Expertise Identification with Data Analytics

SAND2016-3470C

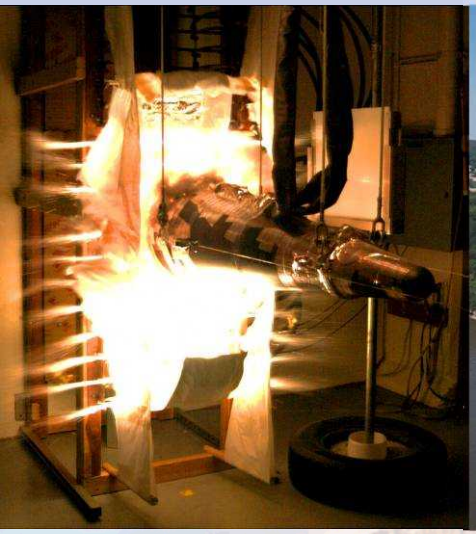
Presented by Dann Barnes and Gary Huang
Knowledge Systems Organization
Sandia National Laboratories

SAND2016-NNNN

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Four Mission Areas

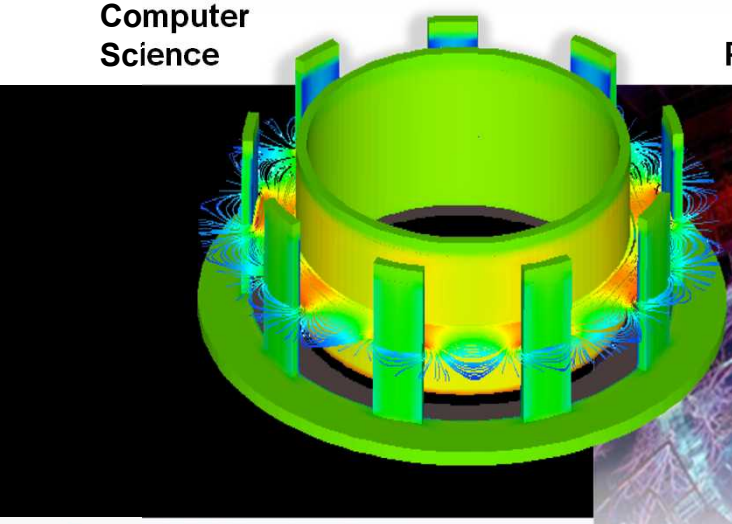
- Nuclear Weapons
- Defense Systems and Assessments
- Energy, Climate, and Infrastructure Security
- International, Homeland, & Nuclear Security



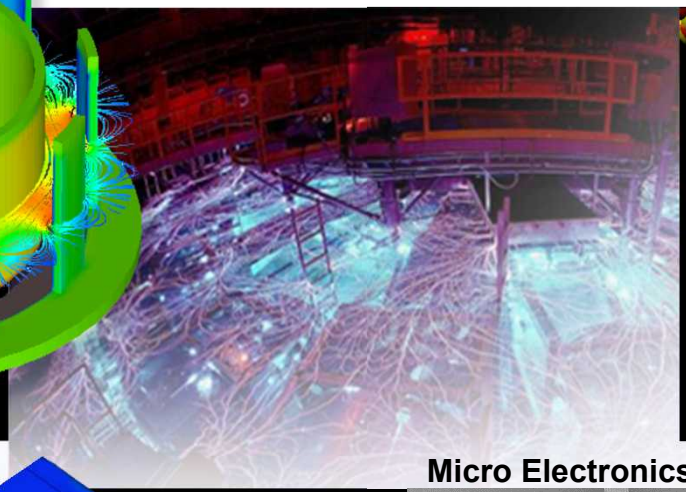
Enabled by Strong Science and Engineering

Research Disciplines

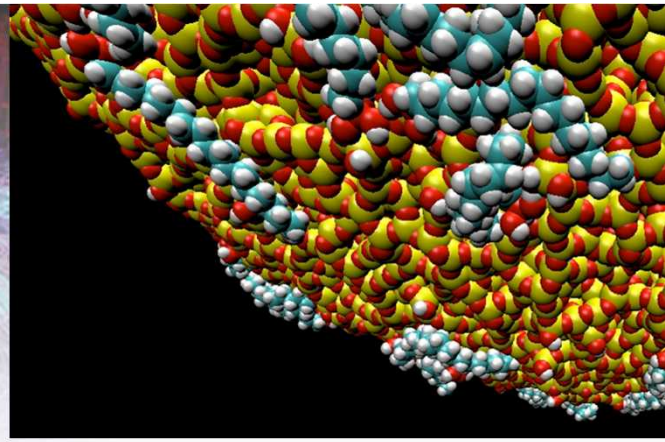
Computer Science



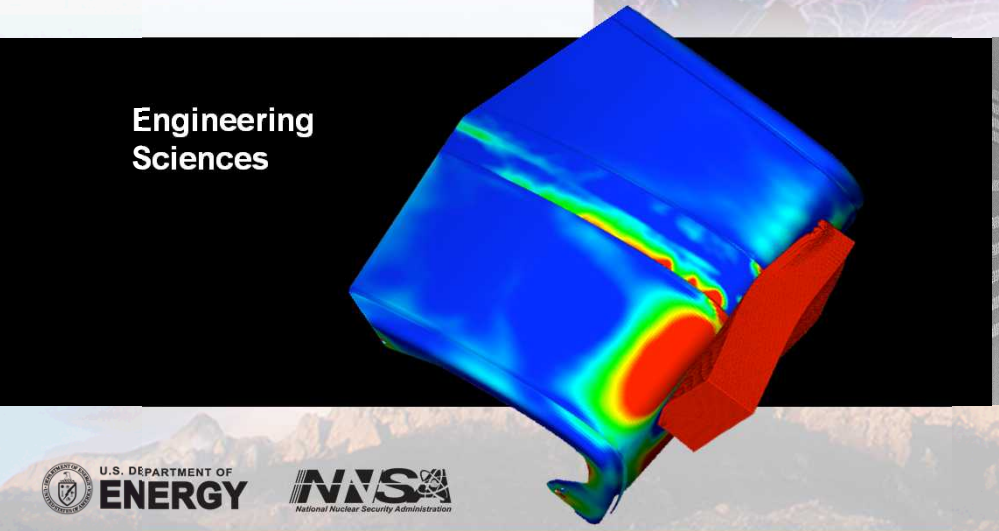
Pulsed Power



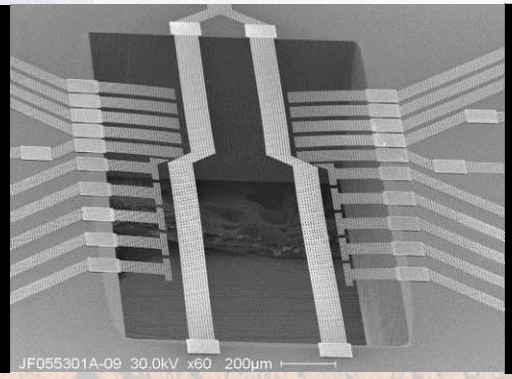
Materials



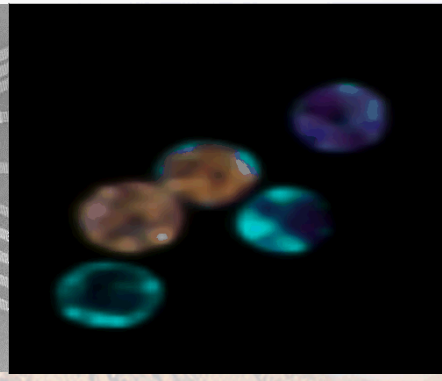
Engineering Sciences



Micro Electronics



Bioscience



Knowledge Systems Department

- Enterprise level technical solutions for information analytics and search.
- Develop and incorporate advanced techniques in content analytics and search into our information systems to improve usefulness of information and to improve the ability of the workforce to find the information they need to perform their jobs.

Expertise Identification Tool Drivers

An expertise identification tool enables Sandia and its workforce to more effectively, efficiently, and accurately network and respond to questions related to our expertise.

- We can better respond to external requests regarding our capabilities
- It will enable identification of internal collaboration partners and reduce duplication of effort.
  - Who can do it or help me do it?
  - Has this been done before?
  - How do I find experts (individuals or organizations) that I might collaborate with?
- When somebody leaves, what knowledge and skills are they taking with them? Do we need to fill in? Who else at the lab has this expertise?
- How have Sandia's areas of expertise changed over time?

Expertise Identification Project

The Problem

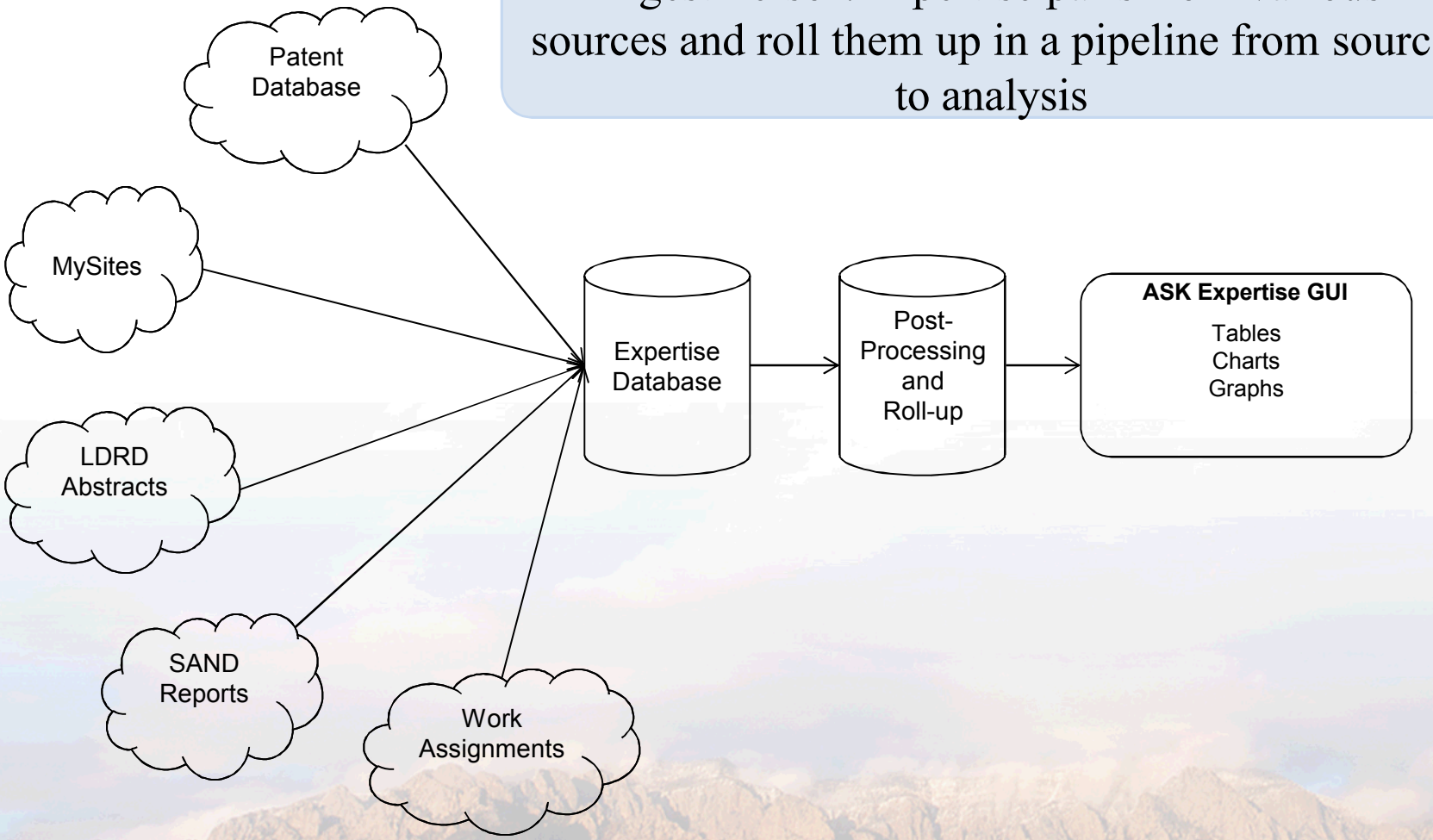
- **Previous attempts to identify expertise have failed because they required staff to manually enter their own information about knowledge, skills, and abilities. Many staff will not enter their information and, if they do, it quickly becomes out of date.**

A Solution

- **Extract expertise from multiple, existing, free-text information sources and present results in an interactive display.**
- **A key advantage of this approach is that it is self-maintaining. It uses information already in our environment, and it is kept current through normal work process.**

"Big Picture"

Ingest Person/Expertise pairs from various sources and roll them up in a pipeline from source to analysis



Sources of information

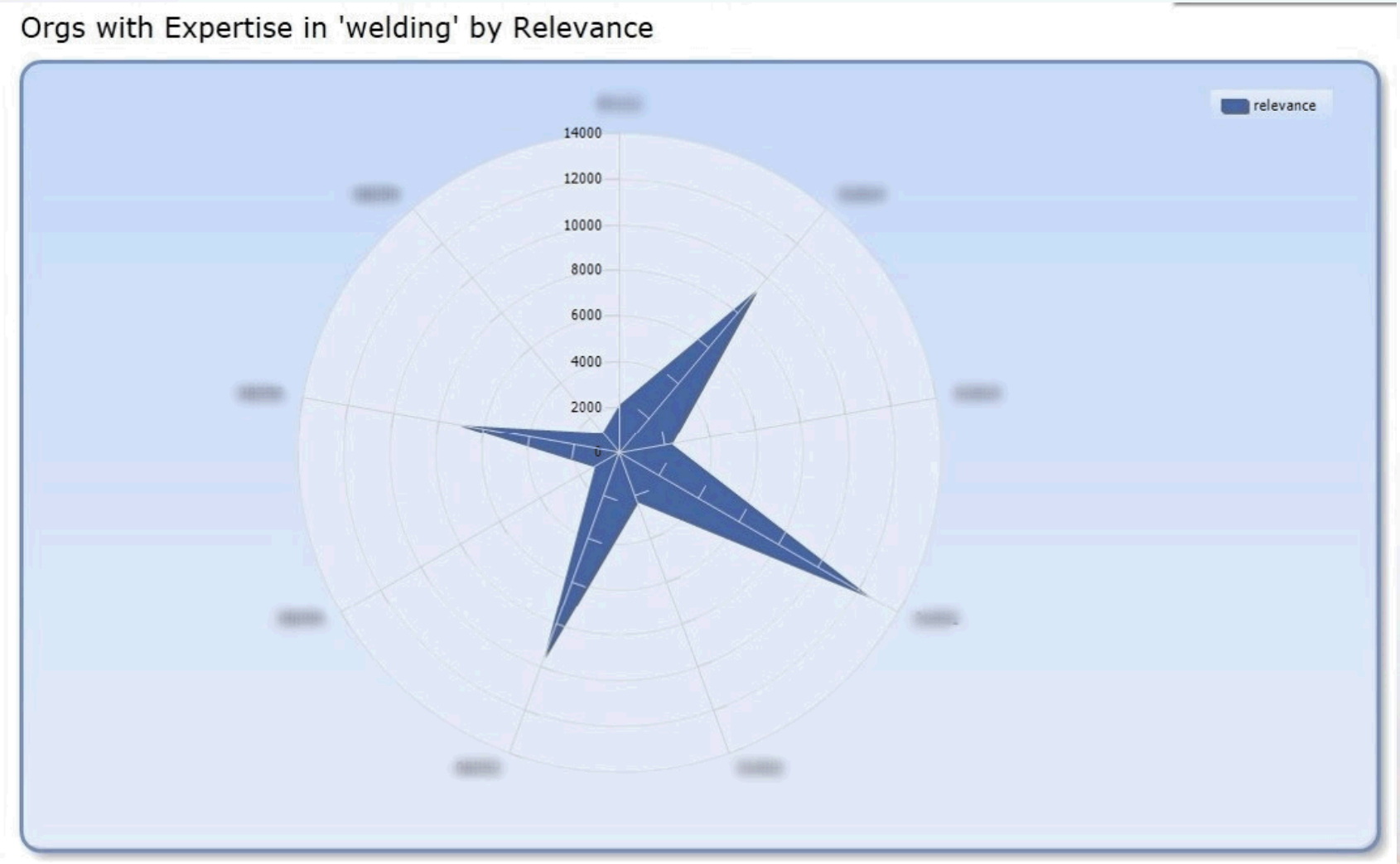
Currently using

- **Published reports (SAND)**
- **Microsoft SharePoint MySite entries**
- **Laboratory Directed Research and Development (LDRD) abstracts**
- **Patent Database**
- **Work Assignments**

Other potential sources of information

- Department names
- Job titles
- Web pages and SharePoint sites
- Book and Report requests
- Degrees
- Training course completions
- Conference attendance
- Resumes
- Email messages and distribution lists

Included a general search capability



A deeper look at a particular organization

Employees with Expertise in 'welding'

Org	Employee	Status	Relevance
[redacted] - Comp Materials & Data Science		Active	3,777
		Active	3,654
		Inactive	1,345
		Inactive	488
Total			9,264

Documents Related to 'welding'

Employee	Status	Source	Doc ID	Title	Relevance
	Active	SAND	2012-7614	Coupling 3D Quantitative Interrogation of Weld Microstructure with 3D Models of Mechanical Response	1,331
			2012-4467	3D Characterization-Aided Modeling of Weld Deformation of 304L Stainless Steel	1,259
			2012-7672	Weld Porosity Characterization in Three-Dimensions within 304L Stainless Steel	1,187
	Total				3,777

Clicking the document ID brings up the document

Coupling 3D Quantitative Interrogation of Weld Microstructure with 3D Models of Mechanical Response

SAND2012-7614A

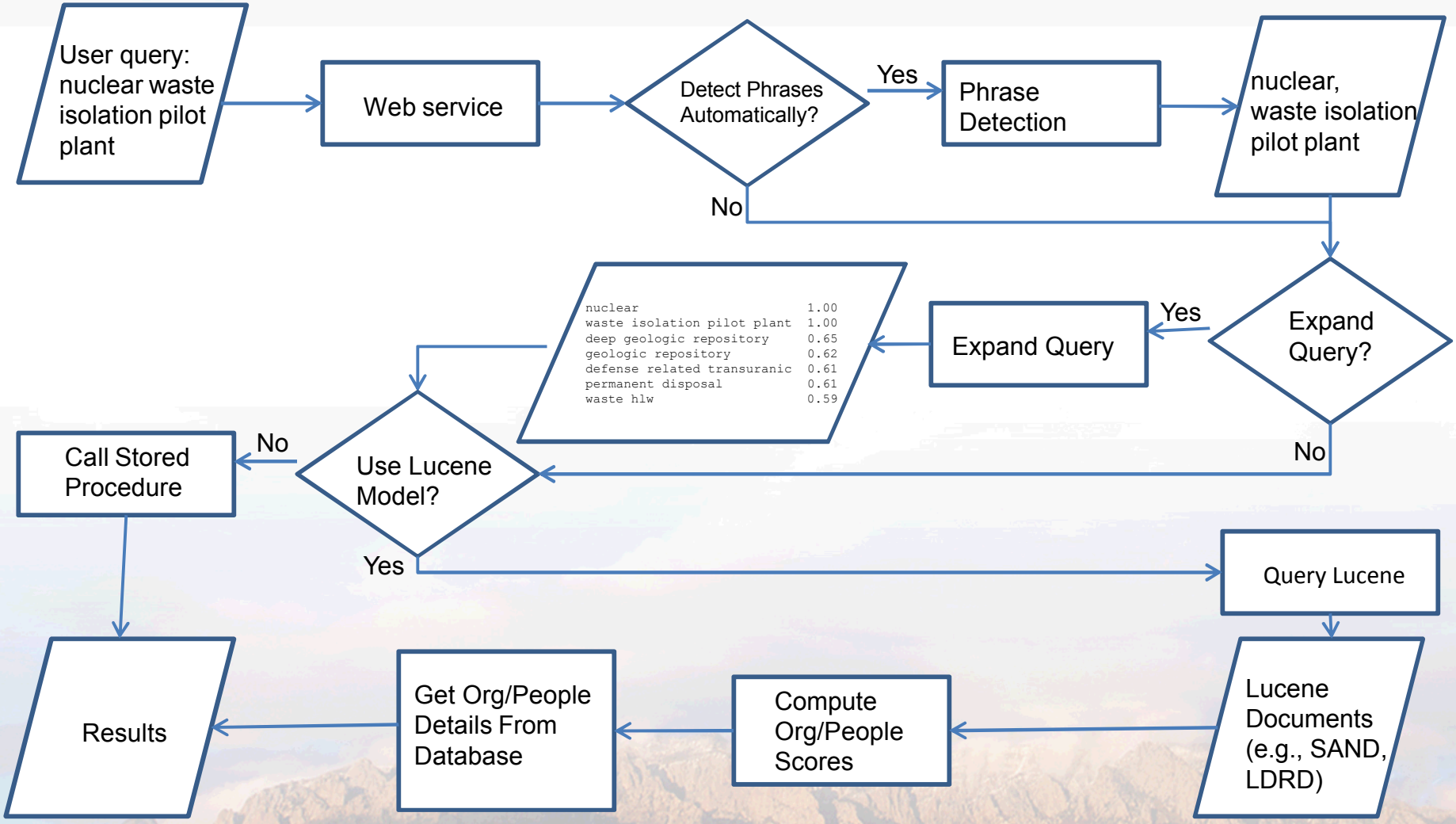
Sandia National Laboratories
PO Box 5800, Albuquerque, NM 87185-0346

In this study, laser welds of 304L stainless steel machined using a continuous wave Nd:YAG laser under two separate focusing lenses, under six sets of delivered power (ranging from 200 – 1200W) and across 5 separate speeds (10 – 80 in/min) are examined using non-destructive means of micro-computed tomography (iCT). Quantitative characterization of the size, shape, frequency and directionality of the voids show marked trends influenced highly by their welding history. These fully characterized, three-dimensional domains are then utilized as direct inputs for finite element analysis to better understand the varied effects of porosity on mechanical response in this highly ubiquitous material system.

ASK Expertise 1.0 and Beyond

- **The prototype Expertise Identification tool was rolled into the broader Analytics for Sandia Knowledge (ASK) project**
  - Drill down to source documents temporarily disabled until sensitivity issues have been resolved
- **Enhancements**
  - Search for expertise using SQL Server full-text index functions
    - freetext – very broad matching criteria
    - freetexttable – includes SQL Server-generated rankings
    - contains – more targeted results
    - containstable – includes rankings
  - Common look and feel of the ASK user interface
  - Include related terms in search
  - Automatic phrase identification
  - Moving to a Lucene-based search engine

Big Picture Workflow



Query Expansion

- **Use terms related to the user query to retrieve a larger set of results**
- **Using generic thesaurus/stemming ignores the nature of domain-specific data**
  - Acronyms abound at Sandia
- **Build a system to learn related terms from Sandia-specific data**

word2vec

- **Unsupervised neural network algorithm for text processing**
- **Learns to represent words as vectors, such that words appearing in similar contexts have higher cosine similarity**

Sweden →

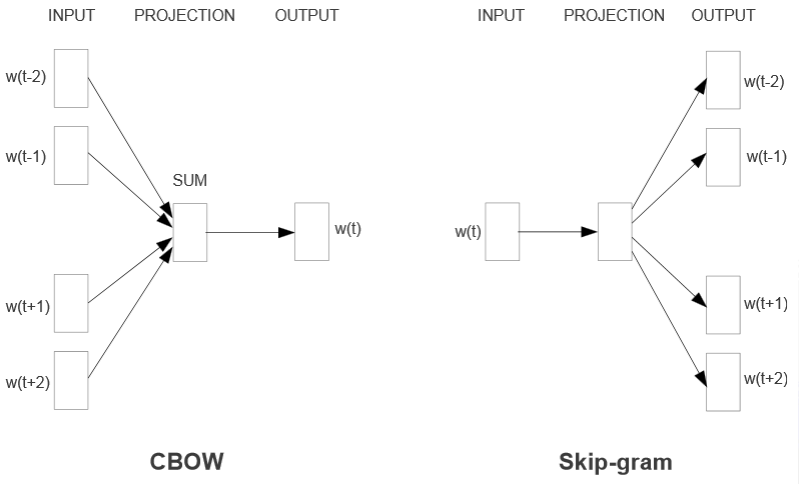
Word	Cosine distance
norway	0.760124
denmark	0.715460
finland	0.620022
switzerland	0.588132
belgium	0.585835
netherlands	0.574631
iceland	0.562368
estonia	0.547621
slovenia	0.531408

* deeplearning4j.org

- **Published by Google in 2013**

word2vec

- **Two-layer neural network**



* Mikolov et al., 2014

Query Expansion

- **Examples of related terms from trained model**

User Query	Related Terms
pki	entrust two factor authentication classified desktop
biological agents	biological warfare agents toxins bioterrorism
ICADS	sabrs gnt base pafb

Detecting Phrases in Queries

- **Users can use quotes to search for exact phrase matches**
- **Alternatively, system can automatically identify phrases in queries**
  - Helps find more meaningful related terms
  - Phrase detection is data driven, using word2phrase from Google [Mikolov et al., 2014]

User Query	Related Terms w/o Phrase Detection	Related Terms with Phrase Detection
reinforcement learning	skills alc 2015 learn	unsupervised unsupervised learning kx trees
Minuteman III	ii gps gbd block iv	gbsd strategic deterrent payload transporter

Using Apache Lucene for Ranking

- **Popular, free open-source software text search engine**
- **Allows fine-grained control**
  - Data & query processing
  - Choice of ranking models
  - Custom boosting based on document field/source
    - e.g., give more weight to matches in document title than document body

Lucene Index Layout

Each item in the Lucene index roughly corresponds to what we intuitively think of as a document:

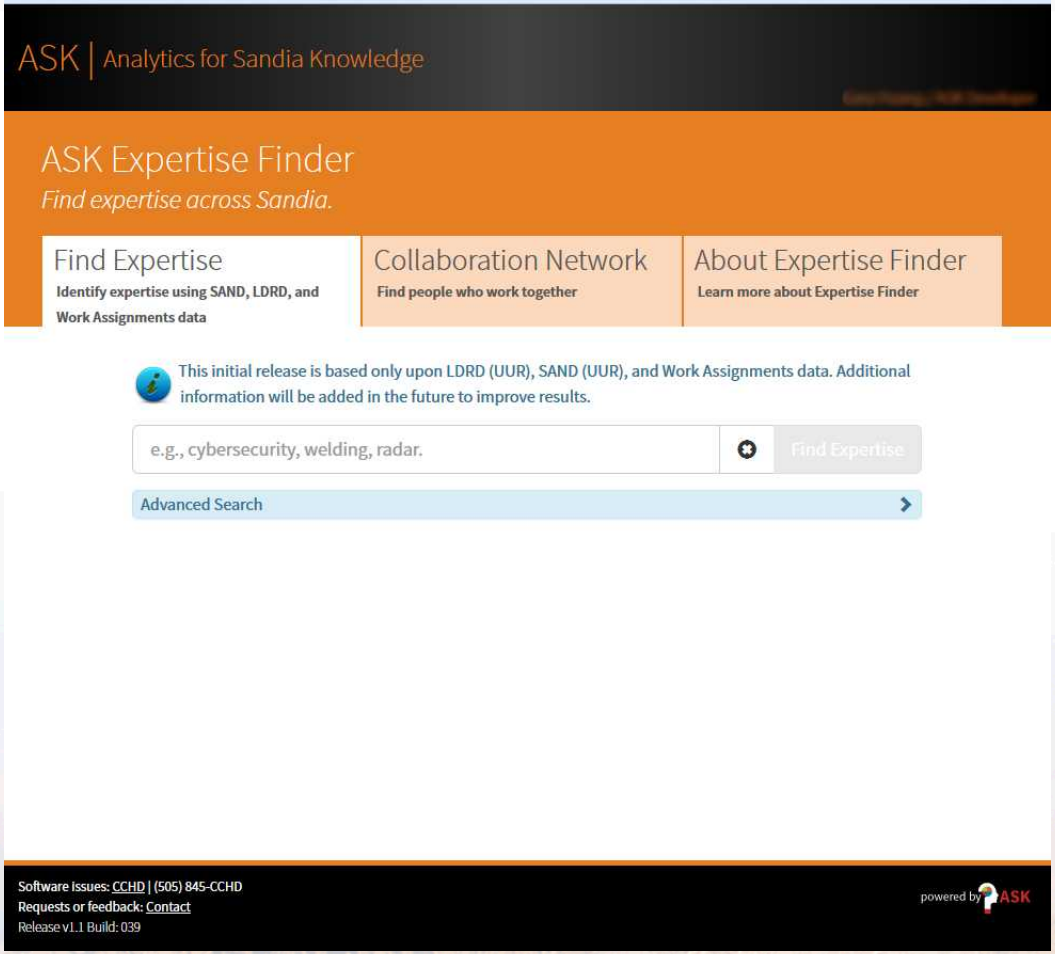
Field	IdfpoPSVBNtxx#txxDtxx	Norm	Value
abstract	Idfp--S--N--------	109	molecular simulations of liquid/vapor phase equilibria for single component and binary mixtures of nanoparticles m. a. horsch, p. j. in't veld, j. lechman, and g
doc_type	Id----S---#i32----	---	100
id	Id----S----------	---	100 2007-0178A
keywords	Idfp-----N--------	---	<not present or not stored>
main_conten	Idfp-----N--------	---	<not present or not stored>
org_number	Id----S----------	---	[redacted]
org_number	Id----S----------	---	[redacted]
org_number	Id----S----------	---	[redacted]
org_number	Id----S----------	---	[redacted]
rand_snl_id	Id----S----------	---	[redacted]
rand_snl_id	Id----S----------	---	[redacted]
rand_snl_id	Id----S----------	---	[redacted]
rand_snl_id	Id----S----------	---	[redacted]
title	Idfp--S--N--------	116	Molecular Simulations of Liquid/Vapor Phase Equilibria for Single Component and Binary Mixtures of Nanoparticles
year	Id----S---#i64----	---	2007

- Each document can have ≥ 0 associated Organization IDs, employee IDs
- Same org number can appear multiple times, to give more credit to orgs with more contributors
- Employee IDs can include inactive employees, useful for expertise trending
- Detailed org/people info are not in the index
  - Keeps index lean and simple
  - Auxiliary info for relevant documents is retrieved from database

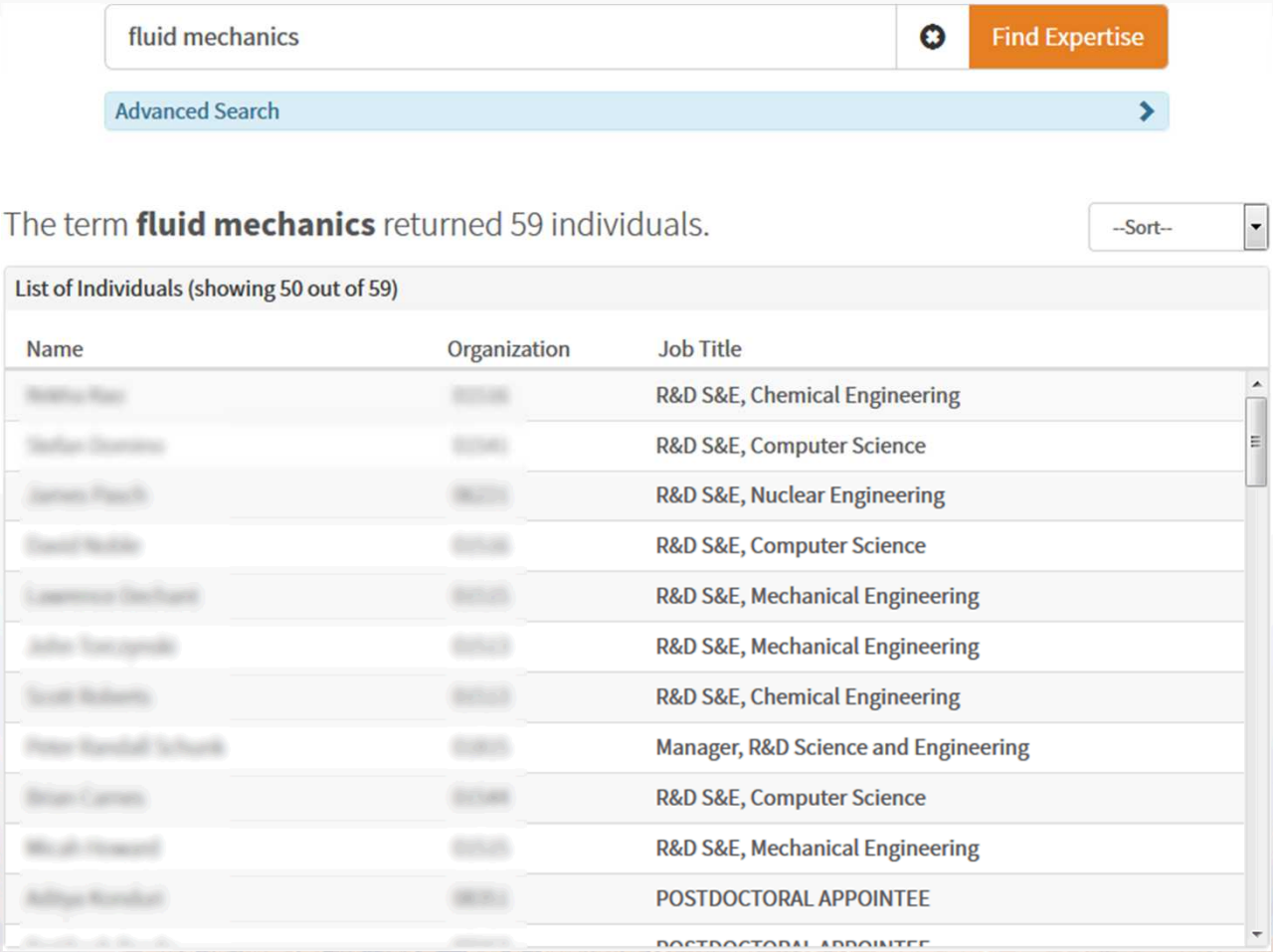
Data Access by User Role

- **Four user roles are currently defined for Expertise Finder, with varying levels of data access**
  - Data Scientist
  - Data Analyst
  - Managers
  - Member of the Workforce
- **Train a separate term expansion model for each role, and results are filtered by role**

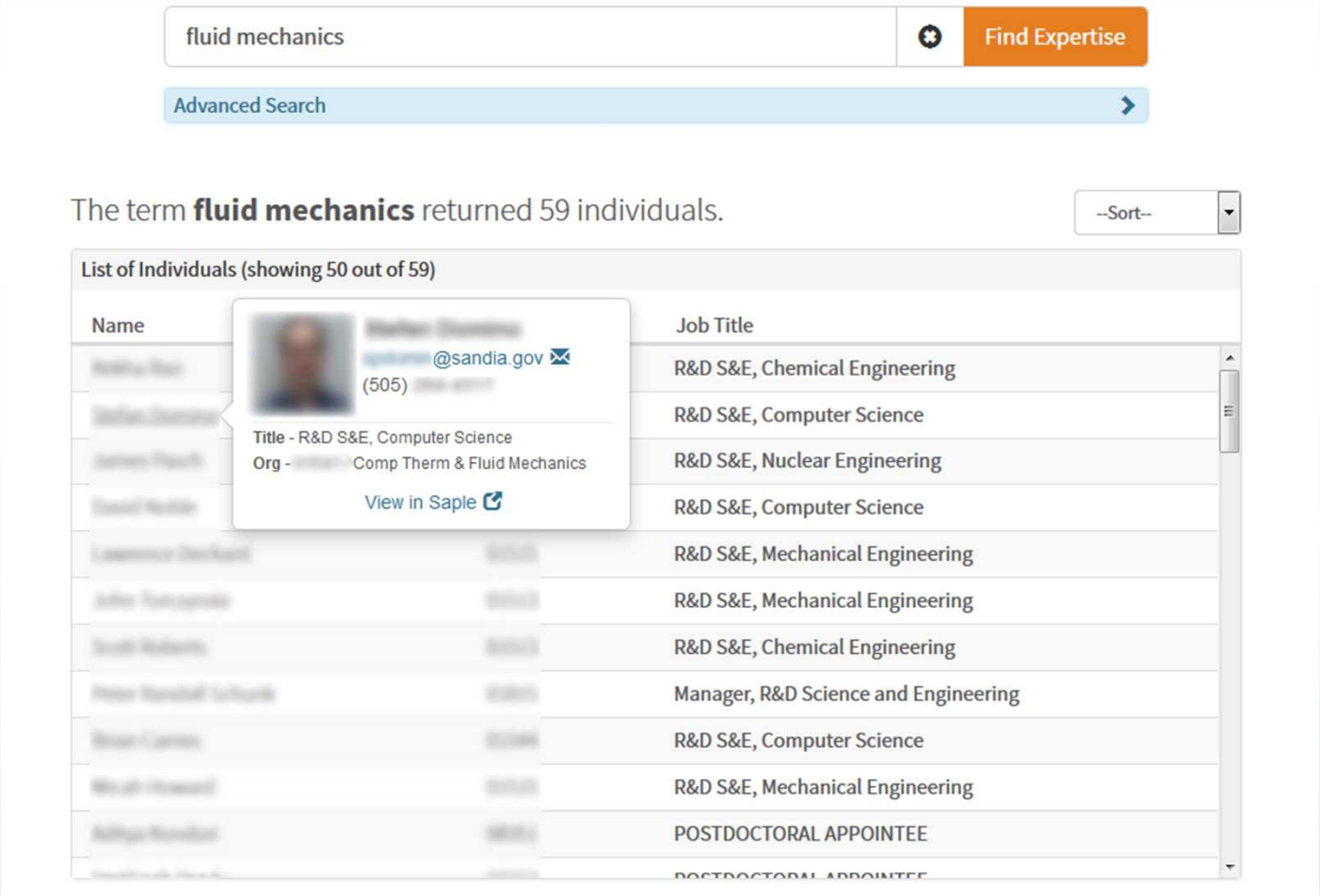
Expertise Finder Home Page



People with Expertise in ‘fluid dynamics’



Contact Information is Displayed when Hovering Over Result

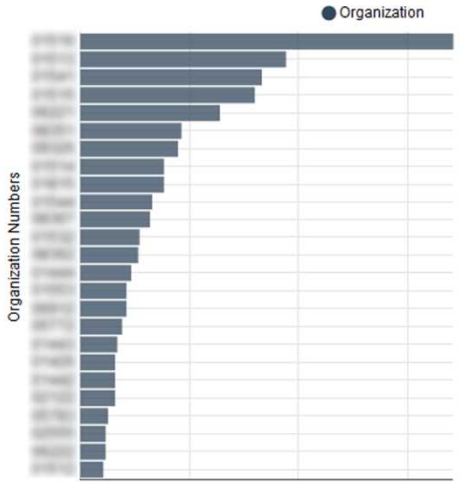


Organizations with Expertise in ‘fluid dynamics’

[redacted]	[redacted]	R&D S&E, Mechanical Engineering
[redacted]	[redacted]	POSTDOCTORAL APPOINTEE
[redacted]	[redacted]	POSTDOCTORAL APPOINTEE

The term "fluid mechanics" returned 40 organizations.

Chart of Expertise by Organization (Top 25 organizations at Sandia)



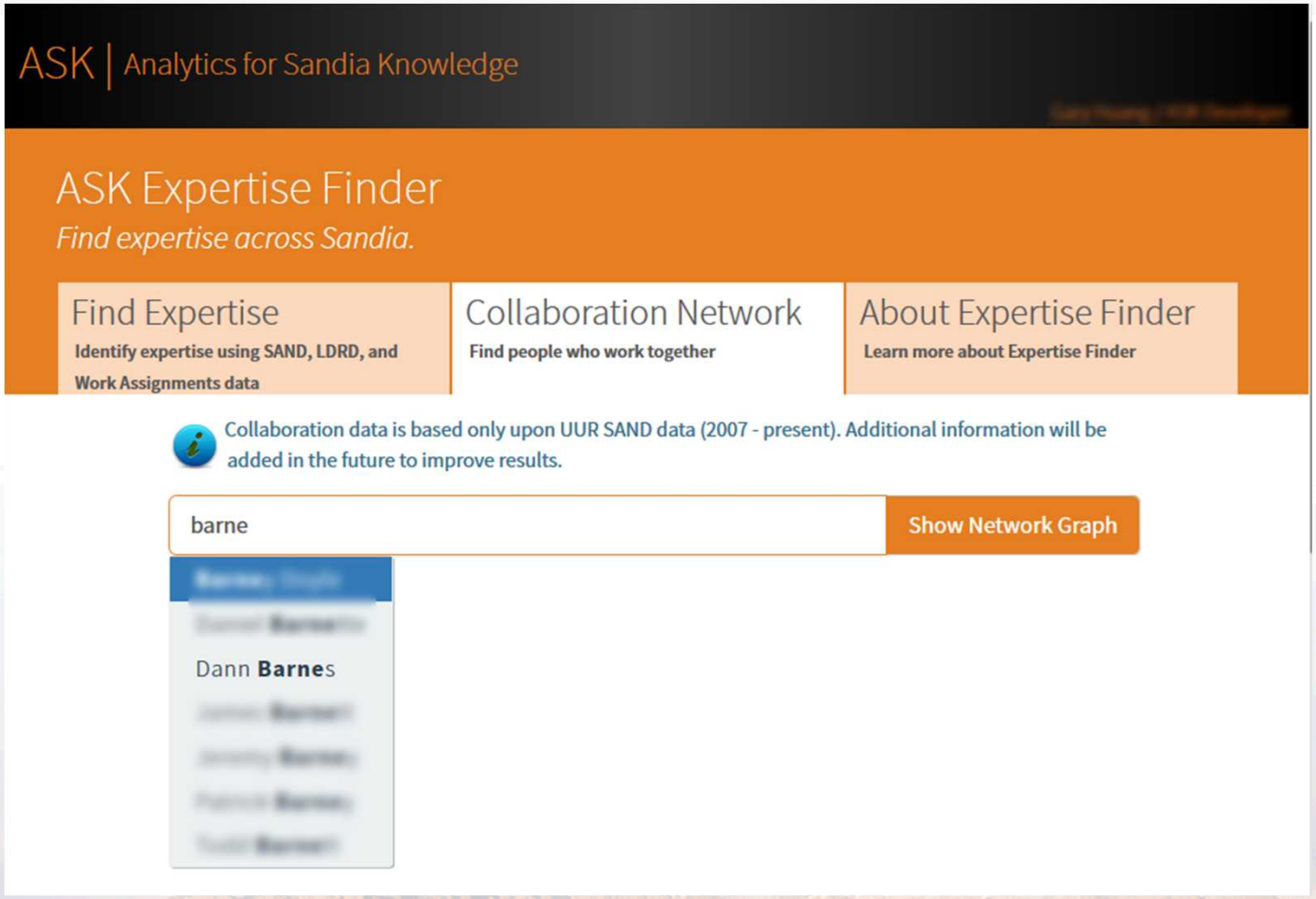
List of Expertise by Organization (showing 40 out of 40)

Name	Organization
Fluid and Reactive Processes	[redacted]
Thermal/Fluid Component Sci.	[redacted]
Comp Therm & Fluid Mechanics	[redacted]
Aerosciences Department	[redacted]
Advanced Nuclear Concepts	[redacted]
Reacting Flow Research	[redacted]
Scientific APPS & USER SUPPORT	[redacted]
Thermal Sciences & Engineering	[redacted]
Advanced Materials Laboratory	[redacted]

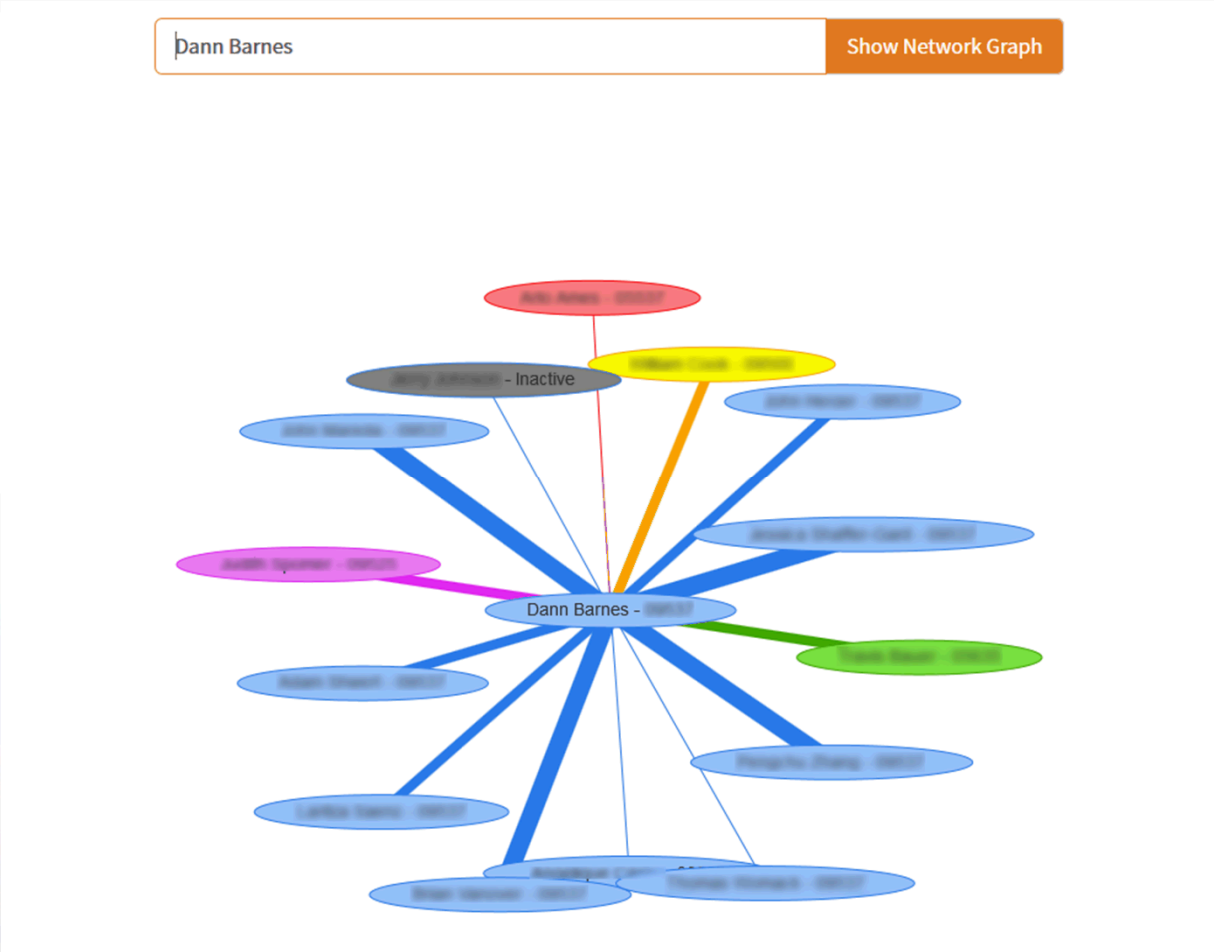
Software issues: CCHD | (505) 845-CCHD
Requests or feedback: Contact
Release v1.1 Build: 039

powered by ASK

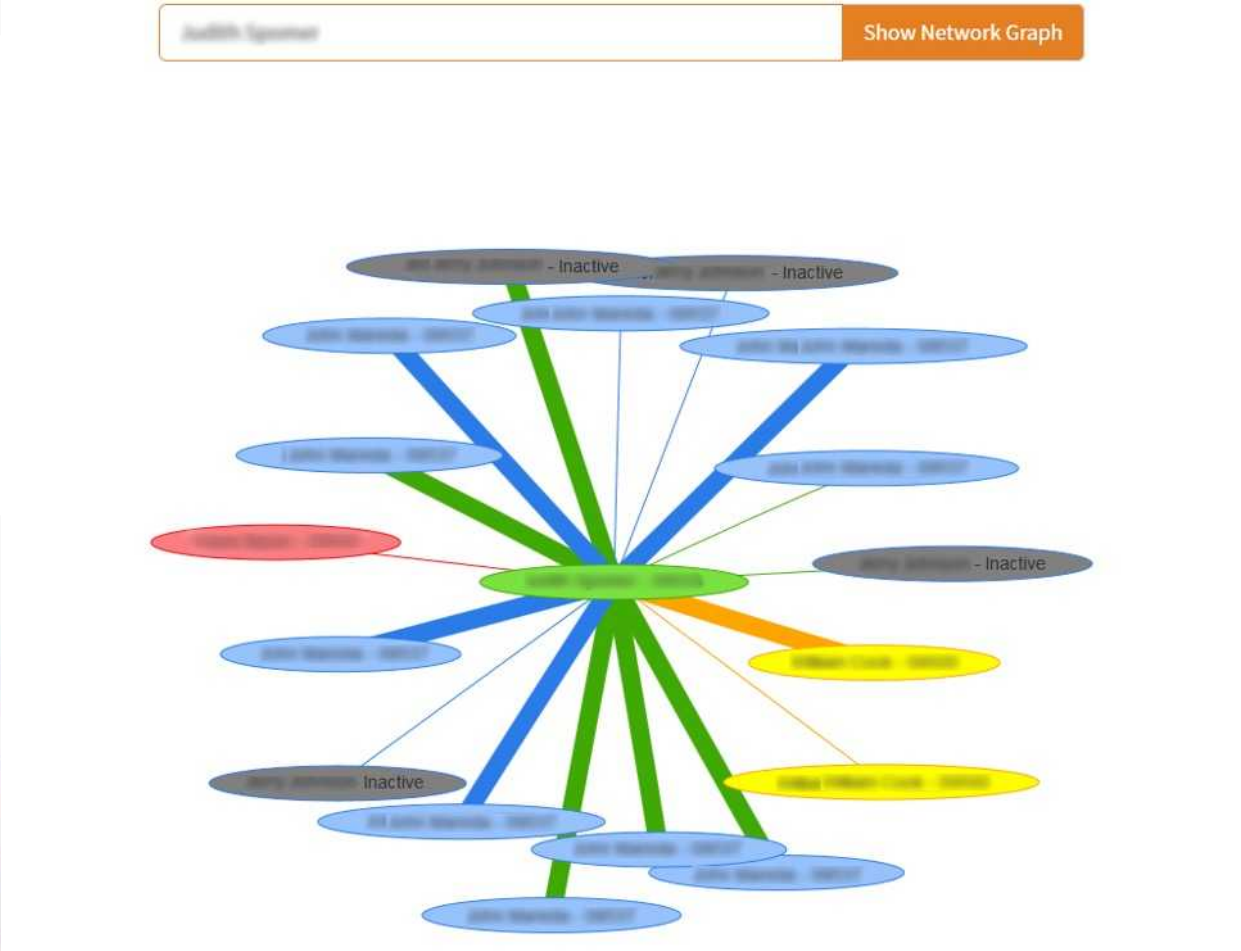
Visualizing Collaboration Networks



Coauthor Network Graph



Clicking on a Node Redraws the Graph



Measuring & Tracking Quality

- **Interview managers across Sandia**
  - Get their judgment on Expertise Finder results for the areas they know about

- **Results from Feb 2016:**

Judgement	Count	%
Good	55	52
Mixed	27	26
Poor	23	22

Future Work

- Add more sources of data
- Gather more user feedback on result accuracy
- Tweak the ranking model to improve results
- Incorporate Expertise Finder results into Sandia's enterprise search results
- Enable ability to drill down to source documents