Very person-dependent
- Redundant and manual

- Data in one place with governance
- But how are they related?

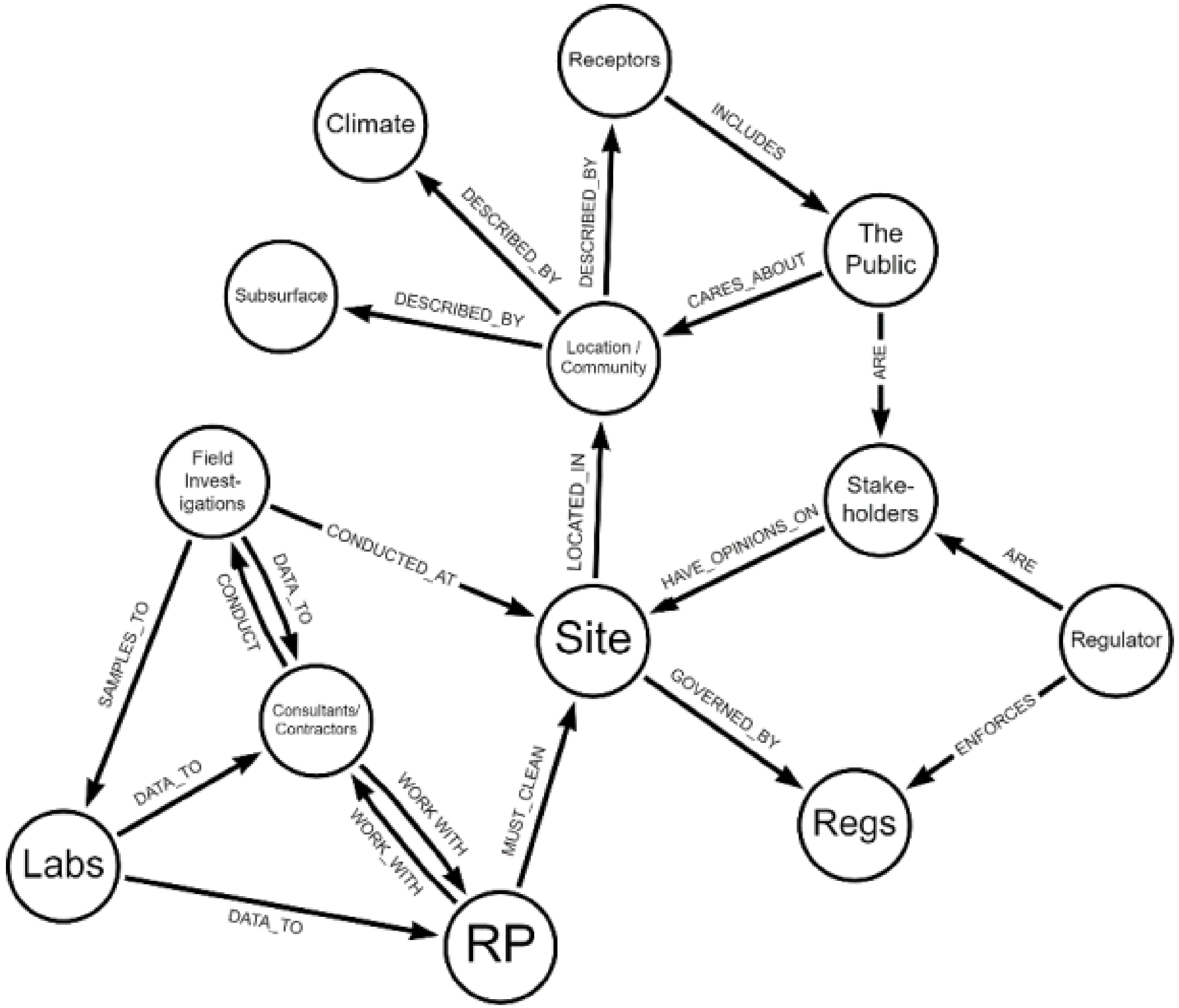
- “A Living CSM”
- Programmatic holistic **question answering**

Toward a knowledge graph of contaminated sites

What and Why?

- A graph database (aka, "graph") stores nodes and relationships instead of tables or documents.
- It's NoSQL – so we are flexible to site-specifics
- The extent to which we understand or can "quantify" a site can be explained by the richness of this graph (how many nodes, node types, and relationship types)
- Aka, an "Augmented, Living Site Conceptual Model"
- Each site's graph may look different, and graph-level measures of complexity or connectedness could be meaningful

Example graph of a site



 I

Riddle me this

Answering (the right) questions is the heart of regulatory compliance

Our ability to navigate regulations depends on:

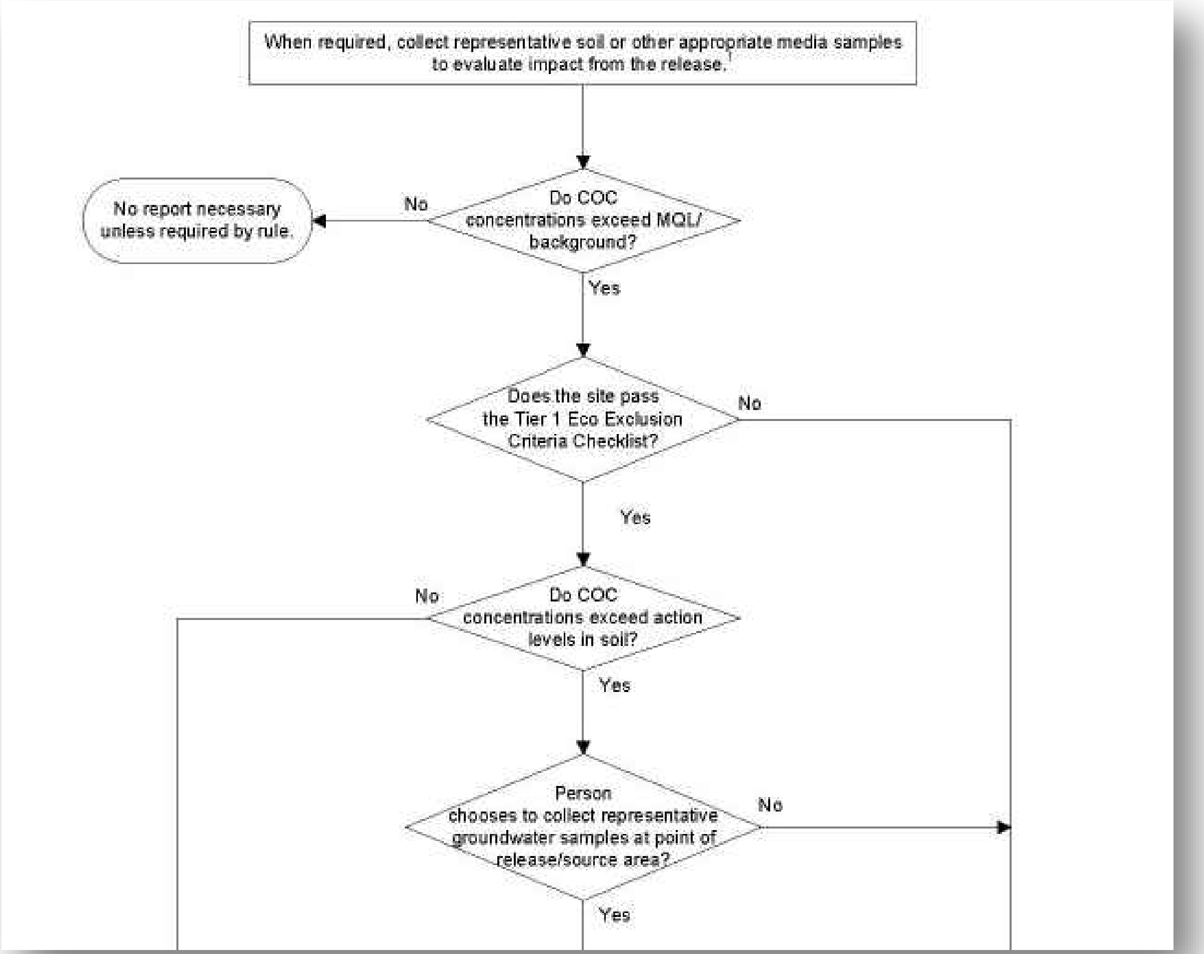
- Knowing what questions need to be answered
- Answering those questions accurately

With the correct data in place, connected and structured properly, answering questions is simply a matter of processing the data.

"*But **processing text is hard***" you say. True.

That's where Large Language Models (LLMs) come in…

Example: Determining regulatory jurisdiction



What if we could navigate this ↑ programmatically?

My Site Might be Contaminated. What Do I Do? - Texas Commission on Environmental Quality - www.tceq.texas.gov. Texas Risk Reduction Program (TRRP) Guidance document

LLMs can be a great interface to (and between) data

Large Language Models (LLMs)

Generative AI models that accept and output text that looks human-generated

Use Cases:

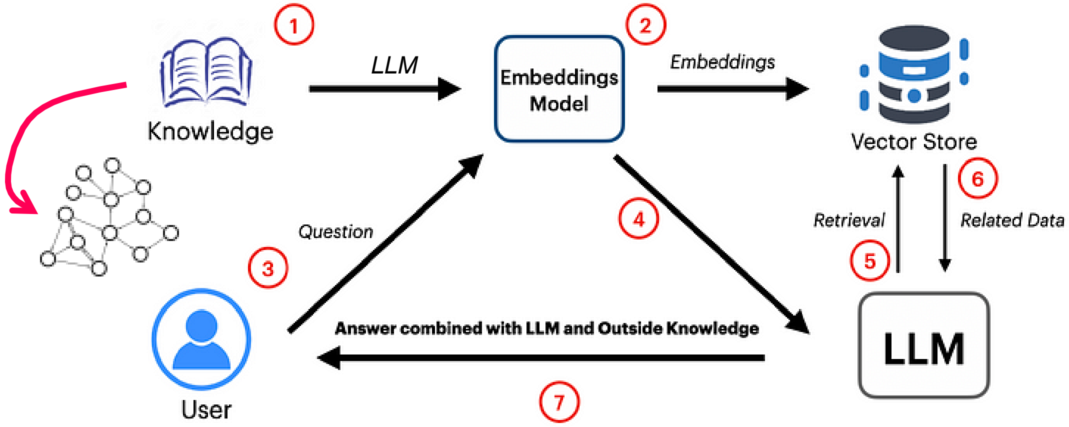
- Text summarization
- Chatbots
- Code generation
- Question Answering (**limited**)

Limitations/Risks:

- Hallucinations
- How to "focus" them on specific information?

Retrieval-Augmented Generation (RAG)

A way to use LLMs to answer specific questions of a specific dataset



Improvements over vanilla LLMs:

- Accuracy
- Informity
- Relevance
- Versatility
- Niche Dataset
- Reduce Hallucinations

Challenges:

- Conflicts between documents
- Over-confidence
- Validation is subjective
- Timeliness of Information
- Hallucination still exists

The best LLMs of 2024 | TechRadar; Introduction to Large Language Models | Machine Learning | Google for Developers I

Exploring the RAG design parameter space

Research Goal:

– Define “accuracy” for LLM-based Question Answering systems

– Use this to metric to optimize the systems ability to solve problems

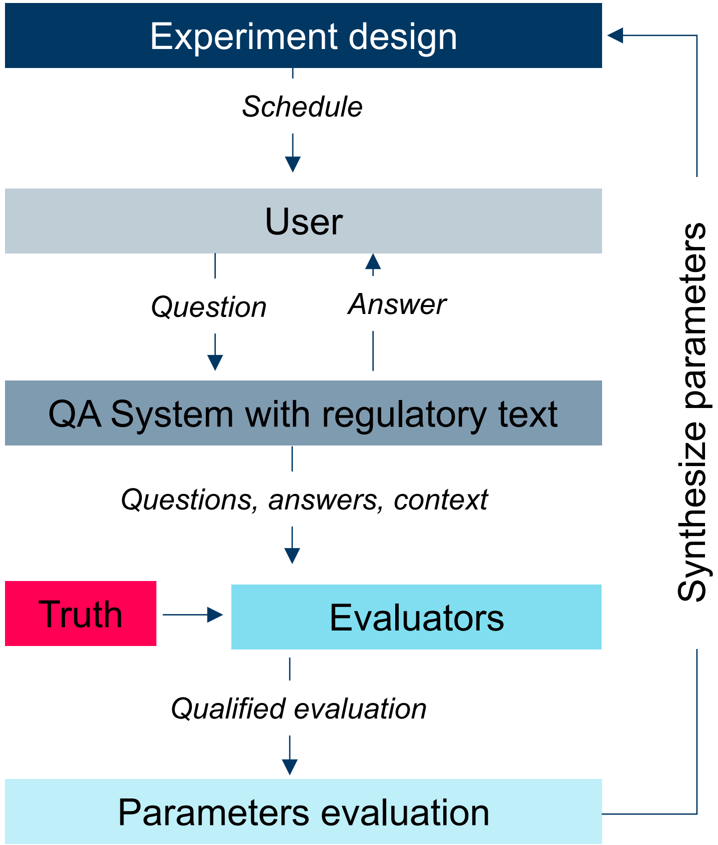
Experimental design

– Regulatory text with FAQ document as ground truths

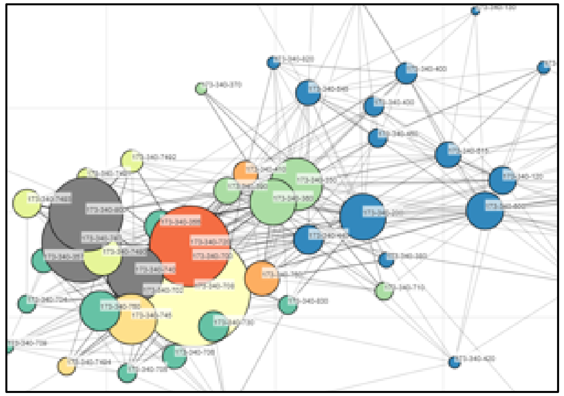
– Custom model configuration

– 40 experiments

– Adopted RAG evaluation framework



Network Graph of Regs



Parameters explored

– **Prompt Engineering:** How we structure meta-instructions to the LLM

– **Chunking approach:** How the graph is constructed from the text

– **Num. examples:** How many few-shot examples are provided to the LLM

– **Temperature:** How “creative” responses are

Evaluate Against

– **Relevance:** Ability to capture key points

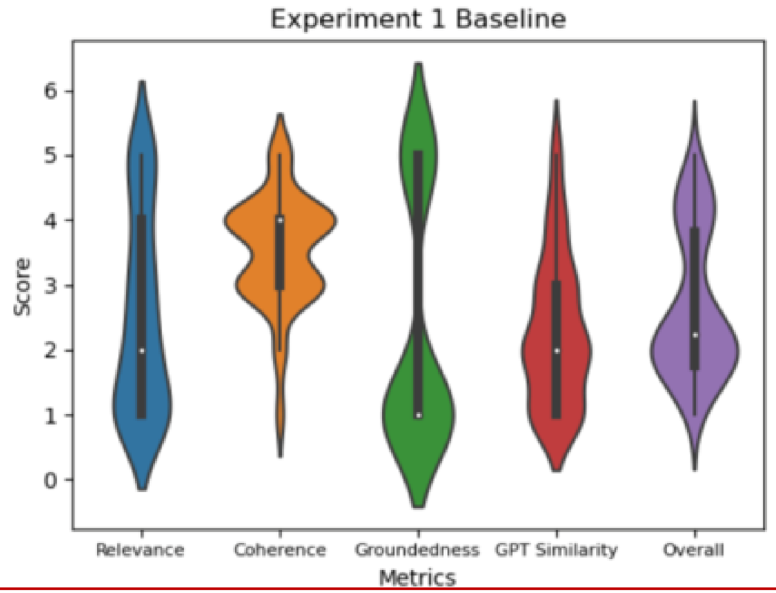
– **Coherence:** How human-like the response it

– **Groundedness:** Similarity between AI claims and ground-truth

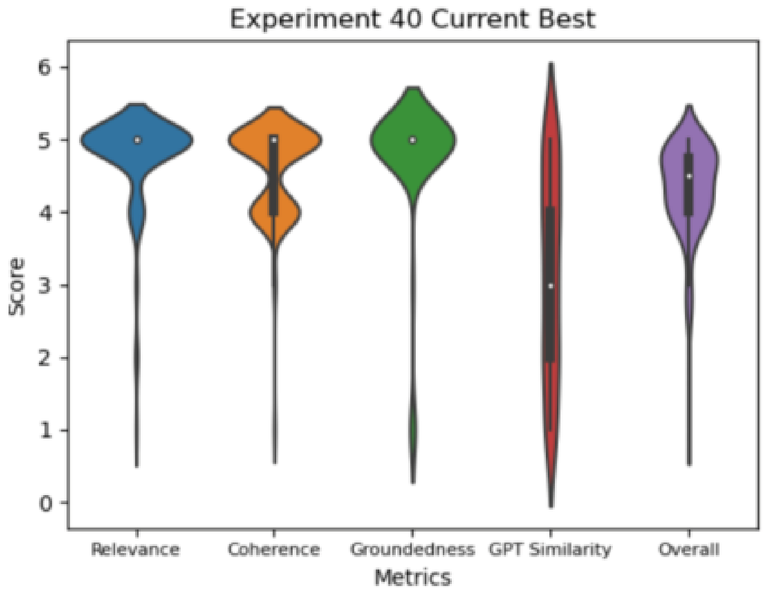
– **Dontknowness:** How confident is it when it’s wrong

RAG evaluation – Qualifying improvement

Baseline Model



Current Best Model



Q52: Are the emissions factors listed in Table W-1A for both leaking components and non-leaking components? How do you calculate emissions from leaking components if onshore petroleum and natural gas source are not required to monitor components?

Baseline

The text contains information about calculating annual emissions of CO2, CH4, N2O, and fluorinated GHGs from various source categories using specific equations and methodologies. It also mentions the monitoring and reporting requirements for different types of facilities. However, there is no specific mention of emissions factors for leaking and non-leaking components in Table W-1A, nor is there a direct calculation method provided for emissions from leaking components for onshore petroleum and natural gas sources. Further analysis or additional information may be required to address the specific question about emissions factors for leaking components and the calculation of emissions from onshore petroleum and natural gas sources.

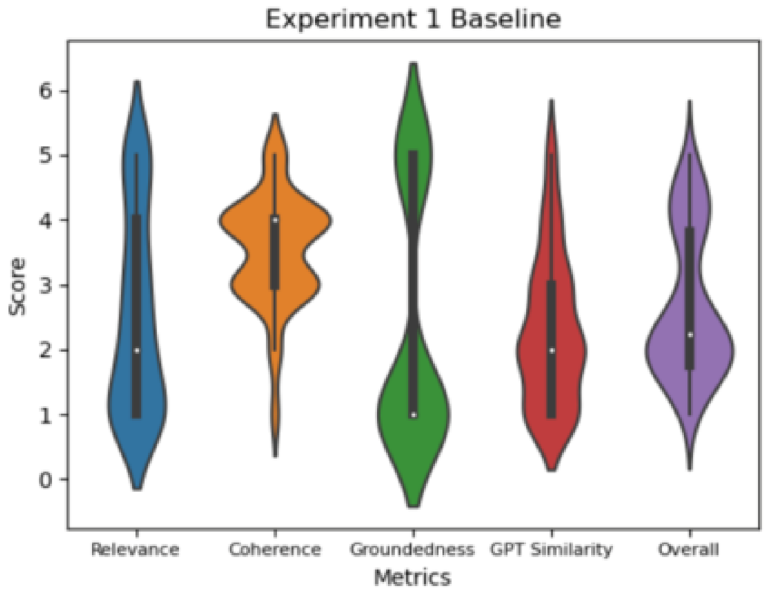
1

Relevance

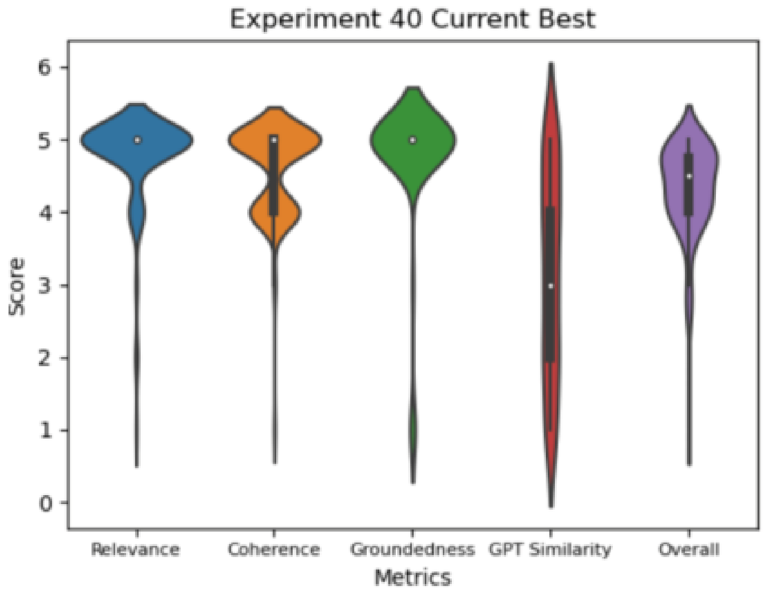
The answer completely lacks relevance to the question about emissions factors and calculation of emissions from leaking components for onshore petroleum and natural gas sources.

RAG evaluation – Qualifying improvement

Baseline Model



Current Best Model



Q52: Are the emissions factors listed in Table W-1A for both leaking components and non-leaking components? How do you calculate emissions from leaking components if onshore petroleum and natural gas source are not required to monitor components?

Current Best

The emissions factors listed in Table W-1A provide population emission factors for both leaking and non-leaking components. To calculate emissions from leaking components for onshore petroleum and natural gas sources where monitoring is not required, you can use the population emissions factors for leaking components listed in Table W-1A. The calculation involves counting all major equipment listed in Table W,Äì1B and Table W,Äì1C, and then multiplying the major equipment counts by the average component counts listed in the respective tables. These factors provide the emissions rate per unit of activity for different types of equipment and processes. The specific equations and methods for calculating emissions are determined based on the type of facility, equipment, and geographic location, as detailed in the regulations. No leak detection of emissions is required for onshore production. For further details on the concept of population emission factors, please refer to section 6(d) of the Technical Support Document

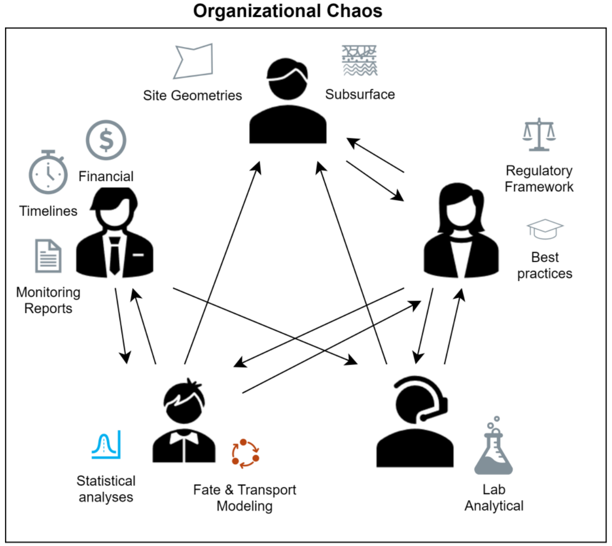
5

Relevance

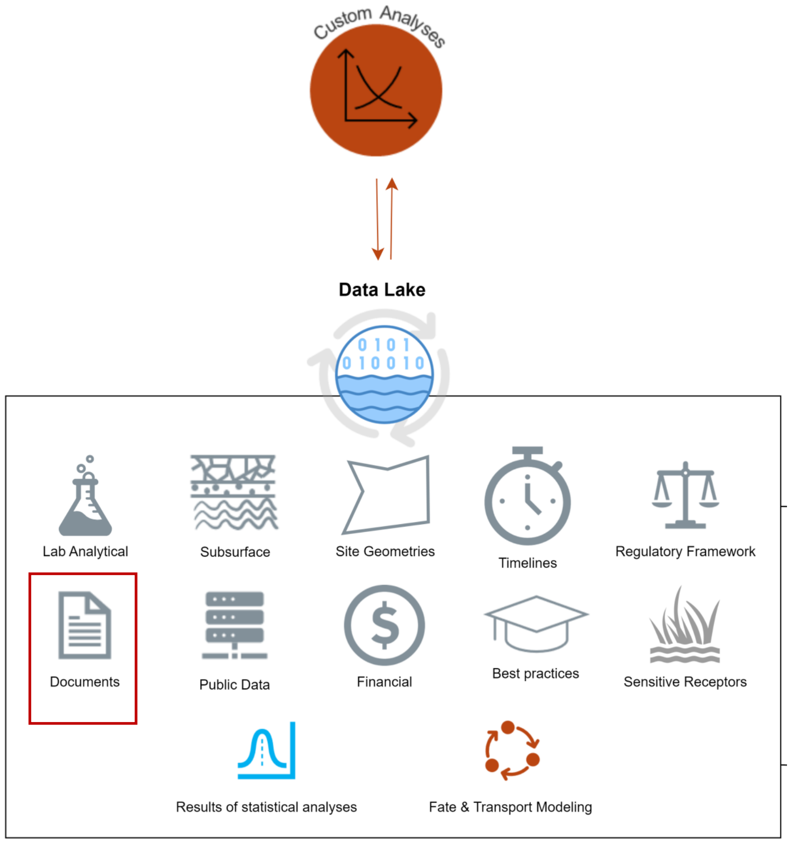
The answer provides a clear and accurate response to both parts of the question, addressing whether the emissions factors listed in Table W-1A are for both leaking and non-leaking components and how to calculate emissions from leaking components if monitoring is not required for onshore petroleum and natural gas sources. The answer also includes additional details on the calculation process and references the relevant regulations and technical support document for further information.

How we are building it

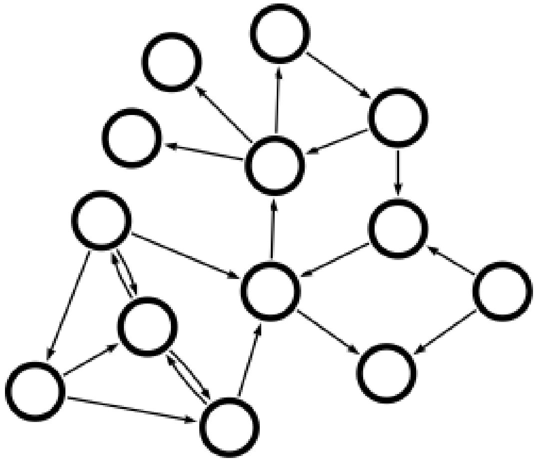
From chaos to connection



??



??



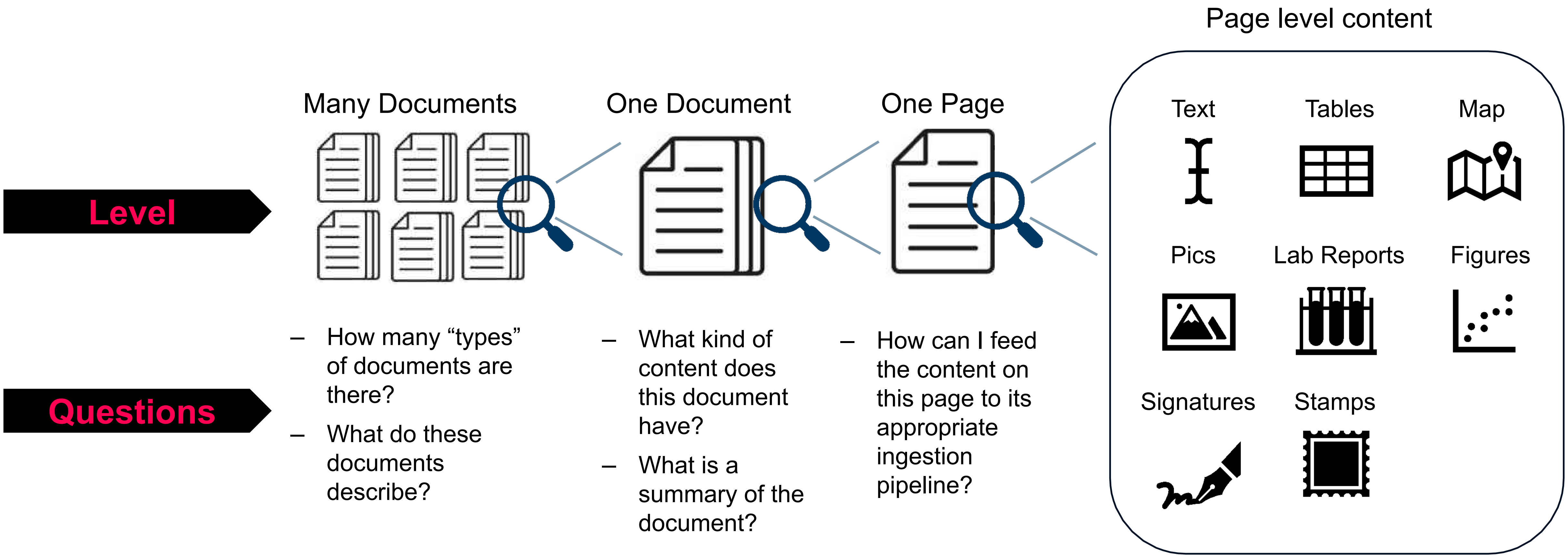
- Current state
- Very person-dependent
- Redundant and manual

- Data in one place with governance
- But how are they related?

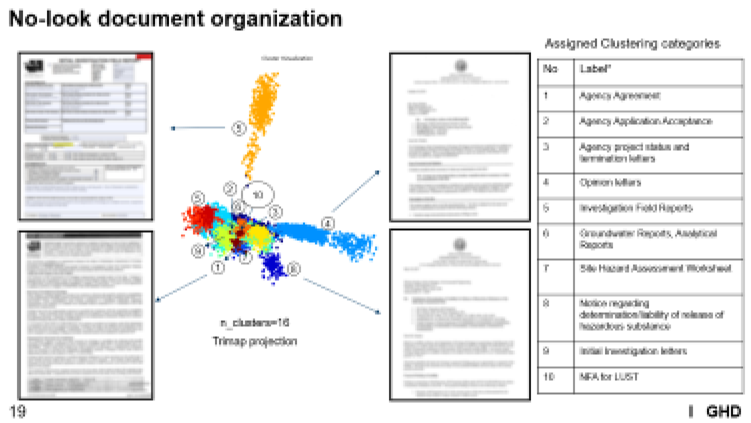
- A Living CSM
- Programmatic holistic **question answering**

Document hierarchy

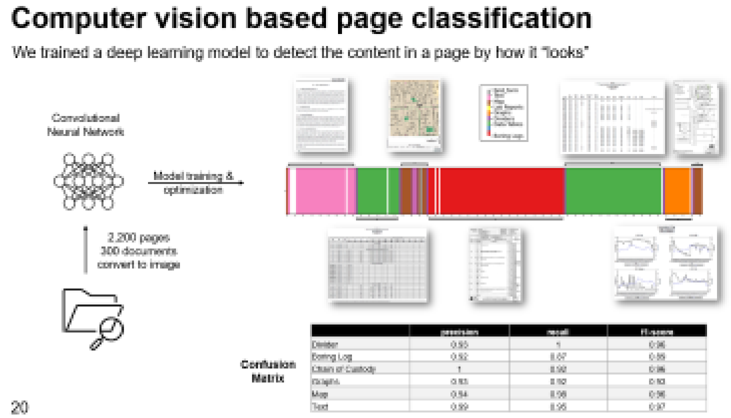
Deconstructing the data within a document is a layered problem, and understanding how documents tend to look for contaminated sites is key. This is much bigger than simply applying OCR.

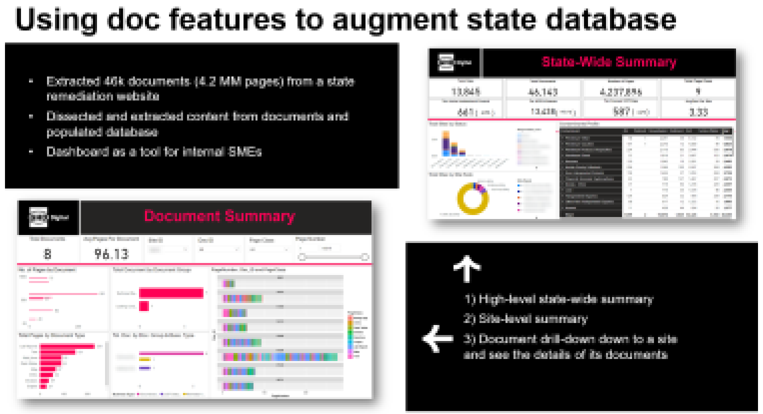


Sequential document deconstruction



Hierarchy Level	Our explorations / solutions
Many documents	– Document clustering
Single document	– Page classification with computer vision
Single page	- Content segmentation
Page content	- Lab report extraction - Excel-based analytical table extraction - Boring log extraction

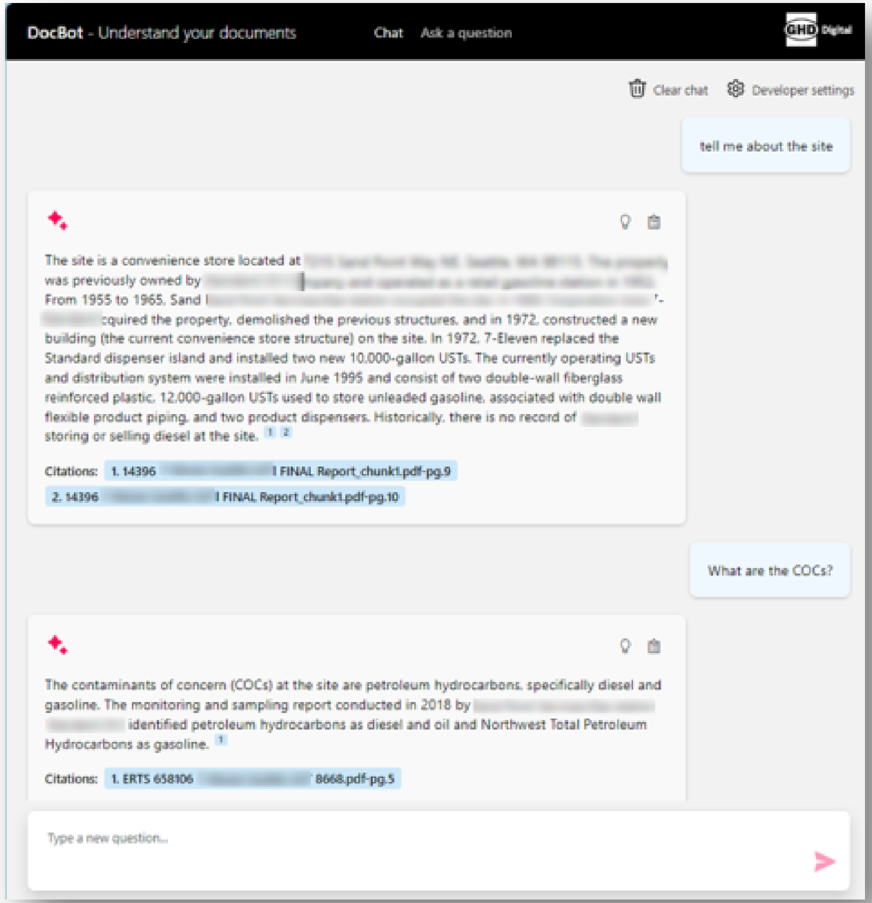




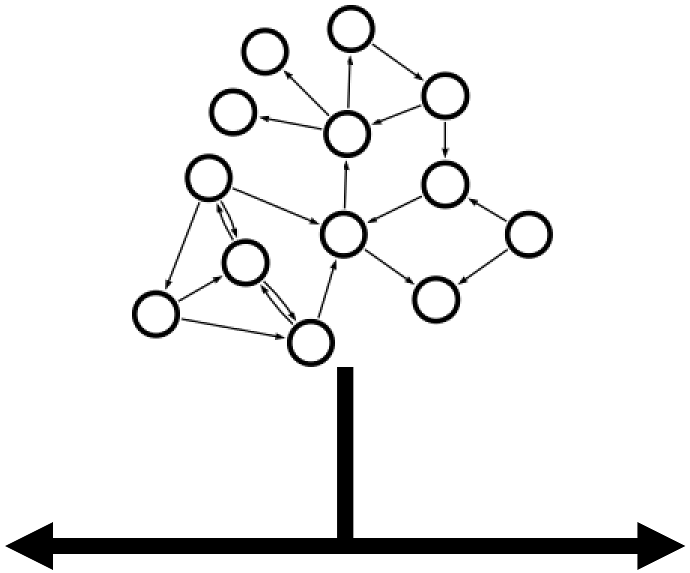
How we are using it

Example linguistic interfaces to site data

Q&A with site docs



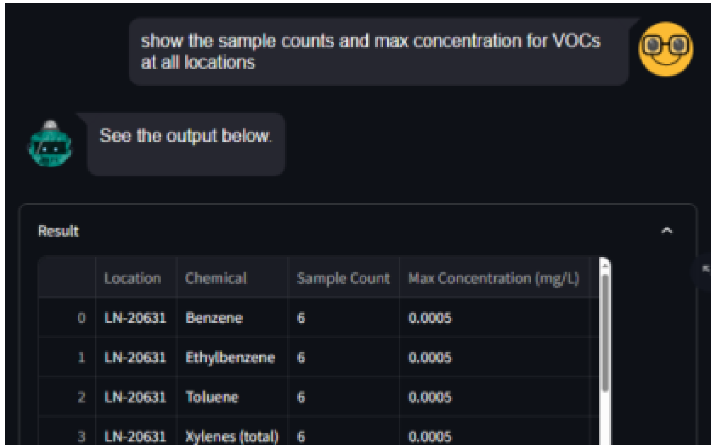
Site graph

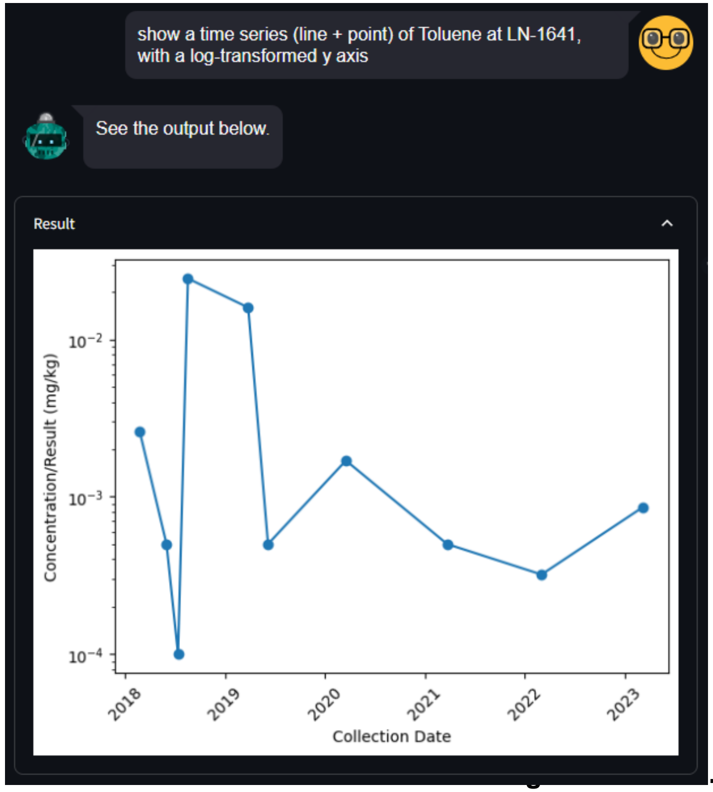


RAG Chat

Agents + Python/Pandas

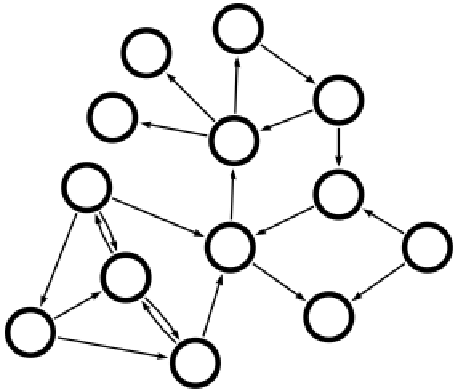
Q&A with tabular data



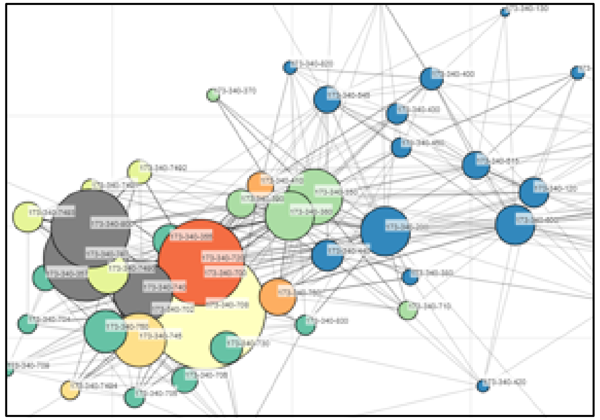


The future of regulatory compliance?

Site graph



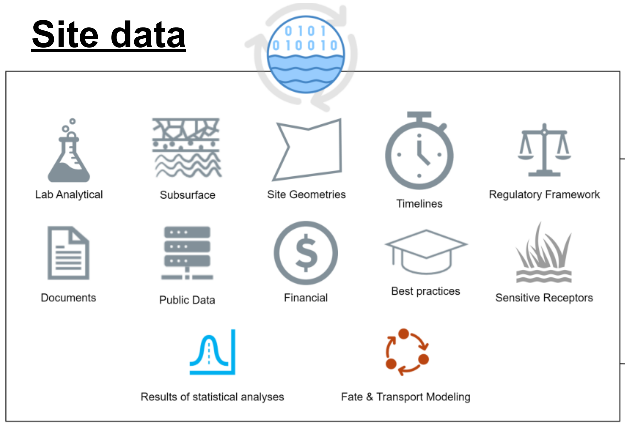
Regulatory graph



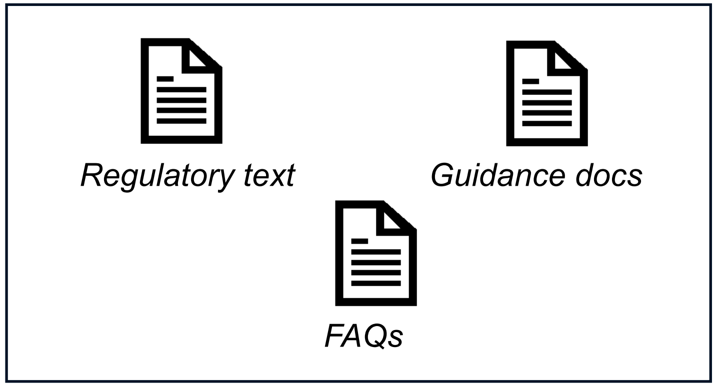
Questions

Answers

Site data



Regulation



Acknowledgements

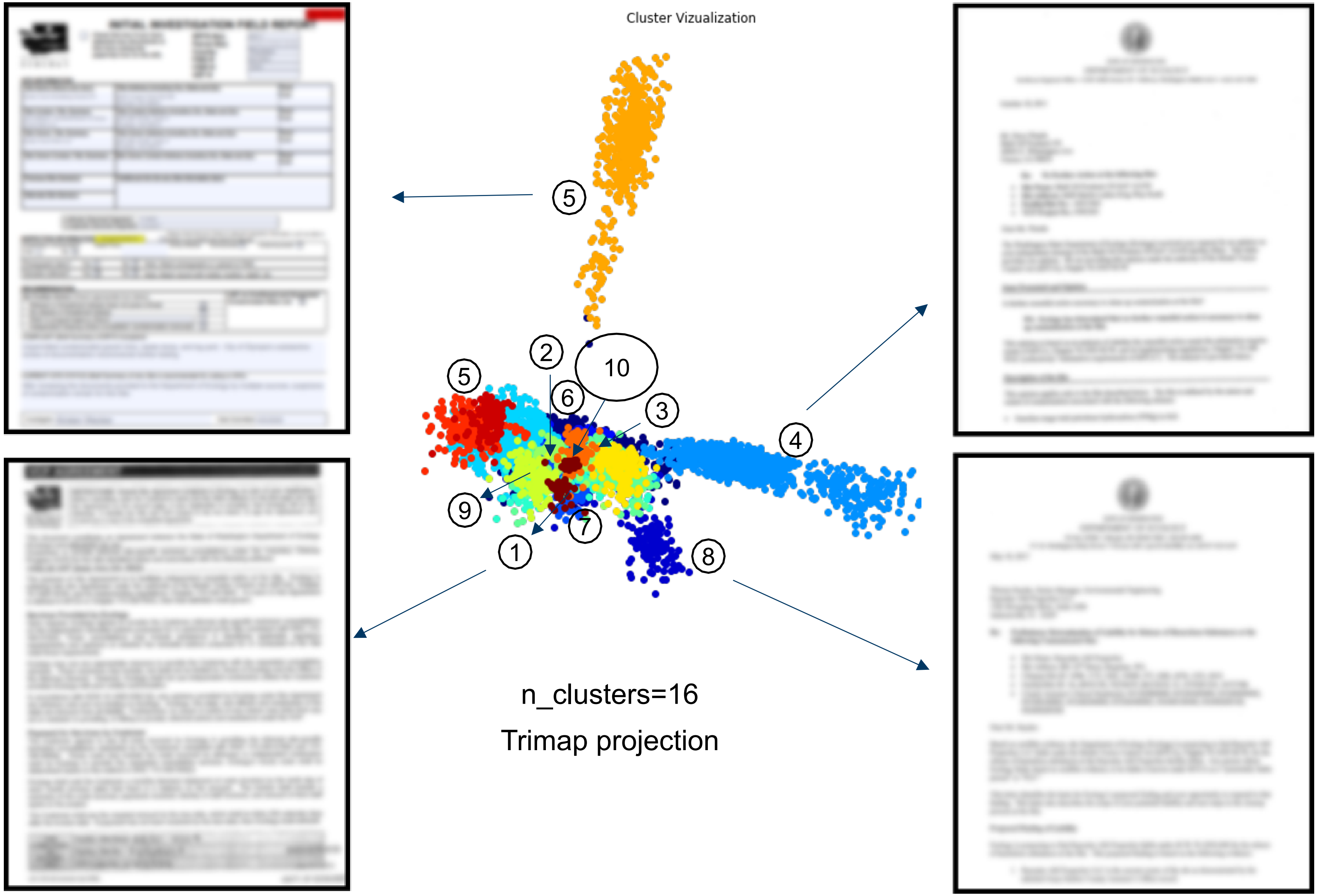
- GHD Environmental Data Science Team
- GHD Contamination Assessment & Remediation Team
- Columbia University Industrial Engineering & Operations Research Program

GHD Digital

*Thank You

→ ghd.com/digital

No-look document organization



n_clusters=16

Trimap projection

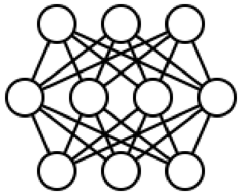
Assigned Clustering categories

No	Label*
1	Agency Agreement
2	Agency Application Acceptance
3	Agency project status and termination letters
4	Opinion letters
5	Investigation Field Reports
6	Groundwater Reports, Analytical Reports
7	Site Hazard Assessment Worksheet
8	Notice regarding determination/liability of release of hazardous substance
9	Initial Investigation letters
10	NFA for LUST

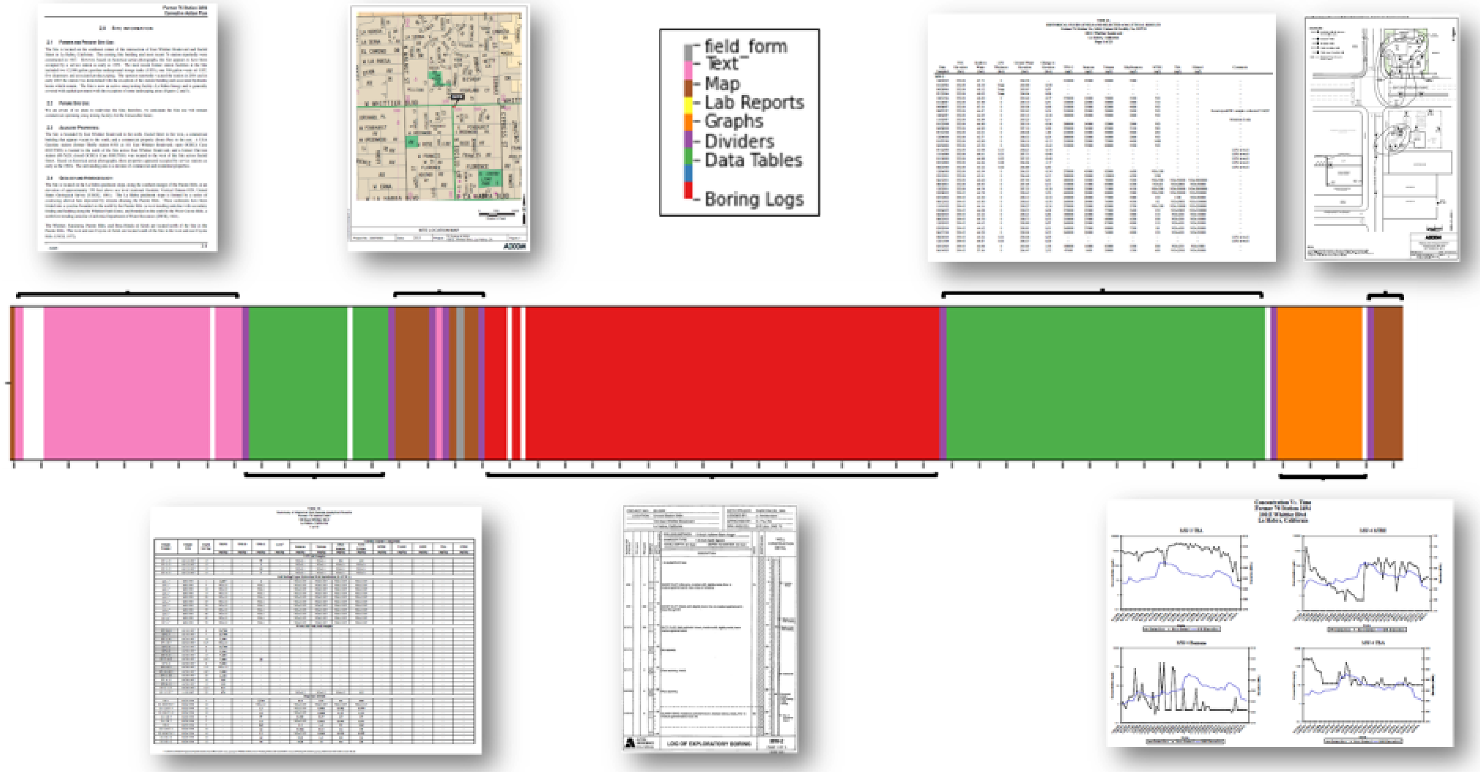
Computer vision based page classification

We trained a deep learning model to detect the content in a page by how it “looks”

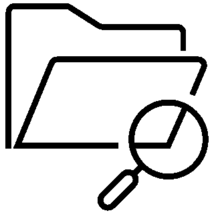
Convolutional
Neural Network



Model training &
optimization



2,200 pages
300 documents
convert to image

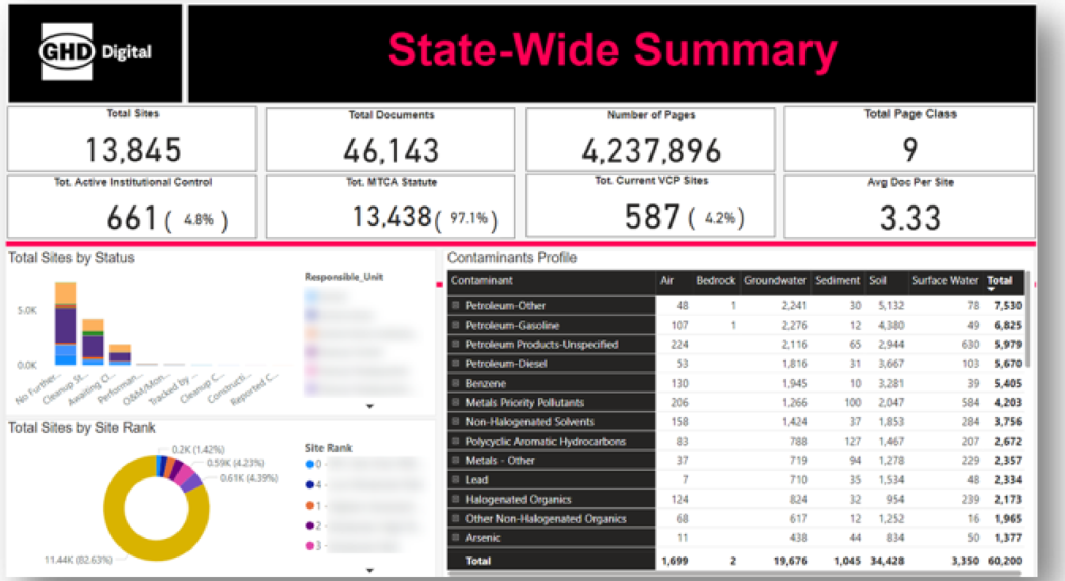


Confusion Matrix

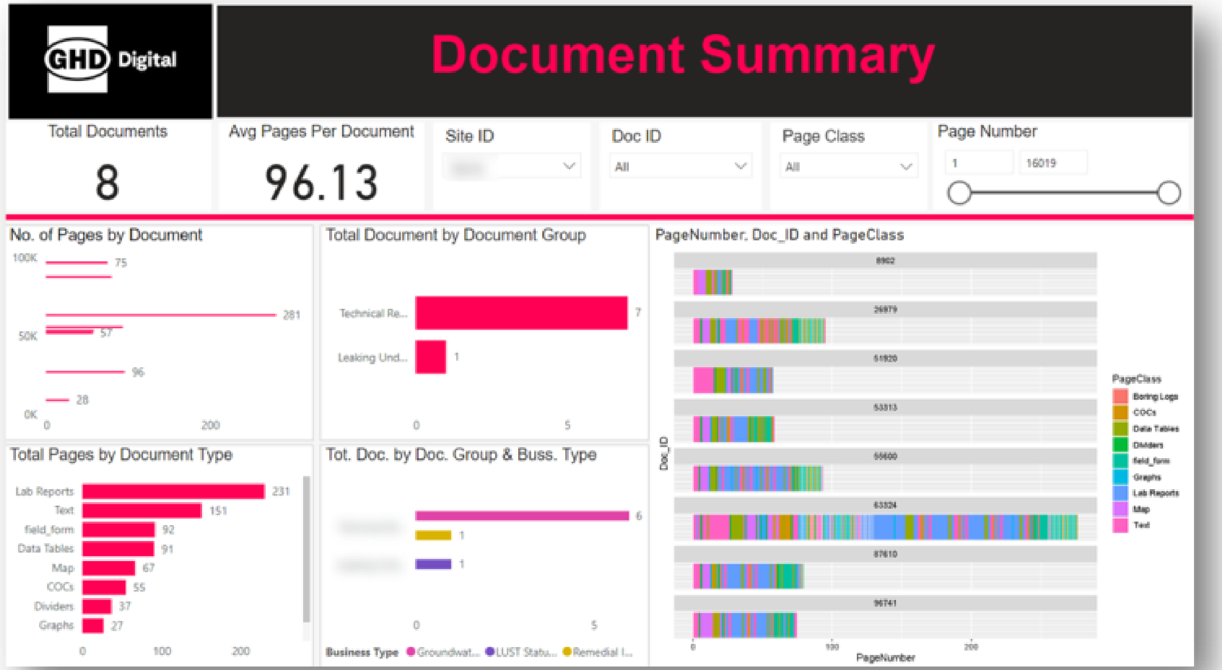
	precision	recall	f1-score
Divider	0.93	1	0.96
Boring Log	0.92	0.87	0.89
Chain of Custody	1	0.92	0.96
Graphs	0.93	0.92	0.93
Map	0.94	0.98	0.96
Text	0.99	0.95	0.97

Using doc features to augment state database

- Extracted 46k documents (4.2 MM pages) from a state remediation website
- Dissected and extracted content from documents and populated database
- Dashboard as a tool for internal SMEs



Contaminant	Air	Bedrock	Groundwater	Sediment	Soil	Surface Water	Total
Petroleum-Other	48	1	2,241	30	5,132	78	7,530
Petroleum-Gasoline	107	1	2,276	12	4,380	49	6,825
Petroleum Products-Unspecified	224		2,116	65	2,944	630	5,979
Petroleum-Diesel	53		1,816	31	3,667	103	5,670
Benzene	130		1,945	10	3,281	39	5,405
Metals Priority Pollutants	206		1,266	100	2,047	584	4,203
Non-Halogenated Solvents	158		1,424	37	1,853	284	3,756
Polycyclic Aromatic Hydrocarbons	83		788	127	1,467	207	2,672
Metals - Other	37		719	94	1,278	229	2,357
Lead	7		710	35	1,534	48	2,334
Halogenated Organics	124		824	32	954	239	2,173
Other Non-Halogenated Organics	68		617	12	1,252	16	1,965
Arsenic	11		438	44	834	50	1,377
Total	1,699	2	19,676	1,045	34,428	3,350	60,200

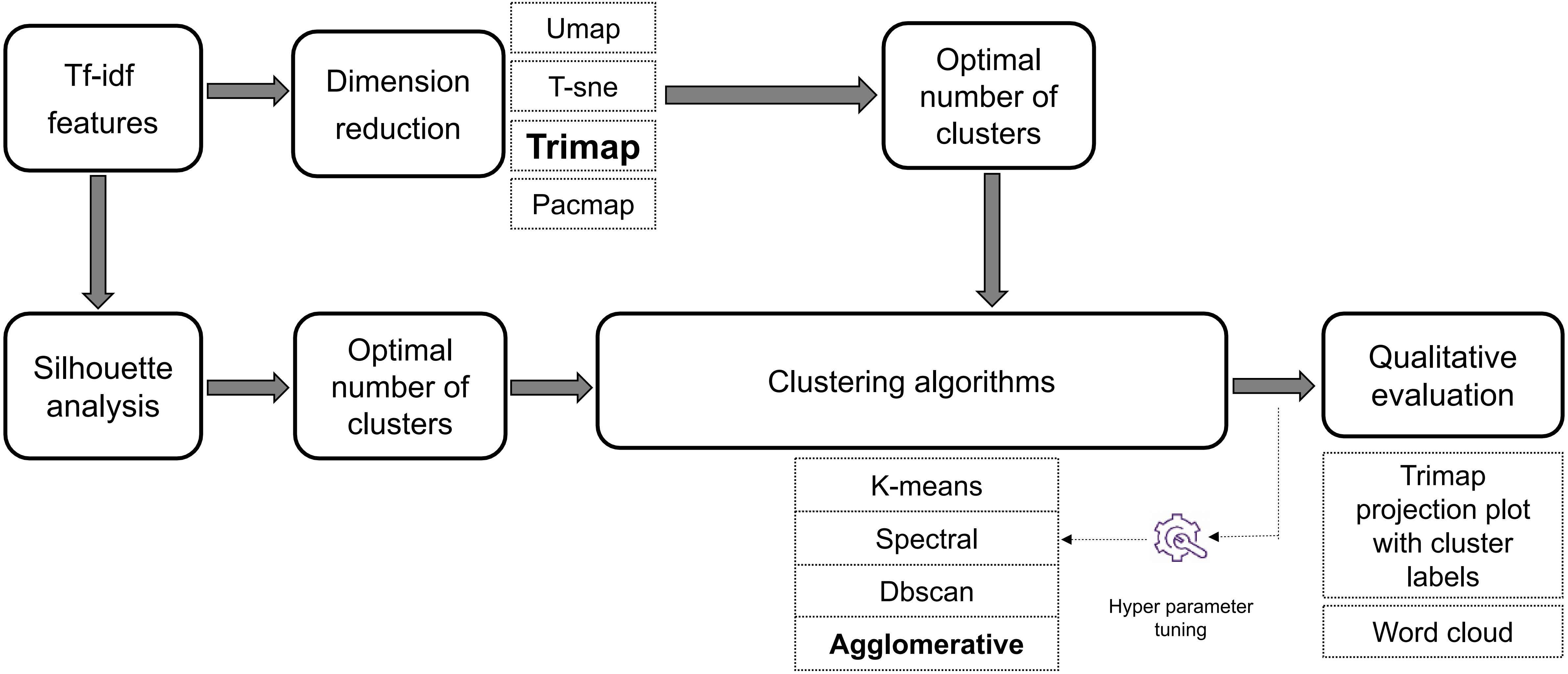


1) High-level state-wide summary

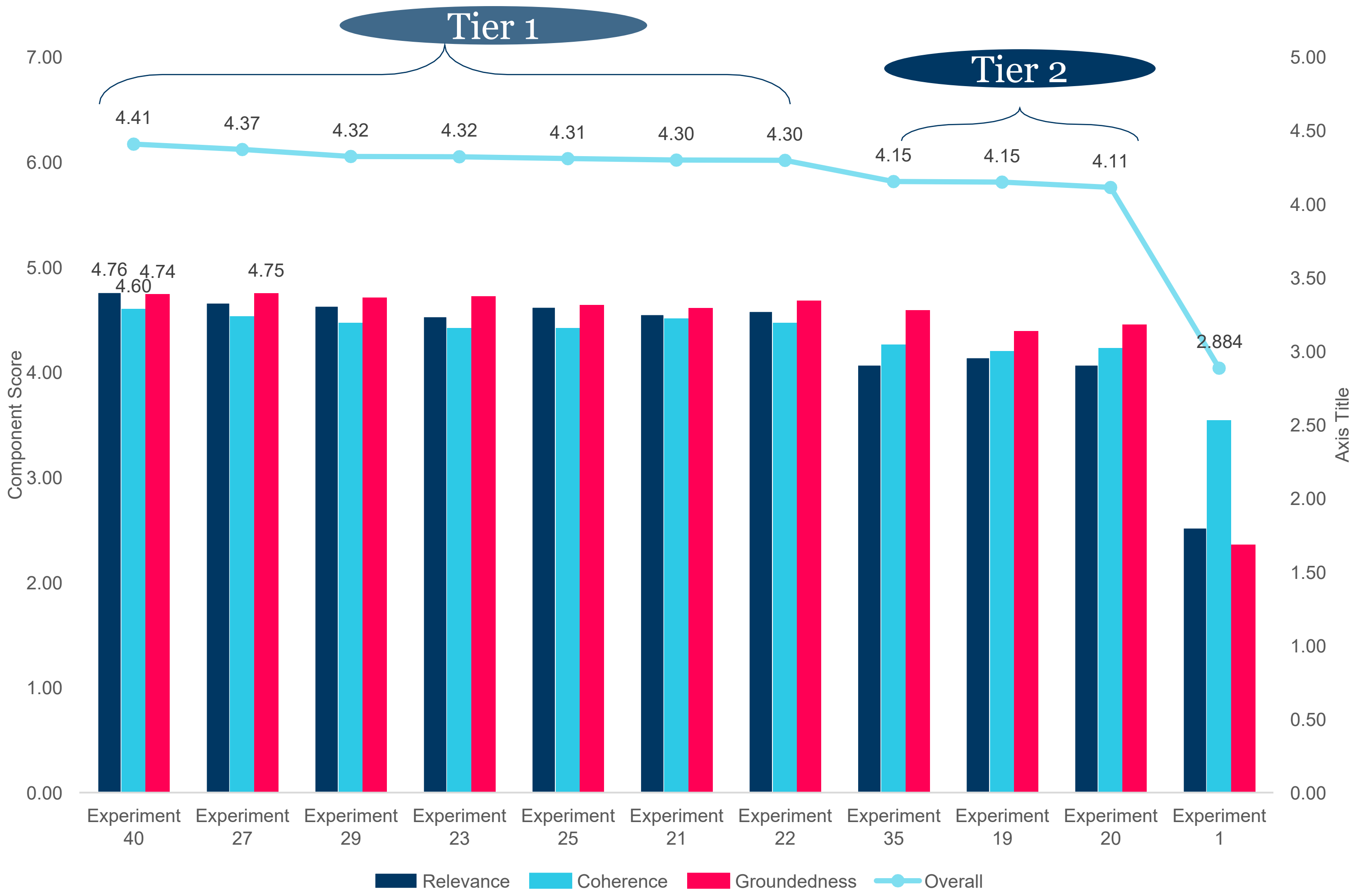
2) Site-level summary

3) Document drill-down down to a site and see the details of its documents

Document clustering – Process



RAG evaluation – Quantifying improvement



Tier 1 model parameters

Exp 40	Paragraph	1-Shot	Config 3	0
Exp 27	Paragraph	1-Shot	Config 3	0.25
Exp 29	Paragraph	1-Shot	Config 3	0.5
Exp 23	Paragraph	3-Shot	Config 3	0
Exp 25	Paragraph	3-Shot	Config 3	0.25
Exp 21	Paragraph	3-Shot	Config 3	0.5
Exp 22	200 Characters	3-Shot	Config 3	0.5

Tier 2 model parameters

Exp 35	200 Characters	3-Shot	Config 3	0
Exp 19	200 Characters	3-Shot	Config 3	0.25
Exp 20	Paragraph	3-Shot	Config 2	0.5

Baseline vs Best

Component	**Experiment 1**	**Experiment 40**	**% incr.**
Relevance	2.52	4.76	***+89%***
Coherence	3.54	4.6	***+30%***
Groundedness	2.36	4.74	***+101%***
GPT-Similarity	2.28	3.16	***+39%***
Overall	2.88	4.41	***+53%***

Case study: Predictive modeling

Predicting Site Closure Timelines for Thousands of Sites with Machine Learning

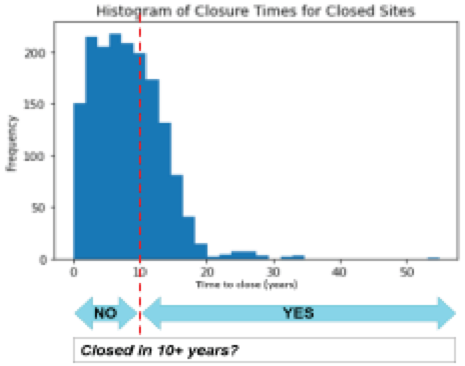
Sky Chrastina
Jonathan Eller, Ph.D.
- Group #B5

Introduction
Contaminated site remediation is costly, invasive, time-intensive and has an environmental footprint all its own. Understanding the scope and timeline of needed activities is crucial for effective remediation with financial sustainability. What if we could know a site's closing date when it opens, which sites are most at risk of overrunning this estimate, and what features about the site contribute to this prediction? We have continued to develop artificial intelligence and machine learning (AI/ML) approaches to quantifying these ideas, and present our progress in this journey

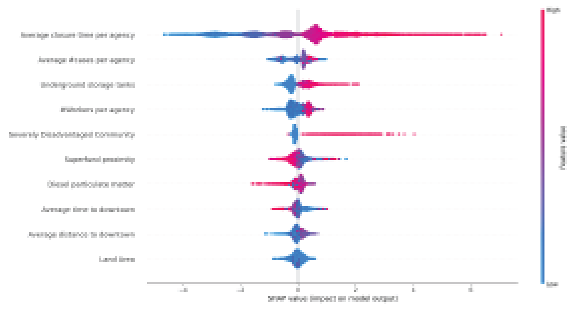
Methodology
Contaminated site remediation is costly, invasive, time-intensive and has an environmental footprint all its own. Understanding the scope and timeline of needed activities is crucial for effective remediation with financial sustainability. What if we could know a site's closing date when it opens, which sites are most at risk of overrunning this estimate, and what features about the site contribute to this prediction? We have continued to develop artificial intelligence and machine learning (AI/ML) approaches to quantifying these ideas, and present our progress in this journey.

Binary Classification
We chose a subset of sites that met data quality conditions and developed ~800 features to describe them. We developed a model to classify site closure against the 80th percentile (10 years). Year 0 predictions achieved 84% accuracy, showing that sufficient knowledge exists at site inception to create a reliable risk flag for outlier site lifetimes.

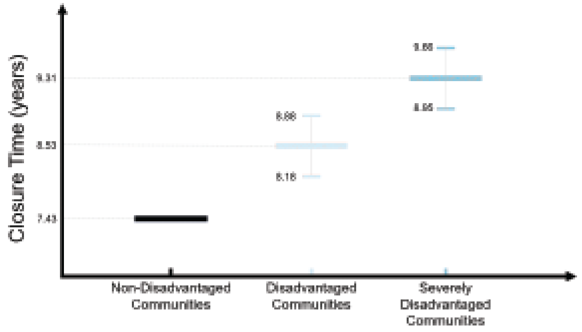
Top Features
Size* of Site
Did site open after 2005
Is site in disadvantaged community
Potential media of concern
Saturated Hydraulic Conductivity
Rating for Paths and Trails
Percent Unemployment
Groundwater Threats
Toxic Release Rating
Framland Classification



Regression



Predicted the exact closure timeline, in decimal years, for closed sites. We used Shapley Additive exPlanations (SHAP) analysis to analyze the most important features. High magnitudes show important features, direction shows the sign of the effect, and the colors indicate positive/ inverse correlation. Sites being located in an economically disadvantaged community had surprisingly strong predictive power.

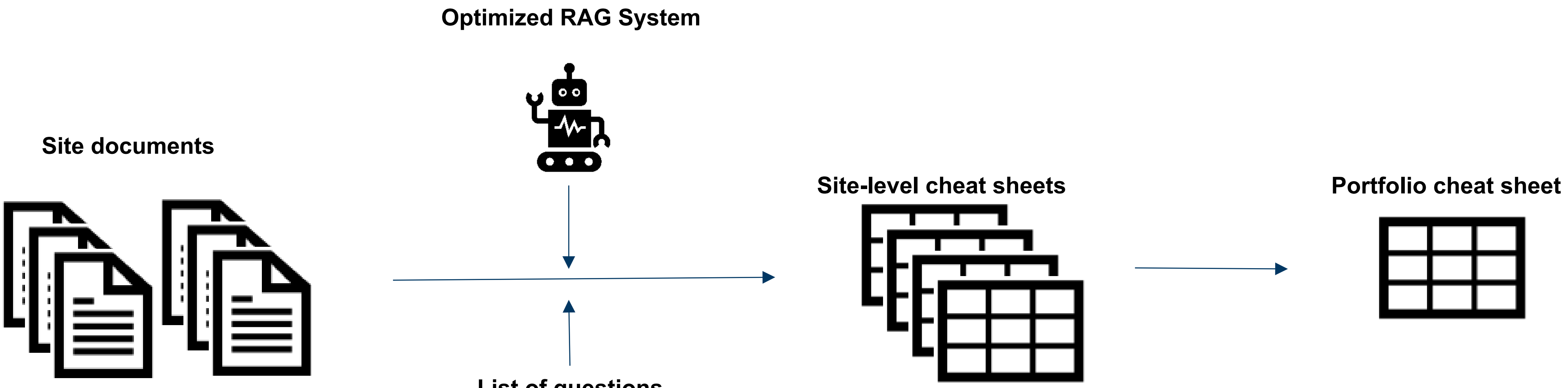


Further Opportunities
- Publication of these findings
- Explore causal explanations for important predictors
- Expand states – dozens of states have contamination – databases with rich information and documents
- Model improvements: Multiclass classification, clustering, continued feature engineering and data augmentation
- Expand sites – include sites with missing data
- Expand features – account for more geospatial, geological, financial, and bureaucratic aspects

Conclusion
Predicting site closures is a complext problem; a data-driven solution thus requires a complex dataset and must include both the scientific, legal, demographic and political contexts. By accounting for all these realities, we can predict how long a site will take to close, very early in its life cycle. These learnings can be utilized by responsible parties, regulators, and other environmental professionals to improve decision-making in a variety of ways.

Case study: Augmenting portfolio knowledge

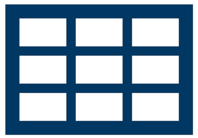
Used optimized RAG system as a back-end engine to build site- and portfolio-level summaries based on an expert-supplied list of questions



Site documents

- Routine Monitoring Reports
- Feasibility Studies
- Closure Requests
- Agency Opinion Letters
- etc…

List of questions



Examples:

- Site location, COC List, depth to water, # wells, etc…
- Regulatory framework
- Presence/Absence of LNAPL
- Remediation system/history