Los Alamos NATIONAL LABORATORY | Information Science & Technology CAPABILITY REVIEW

GRAND UNIFIED FILE INDEX (GUFI)

A Global Meta-Base NOT a Lake House

Gary Grider, ALDSCT
Senior Director for Computing Technology

Brian Atkinson, HPC-Design
HPC Engineer

Amanda Bonnie, HPC-3PO
Science/Engineering Project Leader

May 20, 2026

LA-UR-26-24739

GRAND UNIFIED FILE INDEX (GUFI)

A Global Meta-Base NOT a Lake House

Extremely Scalabe

10-100 Billion items; Over billions of folders

Index Nearly Any Storage System

NFS; localFS; Lustre; HPSS; S3; cloud; win shares;

Crazy Fast Query

Single seconds for users; Single minutes for admins

Access Control Derived from Existing Permissions

Reuse of what already exists; ACL/MLS; POSIX

Multi-Site Access/Deployment

Multiple Different FS; Query betweem multiple locations

Support Many Access Methods via SQL

SQL→ Python; C/C++, ruby, web, MCP, cmd, beyond…

Open Source; Commerical; No Need to Move Data; Award Winning

Available on Git; Embedded in 7 Products; 2 Patents; R&D100 Awarded

2018 R&D 100 WINNER

HPE

HAMMERSPACE

BeeGFS®

Domino

GUFI – Current Deployment

Base GUFI –

- Enable search over file metadata one has access to.
- Ability to sort/limit by metadata feature
  - Size, location, type, etc.
- Full index currently updated **WEEKLY**

Currently deployed and available to users on: *NFS, Lustre, HPSS, Campaign*

GUFI – Current Development

Text & Feature Extraction –

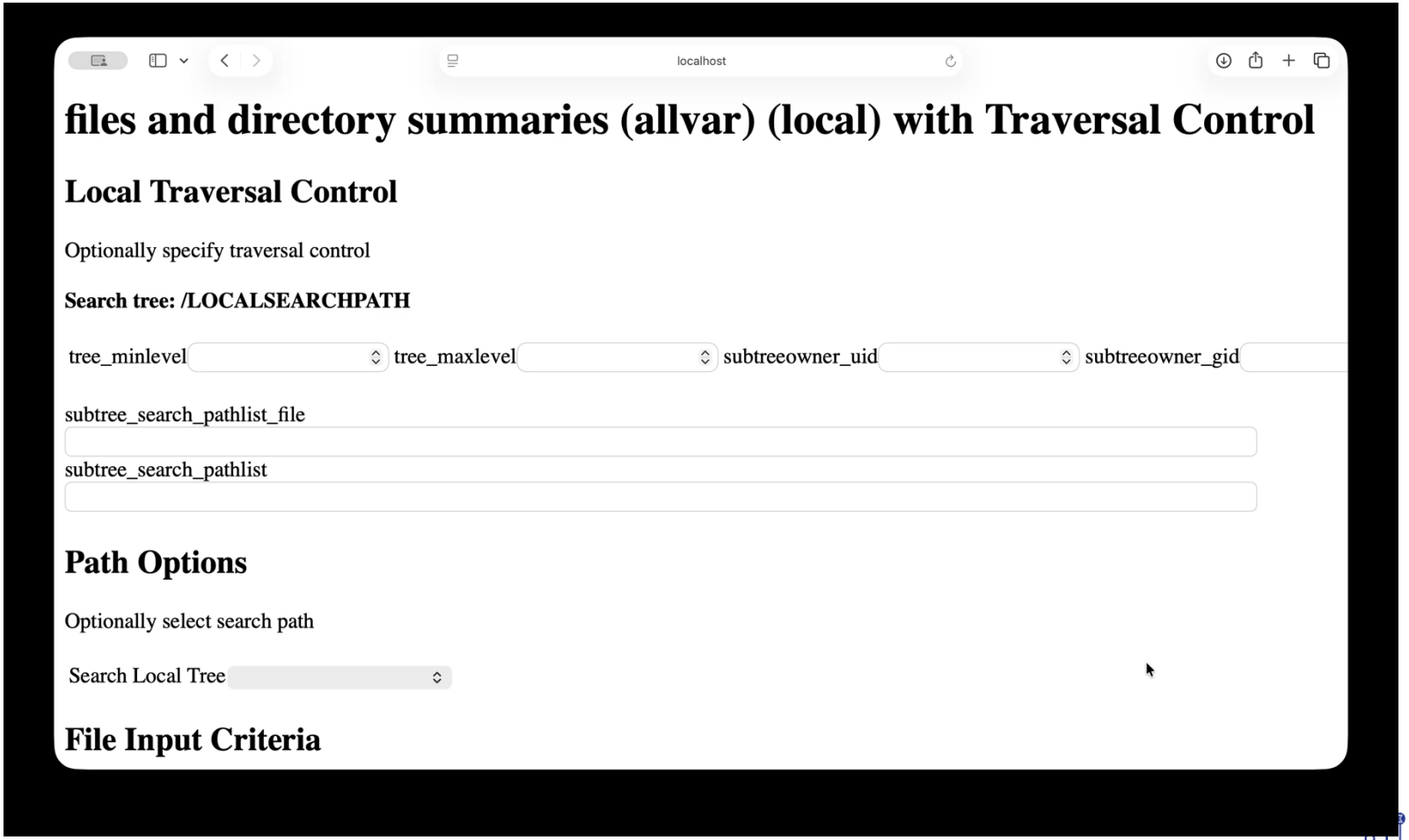
- Enables search over files metatdata **AND** FILE CONTENT one has access to.
- PDF; Microsoft Products; Images(jpg,png); txt/data; code(.py, .c, .cpp) and more.
- Image content is described
  - "A man standing in front of a building"
- Base index is indexed **NIGHTLY**

Vector –

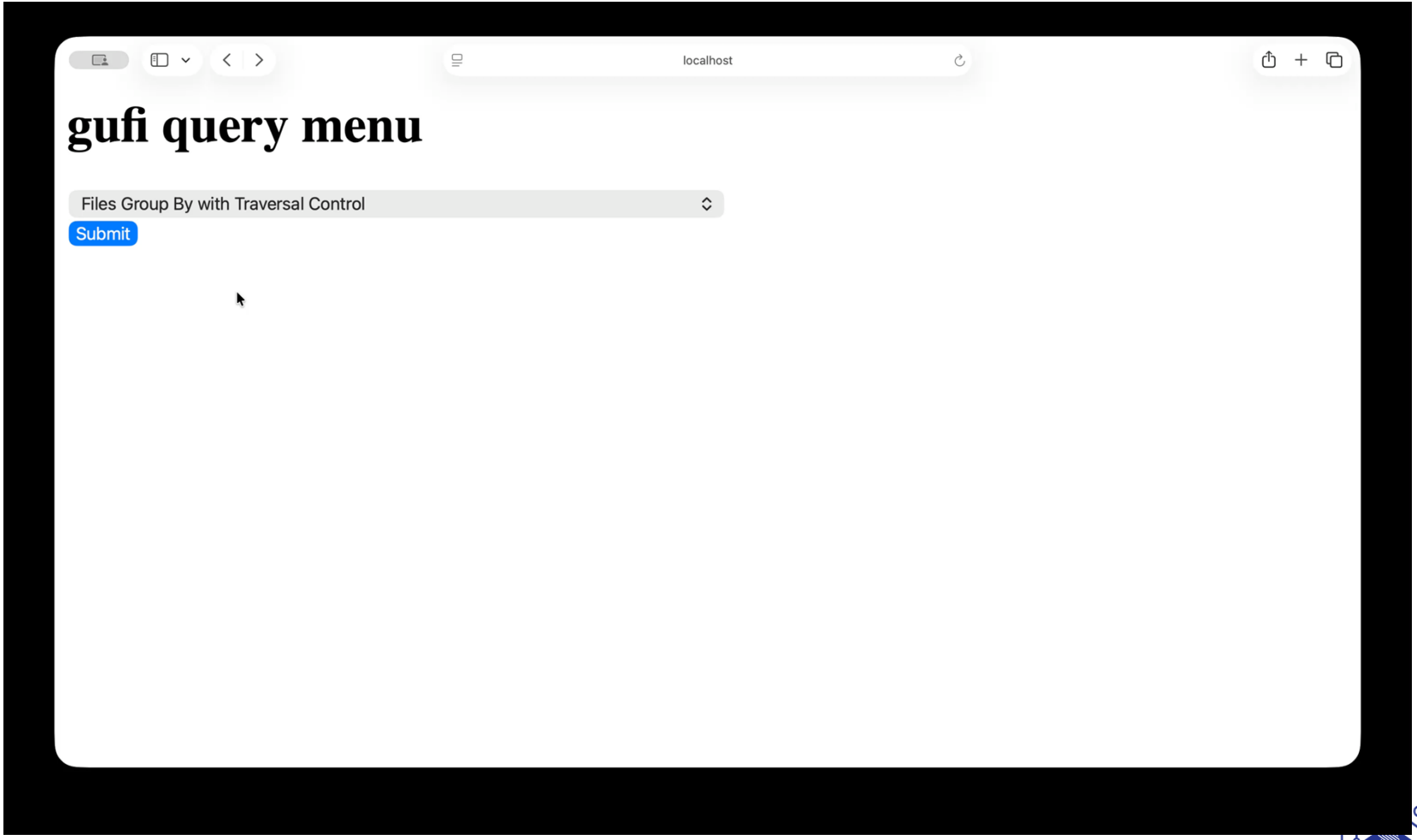
- Creating vectors of file metadata and file contents
- Ultimately enabling data for more than human consumption

DEMOS →

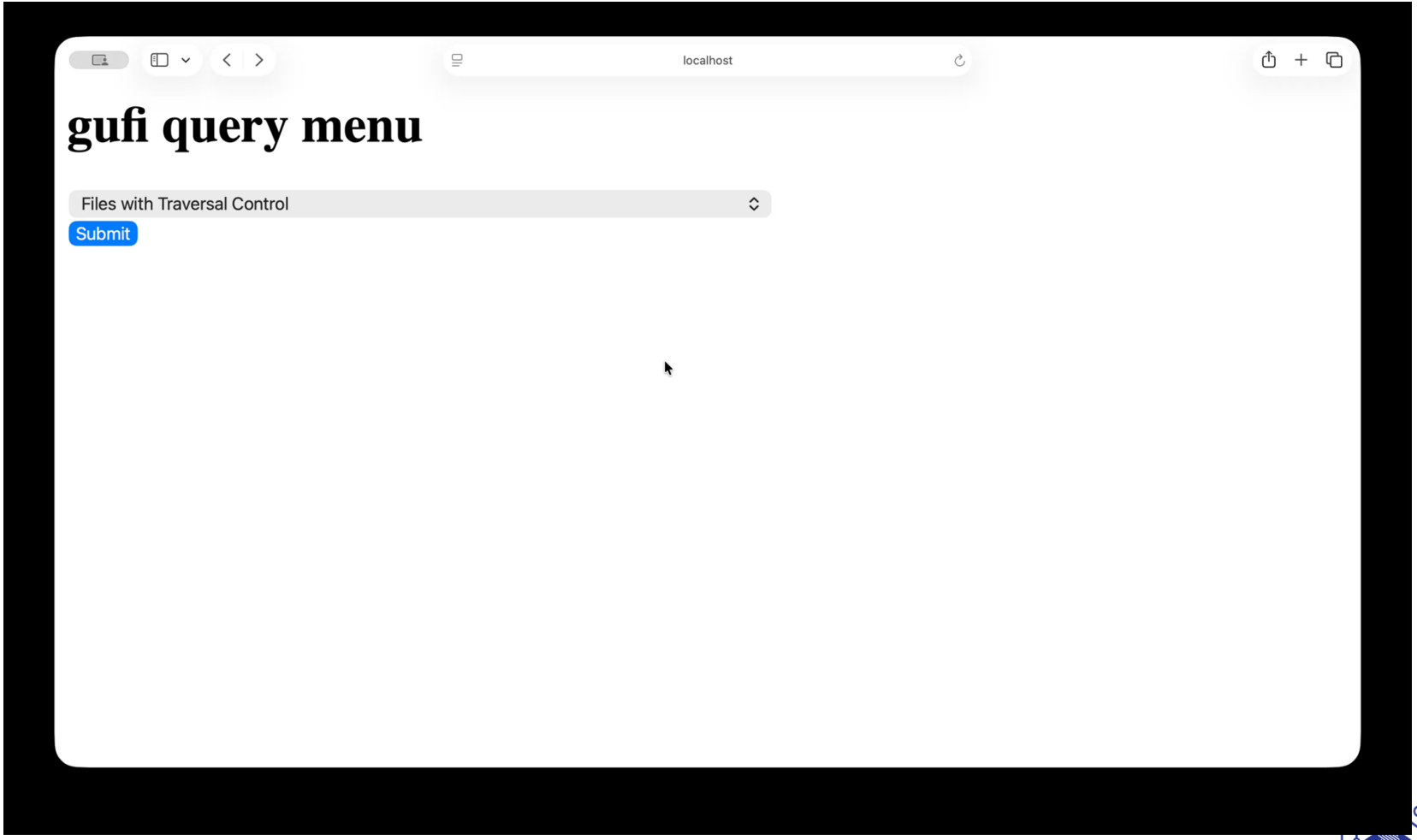
Base + Text & Feature Extraction (All Options)



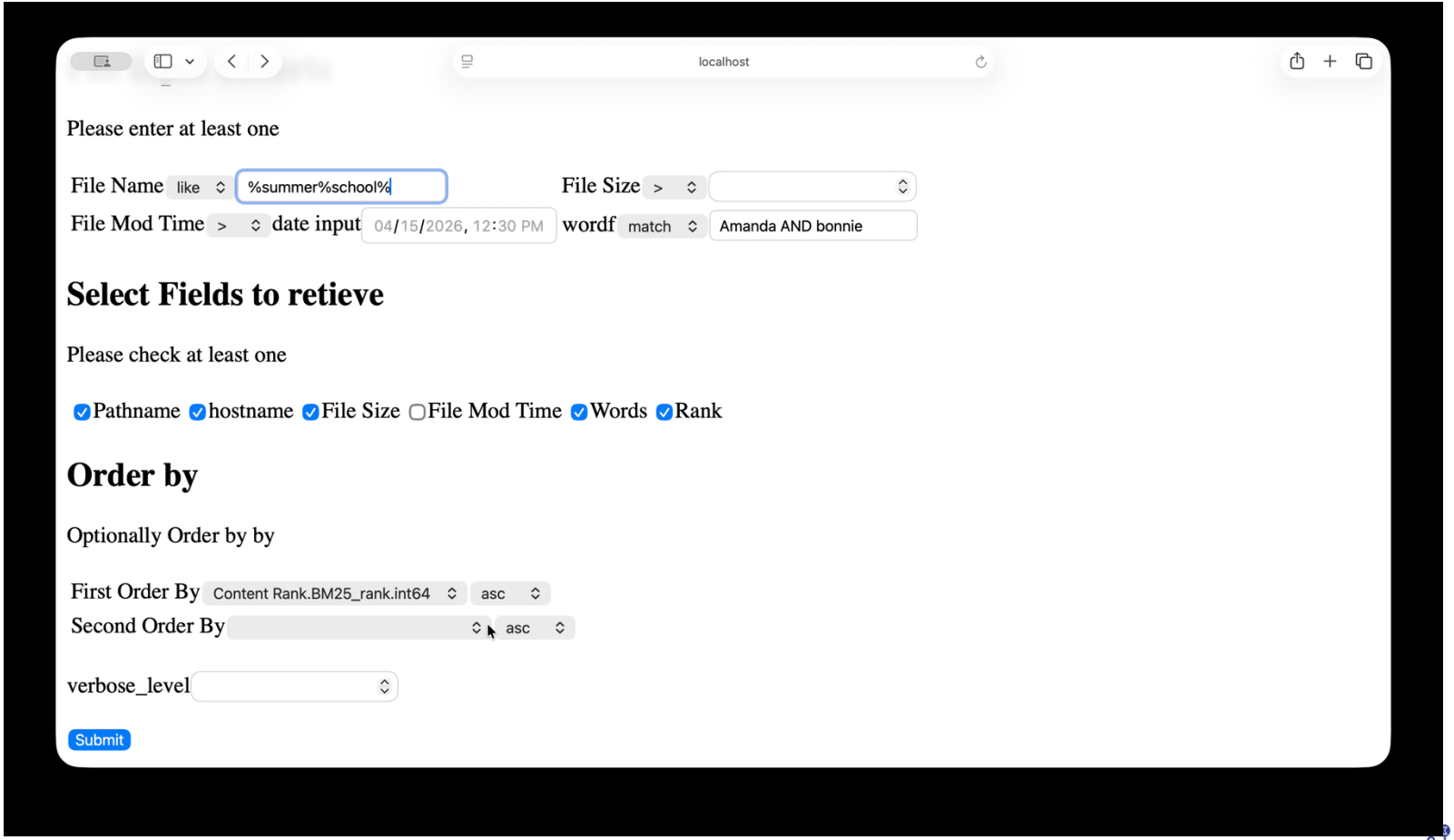
GUFI Query By Group



GUFI Query Over Two Sites

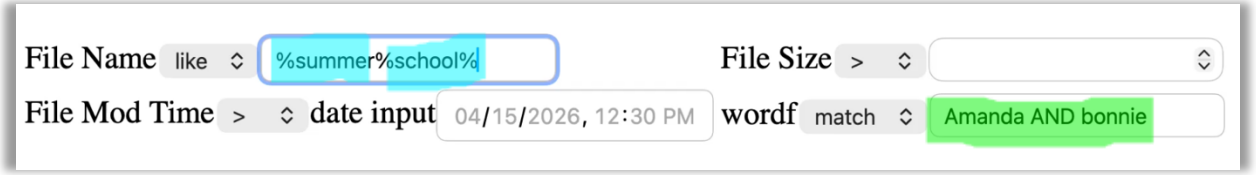


Base + Text & Feature Extractin (Search Demo)



Base + Text & Feature Extraction (Closer Look)

Test System with 70k Files; 7k Directories



pathname	hostname	size	wordf	rank
/search/historyidx/Z-GaryGriderHistory/older-presentations/106273ppt/CCN-9/ISTI/Cluster-Network-Summer-School-FY16/ISTI-Cluster-Network-Summer-SchoolHPC Overview-2016.pptx	regufi.lanl.gov	21639992	...isti summer school director -->amanda<-- -->bonnie<-- nmc management administration andree...	-0.794096830488506
/search/historyidx/Z-GaryGriderHistory/recent-presentations/ISTI-Cluster-Network-Summer-SchoolHPC Overview-2016.pptx	regufi.lanl.gov	21649859	...isti summer school director -->amanda<-- -->bonnie<-- nmc management administration andree...	-4.89971868308536

total rows 2

Base + Text & Feature Extraction (more)

Looking at the database content we list:
File Name, Asset Type, Extracted File Metadata and Text

Note the “derived” and having one or more records for a given file

```
Roadrunner System Overview (LA-UR-08-2087).ppt|derived|0|{"title":
"Roadrunner Project Preview", "author": "ESD  Registered User",
"last_modified_b|
|Document Metadata for Roadrunner System Overview (LA-UR-08-
2087).ppt Title: Roadrunner Project Previ
Roadrunner System Overview (LA-UR-08-2087).ppt|derived|1|{"title":
"Roadrunner Project Preview", "author": "ESD  Registered User",
"last_modified_b|
```

Base + Text & Feature Extraction (more)

Notice how it pulled specific Microsoft metadata from the file

```
ggrider_hpcwire.pptx|derived|0|{"title": "Godfather or Parallel
File Systems", "author": "Grider, Gary", "keywords": "hpc|
|Document Metadata for ggrider_hpcwire.pptx Title: Godfather or
Parallel File Systems Author: Grider,
```

This was a picture with no text or the model didn't have anything to say about it (the actual image was a small logo)

```
ggrider_hpcwire.pptx|derived|1|{"title": "Godfather or Parallel
File Systems", "author": "Grider, Gary", "keywords": "hpc|
|Image /page/1/Picture/0
```

Base + Text & Feature Extraction (more)

It found a picture on page 2 and performed OCR on it

```
ggrider_hpcwire.pptx|derived|2|{"title": "Godfather or Parallel
File Systems", "author": "Grider, Gary", "keywords": "hpc|
|35 HPC Legends Gary Grider Gary Grider Godfather of Parallel
File Systems Image /page/2/Picture/3
```

This is just extracted text

```
ggrider_hpcwire.pptx|derived|3|{"title": "Godfather or Parallel
File Systems", "author": "Grider, Gary", "keywords": "hpc|
|Gary Grider is Senior Director for Computing Technologies at
Los Alamos National Laboratory where he
```

Base + Text & Feature Extraction (more)

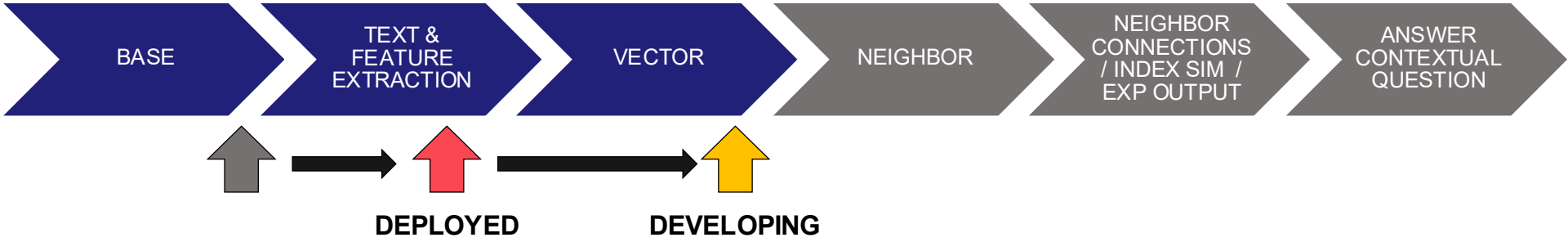
- It is not JUST text files; it is scraping text from all different types of files
  - All Microsoft file types; PDF; txt; image types (.jpg, .png); most code types (.c, .cpp, .py); archives (.tar, .zip); and many more
- All files are torn into parts and chunks and extracted
- On images, it will OCR and ask a model to summarize what is in the picture
  - EX: “Picture of two people standing a podium…”
- Archives, like zip, are unizipped and each file analyzed (even a zip inside a zip!)
- All metadata is traceable back to the source file
- Very extensible so anything that can produce text
  - Image analysis, models, hdf5dumps…
- Even working on array distribution characterizing and changes in array distribution over time

+ Vector (Timing a RAG Query)

+ Vector (RAG Query Closer Look)

```
ggrider@pn2201328 GUFI-MAC-Indexer %
ggrider@pn2201328 GUFI-MAC-Indexer % time ./run_history_vec_conf 'andy white' 'andy AND white' 'who was andy white' '' 2>/dev/null

**** text from gufi using input keyword (full text search) and vector search nearest neighbor to question provided *********

******** top N files ******
file: /Users/ggrider/vaultiq/historyidx/Z-GaryGriderHistory/pictures/evenmorepics/_MG_4025.JPG
snippet: roadrunner -->andy<-- -->white<-- anastasio scc pete dominici pp p
file: /Users/ggrider/vaultiq/historyidx/Z-GaryGriderHistory/pictures/evenmorepics/_MG_4023.JPG
snippet: pete dominici -->andy<-- -->white<-- roadrunner scc fastest earmark anastasio pp...
file: /Users/ggrider/vaultiq/historyidx/Z-GaryGriderHistory/older-presentations/106273ppt/HPC-DO/lanl_computing.pptx
snippet: ...role -->andy<-- -->white<-- marie lead lufd kurt soo -->andy<-- duncan...

**** context *********
roadrunner andy white anastasio scc pete dominici pp p
 pete dominici andy white roadrunner scc fastest earmark anastasio pp p
 computing lanl vision future april goal computing recognized leader high performance computing especially national security scienc
e forefront computer science key elements laboratory computing program scope asc institutional computing sci computing key signatur
es success visible integrated leadership compelling documented strategy robust institutional investment xcommitment including facil
ity plan scope asc computing successful delivery trinity beyond excellent support mission stable reliable computational infrastruct
ure capability capacity machines responsive customer service leadership middleware computing paradigms effective partner llnl snl e
specially aces asc effective exascale partnership scope institutional computing sustained commitment multi year spend plan capacity
 acquisition best class ability explore novel architectures test beds data intensive virtual clusters including partners xoff site
strategic allocation resources tie winning proposals transparent reliable service users responsive trends bi

*************

question: who was andy white
NO CONTEXT


WITH GUFI RAG CONTEXT
 and what was his role in the development of the fastest computer in the world?
Answer:andy white was the chief scientist of the supercomputer center (scc) at the national laboratory (lanl) in new mexico. He was
 responsible for the design and development of the world's fastest computer, which was christened the roadrunner. The roadrunner wa
s a 100-petaflop supercomputer that was built in 1998. It was the first supercomputer to achieve a petaflop performance, which is a
 thousand trillion floating-point operations per second. White was also the director of the supercomputing center at the llnl.
./run_history_vec_conf 'andy white' 'andy AND white' 'who was andy white' ''   8.29s user 1.97s system 115% cpu 8.865 total
ggrider@pn2201328 GUFI-MAC-Indexer %
```

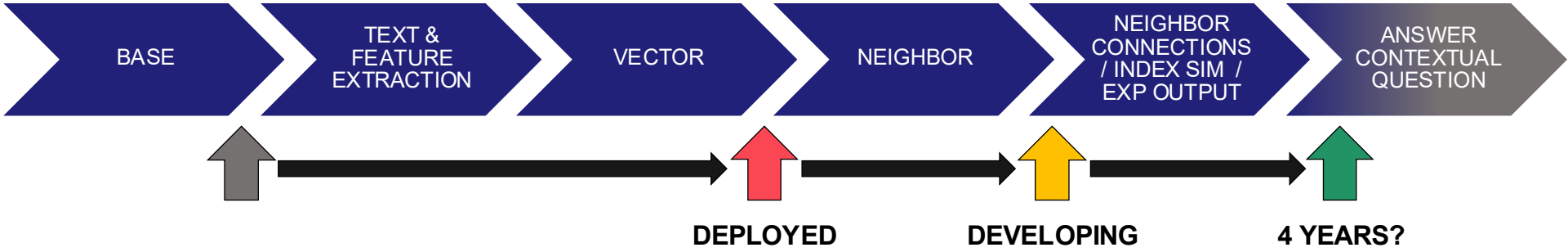
GUFI - Future



Continued Development with Upgrades Along the Way

- Base Extraction
  - Weekly (now) → Nightly (6-9 months) → On Demand (12-14 months)
- Text & Feature Extraction
  - Specific File Types (6-9 months) → Specific Users/Projects (12 months)
  - More Feature Types: movies, audio, etc. (10-12 months)
- Vector
  - Basic RAG (9-12 months) → Model Context Protocol (MCP) (12-16 months)

GUFI – Beyond…

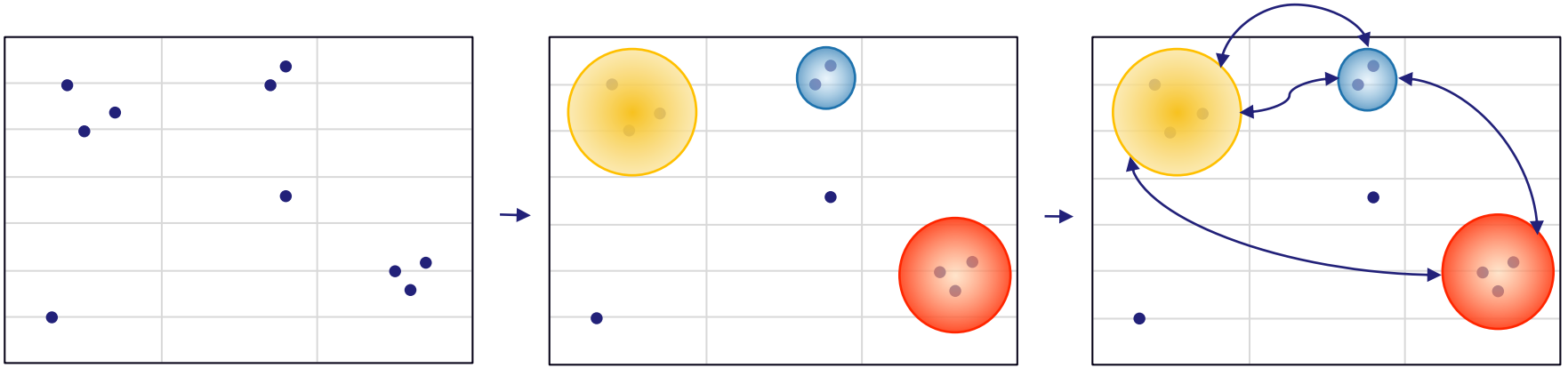


Our first Major Target is to have production deployment with vectors to provide data for more than human consumption.

But can we have more? Can we answer contextual questions with GUFI?

GUFI – Next Steps

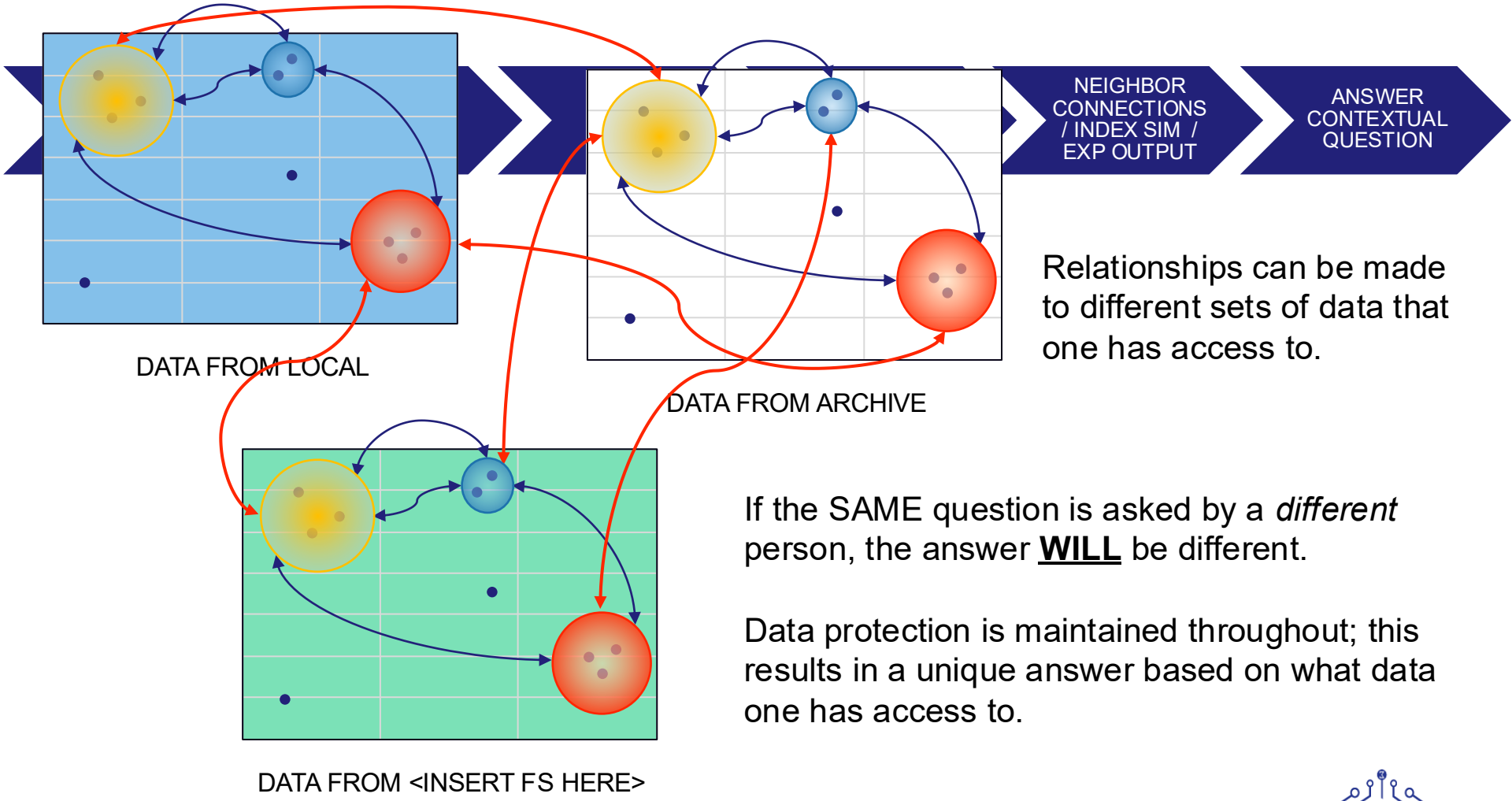
BASE → TEXT & FEATURE EXTRACTION → VECTOR → NEIGHBOR → NEIGHBOR CONNECTIONS / INDEX SIM / EXP OUTPUT → ANSWER CONTEXTUAL QUESTION



GET VECTORS

FIND NEAREST NEIGHBORS & CLASSIFY NEIGHBORS

FIND NEIGHBOR CONNECTIONS



Relationships can be made to different sets of data that one has access to.

If the SAME question is asked by a *different* person, the answer **WILL** be different.

Data protection is maintained throughout; this results in a unique answer based on what data one has access to.

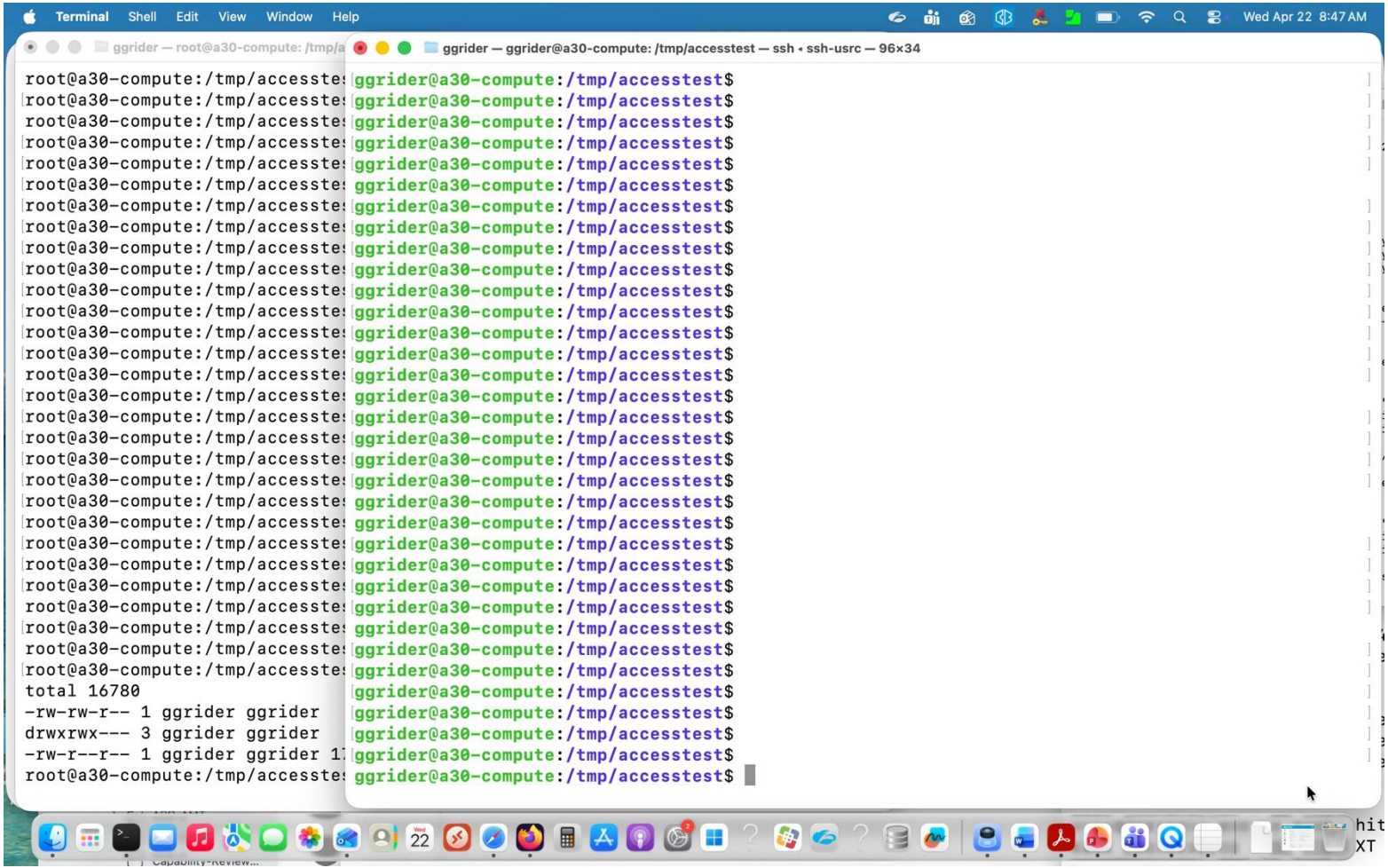
Providing a Narrowing Route to the GOLD

THE GOLD!!

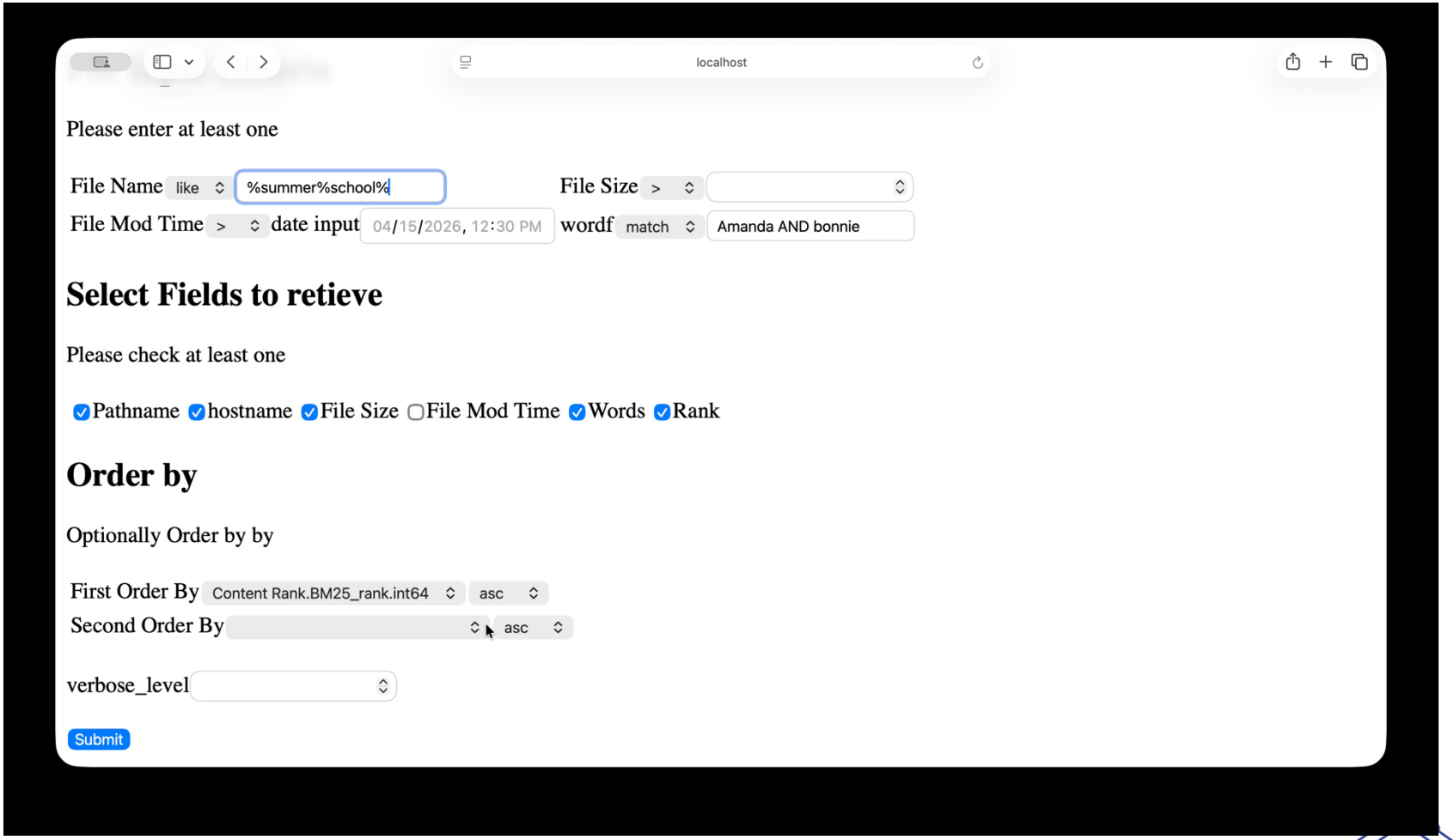
Q&A + Additional Demos

- Query by Group
- Query Over Two Sites
- Simple Acess Control
- Base + Full Text Query
- Domino AI Demo

Simple Acess Control



GUFI Base + Full Text



Domino AI Demo