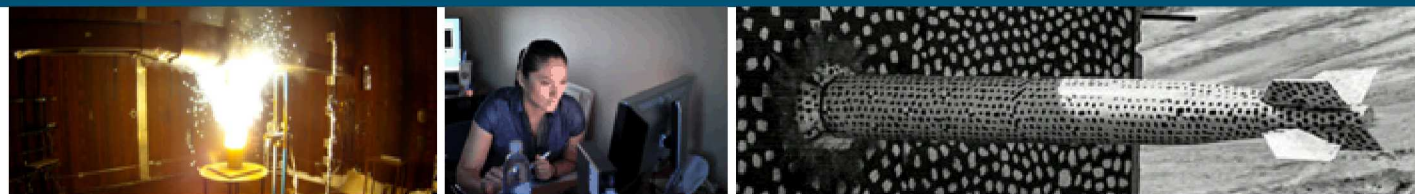


# Smart Personalized Information Retrieval Environment



*PRESENTED BY*

Pengchu Zhang and John Herzer

National Laboratories Information Technology  
Summit 2018



Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

## 2 Enhancing Enterprise Search

### SearchPoint Application Profile

- Deployed on Apache Solr search engine
- Migrating to LucidWorks Fusion
- 360,000 pages indexed
- Average of 13000 queries per day
- 8300 distinct customers submitted queries during the year

### Recent Enhancements

- Applications that Listen
- Link popularity boosting



## Exploring analytics to improve search results

- Click popularity boosting – Aggregate Behavior Model
- Word2Vec for synonym expansion
- Doc2Vec feature extraction for Related SAND reports
- General Reranker is a framework for evaluating these models

Showing 1 - 25 of 5184 results

### Education **Outreach**

<https://sharepoint.sandia.gov/sites/snlci/SitePages/Education.aspx>

Education Outreach

### Public Relations & Strategic Communication (PRSC) (SNL/CA)

<http://info.sandia.gov/centers/8500/prsc>

Design & Publishing Center, **Outreach**, Community **Outreach**, PRSC, Public Relations & Strategic Communication (PRSC), Media

### Division Diversity Council (DDC) Home

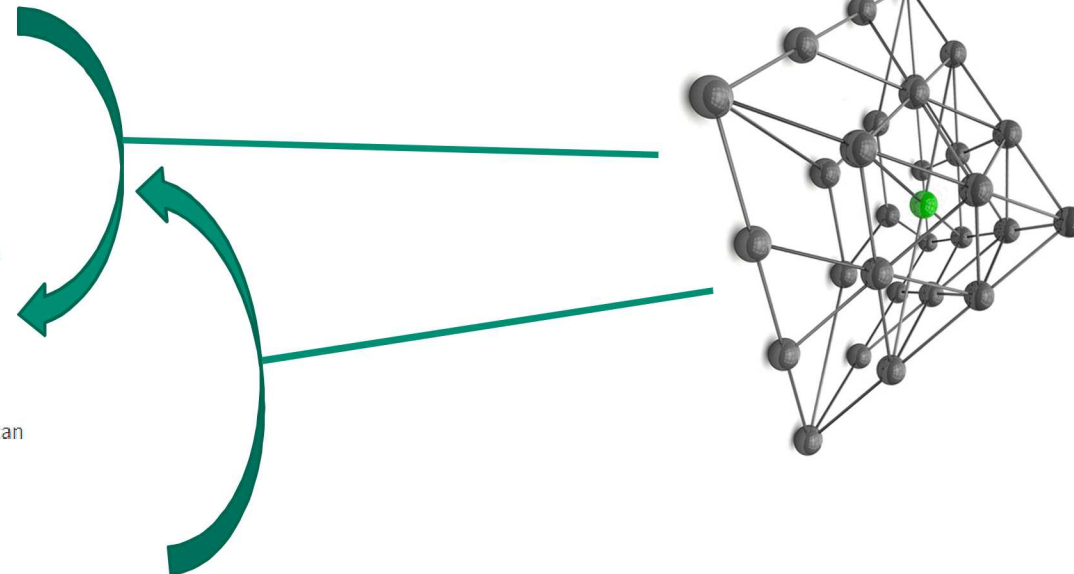
<https://sharepoint.sandia.gov/sites/ddcsnl/Pages/DDC-Groups-Summary.aspx>

Connection, DDC, AAOC, APLC, CWNG, FNNG, GLBT, HLC SWC, MSC, WAG, Diversity, African American **Outreach** Committee, Asian Pacific Leadership

### Pages - **Outreach**

<https://sharepoint.sandia.gov/sites/BLCe/pages/outreach-.aspx>

Site Contents **Outreach** Page Content For Technical Assistance:CCHD(505) 845-2243  
corpInfo



## Presenting personalized recommendations

Multiple components already provide ranking inputs

- Core search engine uses term frequency based ranking (TF/IDF algorithm)
- Rankings adjusted by URL popularity boosting

Worked with the UX team to determine best display options

Decided to display recommendations in a separate panel

- Didn't want another system competing to adjust rankings
- SPIRE recommendations are more course grained
- Want to distinguish them from pure content recommendations

# SAND Recommendation Layout

- SPIRE Recommendations apply to customer not search term
- Verify that results contain at least one SAND from topic of interest

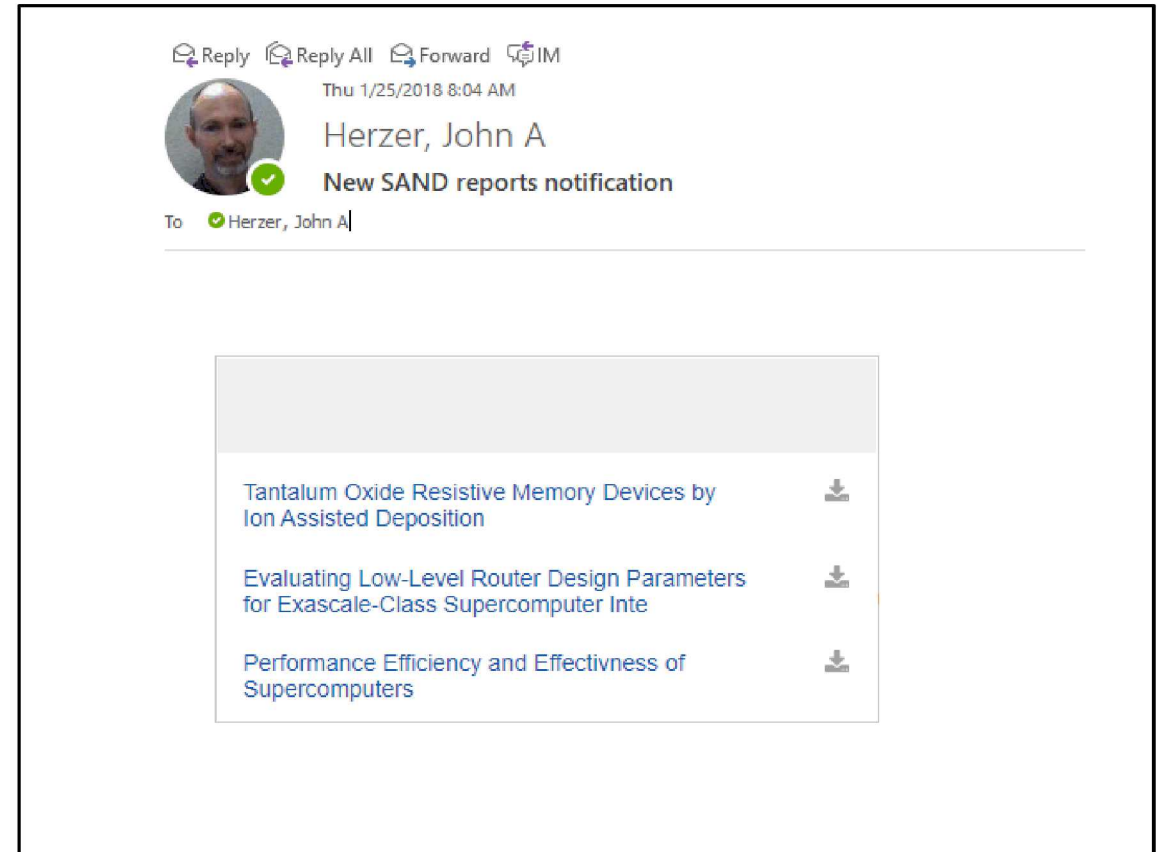
The screenshot displays the SAND search interface. At the top, there is a navigation bar with 'INSIDE', 'SMM', 'Policies', 'Orgs', and 'News'. A search bar contains the query 'semiconductor quantum dots' with a magnifying glass icon. To the right of the search bar is a 'Rate These Search Results' section with five stars. Below the search bar, there are tabs for 'Sandia' and 'FileNet'. The main content area is divided into three sections:

- Filter your results:** A sidebar on the left with a question mark icon, listing various filters: SAND Reports (6429), Sandia External Web (378), Sandia News (310), Sandia Internal Web (250), SharePoint (246), Sandia Videos (8), and Sandia Management Information (1).
- Did you mean:** A section suggesting 'semiconductor quantum dmts' with a question mark icon. Below it, it says 'Showing 1 - 25 of 7622 results'. Three search results are listed:
  - Single-electron-occupation metal-oxide-semiconductor quantum dots formed from efficient poly SAND2017-7551J**  
[http://prod.sandia.gov/sand\\_doc/2017/177551j.pdf](http://prod.sandia.gov/sand_doc/2017/177551j.pdf)  
 [sand] Single-electron-occupation metal-oxide-semiconductor quantum dots formed from efficient poly-silicon gate layout S. Rochette1,2
  - Anisotropic hole g-factors and relevance to photon-to-spin conversion schemes in semiconductor quan...**  
 SAND2017-5794A  
[http://prod.sandia.gov/sand\\_doc/2017/175794a.pdf](http://prod.sandia.gov/sand_doc/2017/175794a.pdf)  
 [sand] quantum dot circuits 2) Authors' list: A. Bogan,1,2 S. A. Studenikin,1 M. Korkusinski,1 G. C. Aers,1 L. Gaudreau,1 P. Zawadzki,1 A. Kam
  - Model for a Semiconductor Quantum-Dot Nanolaser**  
 SAND2014-16567A  
[http://prod.sandia.gov/sand\\_doc/2014/1416567a.pdf](http://prod.sandia.gov/sand_doc/2014/1416567a.pdf)  
 [sand] Model for a semiconductor quantum-dot nanolaser W. W. Chow, 1 F. Jahnke 2 and C. Gies 2 1Sandia National Laboratories
  - Model for a Semiconductor Quantum-Dot Nanolaser**  
 SAND2014-16519PE  
[http://prod.sandia.gov/sand\\_doc/2014/1416519pe.pdf](http://prod.sandia.gov/sand_doc/2014/1416519pe.pdf)  
 [sand] Model for a semiconductor quantum-dot nanolaser Weng Chow,
- Personalized SAND Document Recommendations:** A blue sidebar on the right with a white background, listing recommended documents:
  - Density Functional Analysis of Fluorite-Structured (Ce, Zr)O<sub>2</sub>/CeO<sub>2</sub> Interfaces (Density Funct...)  
SAND2017-4331J
  - Quantum Mechanical Treatment of Electron Broadening  
SAND2017-3254C
  - Studying Si/SiGe disordered alloys within effective mass theory  
SAND2017-2873C
  - Quantum Mechanical Treatment of Electron Broadening  
SAND2017-3549C
  - Two-well terahertz quantum cascade lasers with suppressed carrier leakage  
SAND2017-8933J

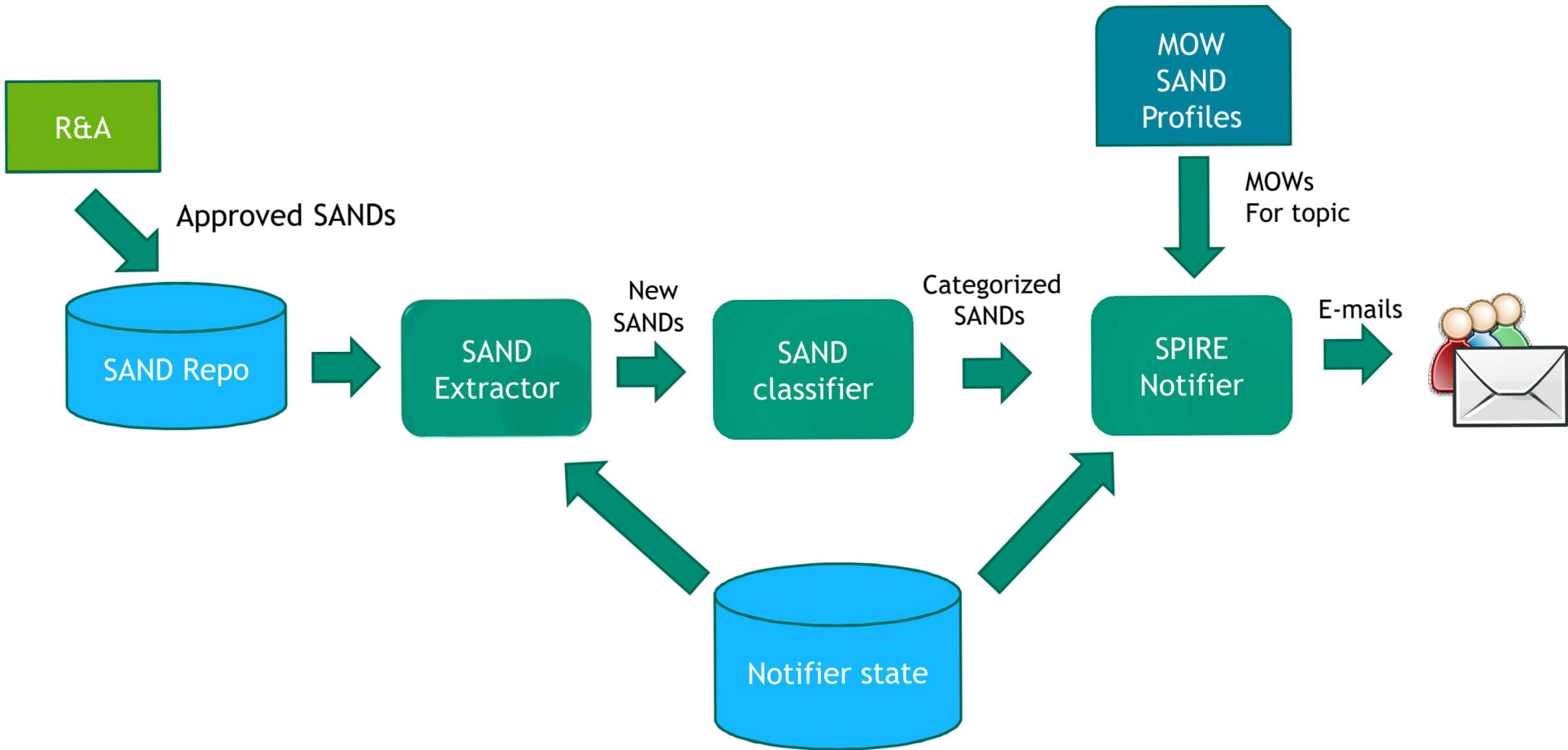
At the bottom of the 'Personalized SAND Document Recommendations' sidebar, there is a 'Give Feedback' link.

## Pushing recommendations to the customer

- Email notifications are the push side of recommendations
- Need to prevent repeating previous recommendations
- Identify optimal frequency and provide opt-out capability

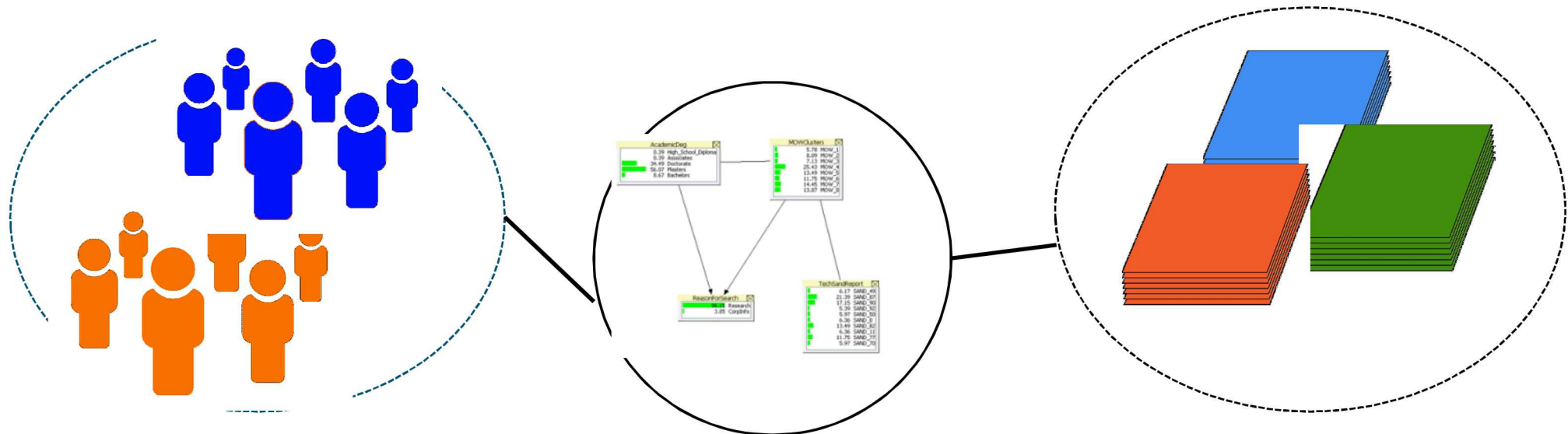


7 SPIRE Notification Process



## Providing personalized content recommendations

- SPIRE – Sandia Personalized Information Retrieval Environment
- Match customers with relevant content based on their information activities
- Group MOWs by their common interests
- Cluster related content by common term usage
- Develop predictive models that show who is likely to want which content

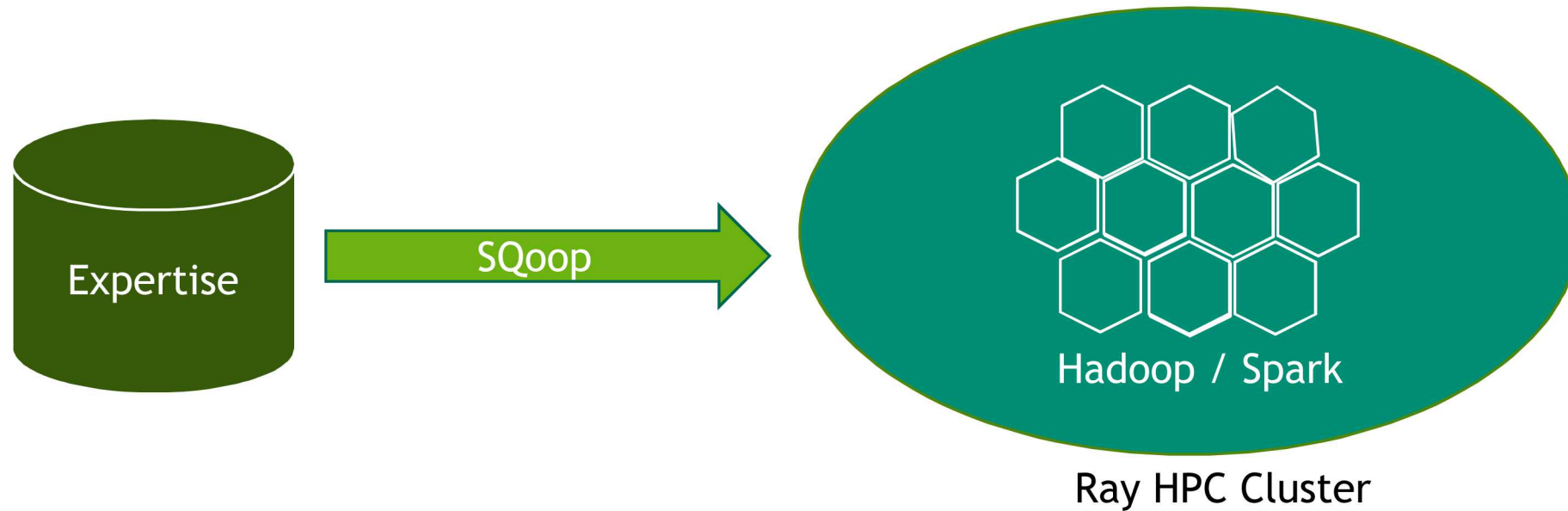




## 9 Profiling MOWs by activity

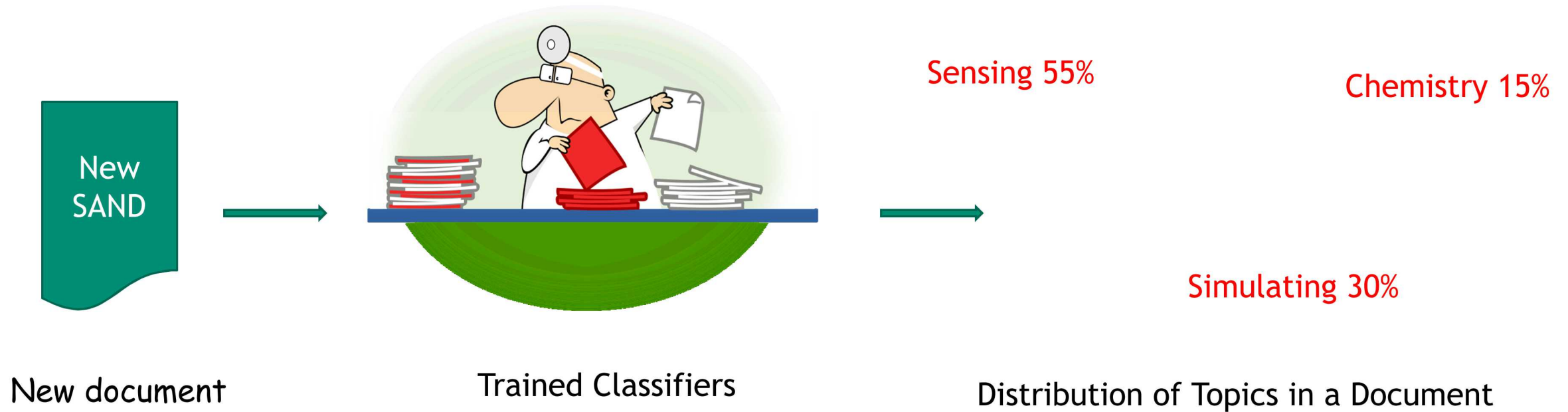
Customers are clustered based on their attributes and information usage

- Personal attributes: education, years at Sandia...
- Documents created: SAND reports, LDRDs, patents, PMF objectives ...
- K-means used to cluster MOWs with  $K=30$



## Build SAND Report Classifiers

1. Classify documents into proper classes
2. Recognize the document class in various formats
3. Recognize the distribution of possible classes in a document



- Collected ~100,000 SAND reports over last 50 years
- Data cleaned and indexed with Apache/Lucene
- Built “Taxonomy” for Sandia Category Guide (SCG)
- Selected the highly ranked documents with SCG taxonomy terms/phrases as the training sets
- Build a Word2Vec language model for embedding representatives of words
- Built a Convolutional Neural Network (CNN)
- Trained the network with various hyperparameters

# Build “Taxonomy” based on Sandia Category Guide (SCG)

## Category (Material Sciences)

### Subcategories:

Ceramics  
plastics  
seals and Adhesives

### Obtain the terms/phrases for subcategories from:

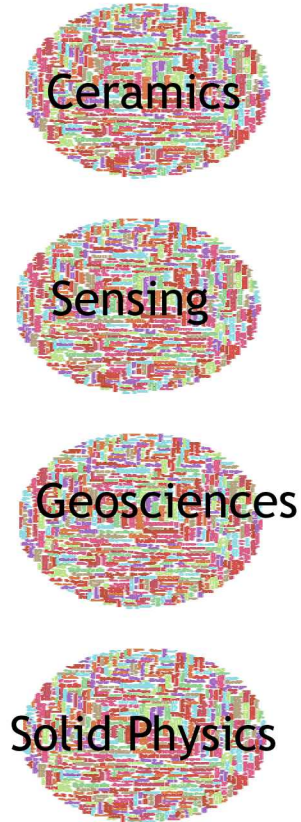
Documents created by Sandia authors  
Wikipedia  
Word2Vec built on Sandia’s documents  
Internal Organization webpages

### Taxonomy Example for “Material Sciences/Composite Materials:

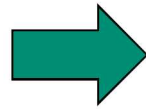
"carbon composite" "carbon fiber" "ceramal" "ceramic composite"  
"ceramic matrix composite" "CFRP" "Chobham armour" "clad metals" "CMC"  
"composite material" "concrete-plastic composite" "fiber reinforced composite"  
"fiber reinforced polymer" "fiberglass" "formica" "FRP" "glasdpolyester"  
"graphite" "GRP" "laminare" "Mallite" "metal composite" "metal matrix composite“...

NOTE: Taxonomy used here is simplified from Jessica Shaffer-Gant’s version

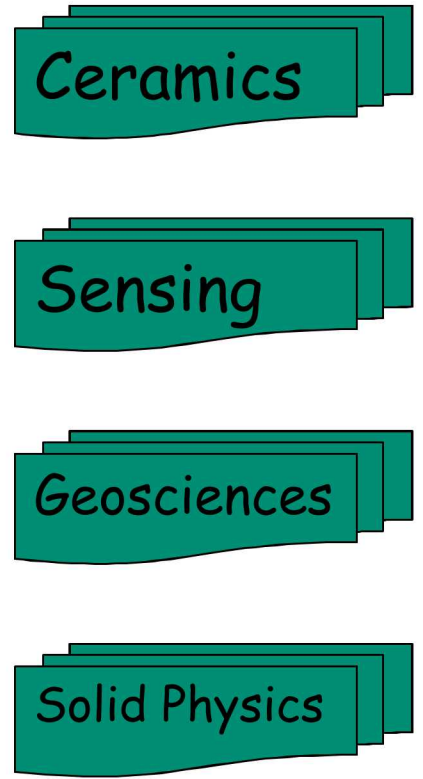
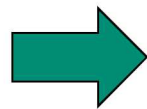
# Collect Sets of Labeled Documents for Training



Terms/Phrases from Taxonomy

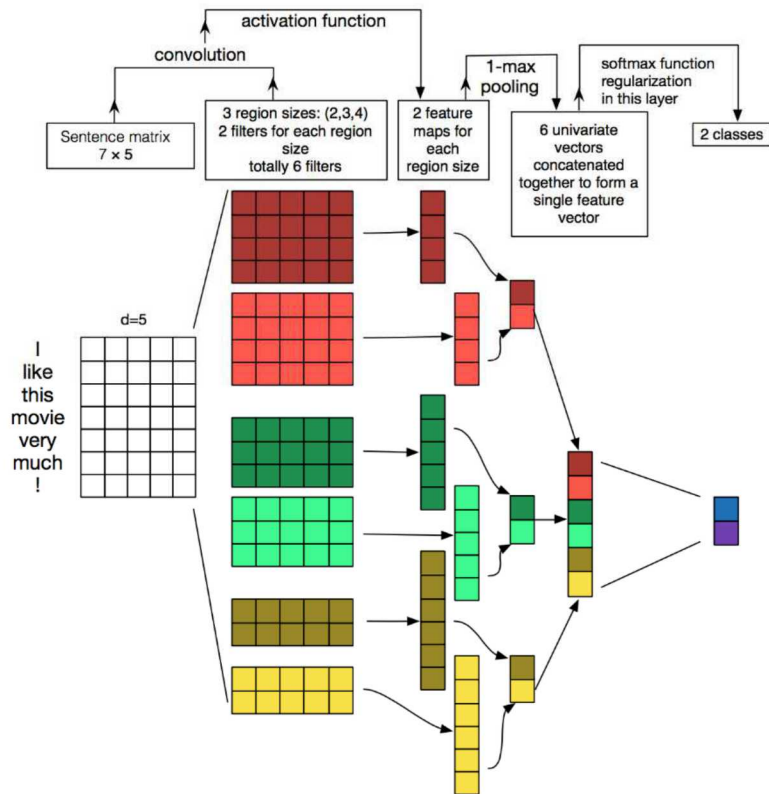


Search Engine



Labeled and Ranked Documents

# Convolutional Neural Network for Text Classification



```
print('Training model.')
```

```
# train a 1D convnet with global maxpooling
sequence_input = Input(shape=(MAX_SEQUENCE_LENGTH,),
dtype='int32')
embedded_sequences = embedding_layer(sequence_input)
x = Conv1D(128, 5, activation='relu')(embedded_sequences)
x = MaxPooling1D(5)(x)
x = Dropout(0.5)(x)
x = Conv1D(128, 5, activation='relu')(x)
x = Dropout(0.5)(x)
x = MaxPooling1D(5)(x)
x = Conv1D(128, 5, activation='relu')(x)
x = Dropout(0.5)(x)
x = MaxPooling1D(35)(x)
x = Flatten()(x)
x = Dense(128, activation='relu')(x)
preds = Dense(len(labels_index), activation='softmax')(x)
model = Model(sequence_input, preds)
model.compile(loss='categorical_crossentropy', optimizer='rmsprop',
metrics=['acc'])
```

Modified from a Keras example

## Examples of Classifiers for Classification

```
prediction = model.predict(data[146:150])
```

```
[37 1] [ 6.892 79.014] % SAND2000-0217.txt thermodynamics atmospheric sciences  
[30 14] [ 6.755 91.043] % SAND2000-0218.txt particle physics electronics and electrical engineering  
[16 31] [ 6.668 60.088] % SAND2000-0221.txt fluid mechanics plasma physics  
[ 9 29] [ 29.797 52.055] % SAND2000-0222C.txt computer architecture optics
```

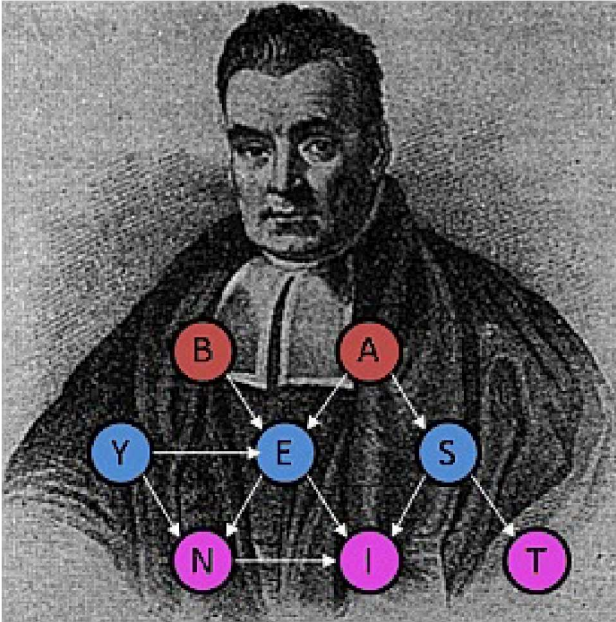
Class ID	Class Distribution	SAND ID	Class Names
----------	--------------------	---------	-------------

# Example of Classifying a Document

[37 1] [ 6.892 79.014] % SAND2000-0217.txt thermodynamics atmospheric sciences







Posterior probability of 'x'  
given the evidence 'E'

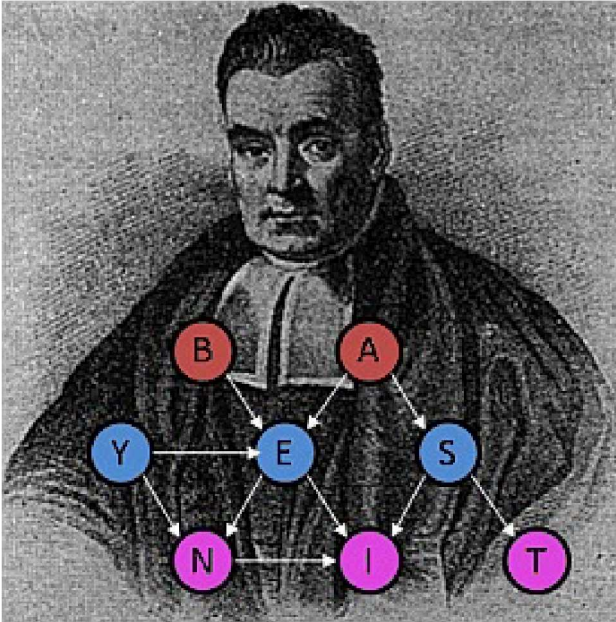
$$P(x|E)$$

Prior probability

Likelihood of the evidence of 'E'  
If the hypothesis 'x' is true

$$= \frac{P(x) * P(E|x)}{P(E)}$$

Prior probability that  
the evidence itself is true



Probability of Lung cancer patients among people

Probability of smokers among lung cancer patients

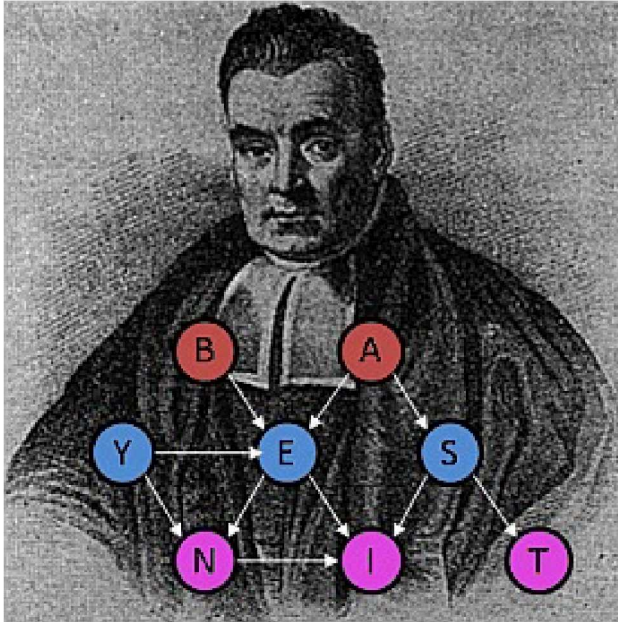
$$P(x|E) = \frac{P(x) * P(E|x)}{P(E)}$$

If one is a smoker, what is the probability he/she may have lung cancer

Probability of smokers among people

$$P(LC|S) = \frac{0.00006 * 0.7}{0.2} = 0.00021$$

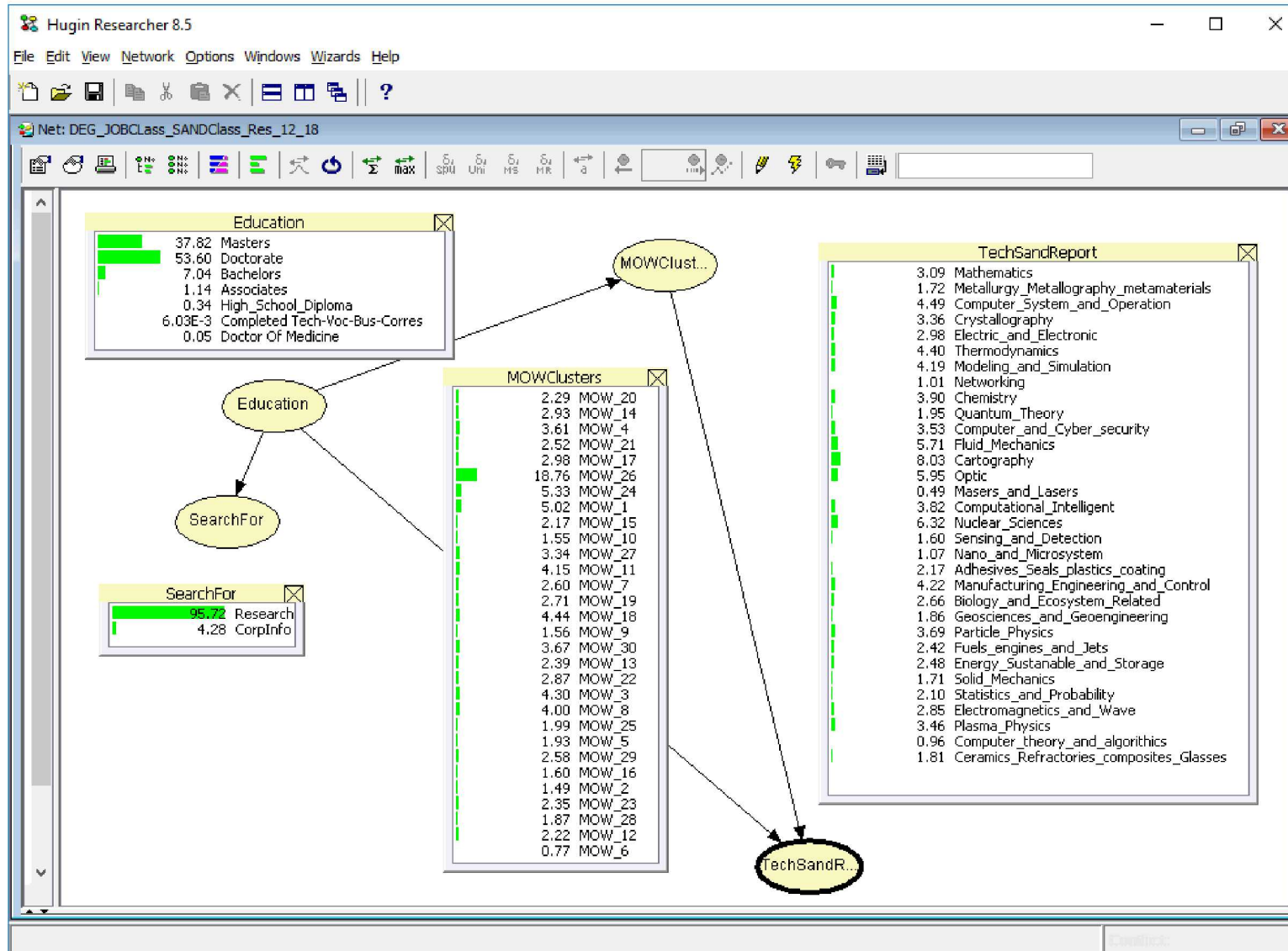
$$\frac{0.00021}{0.00006} = 3.5 = 350\%$$



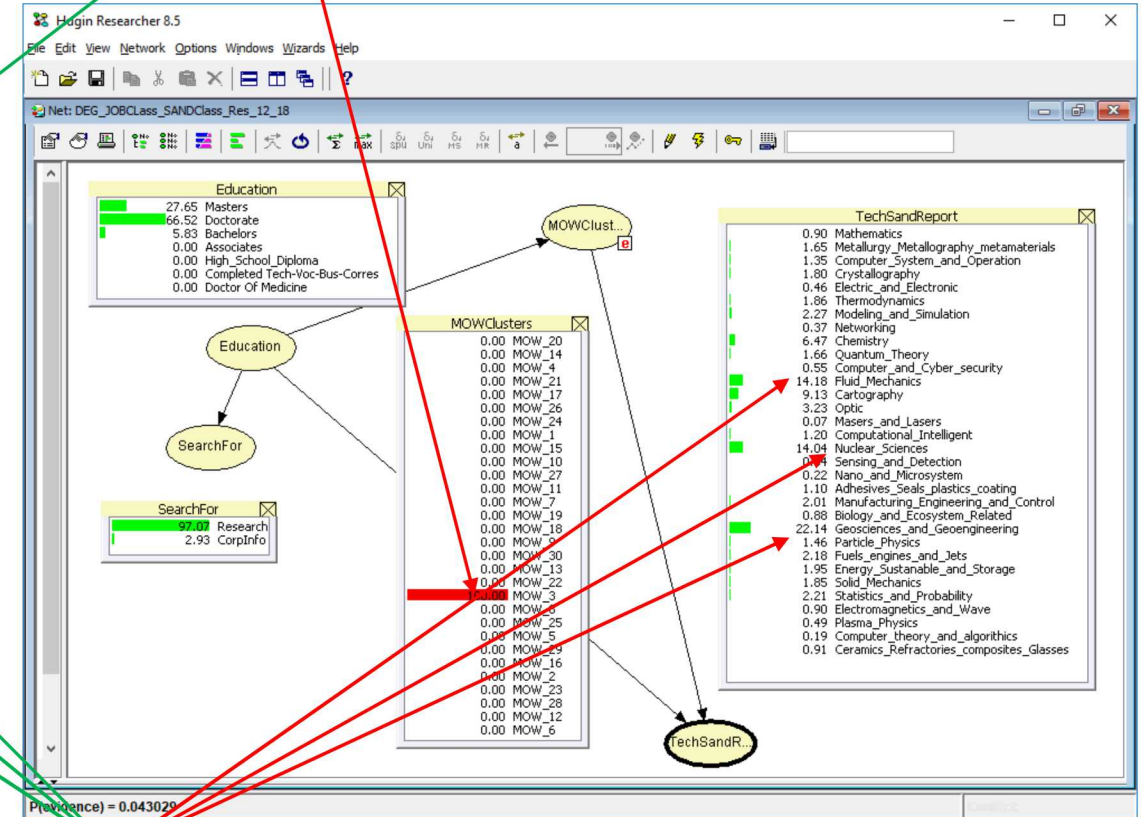
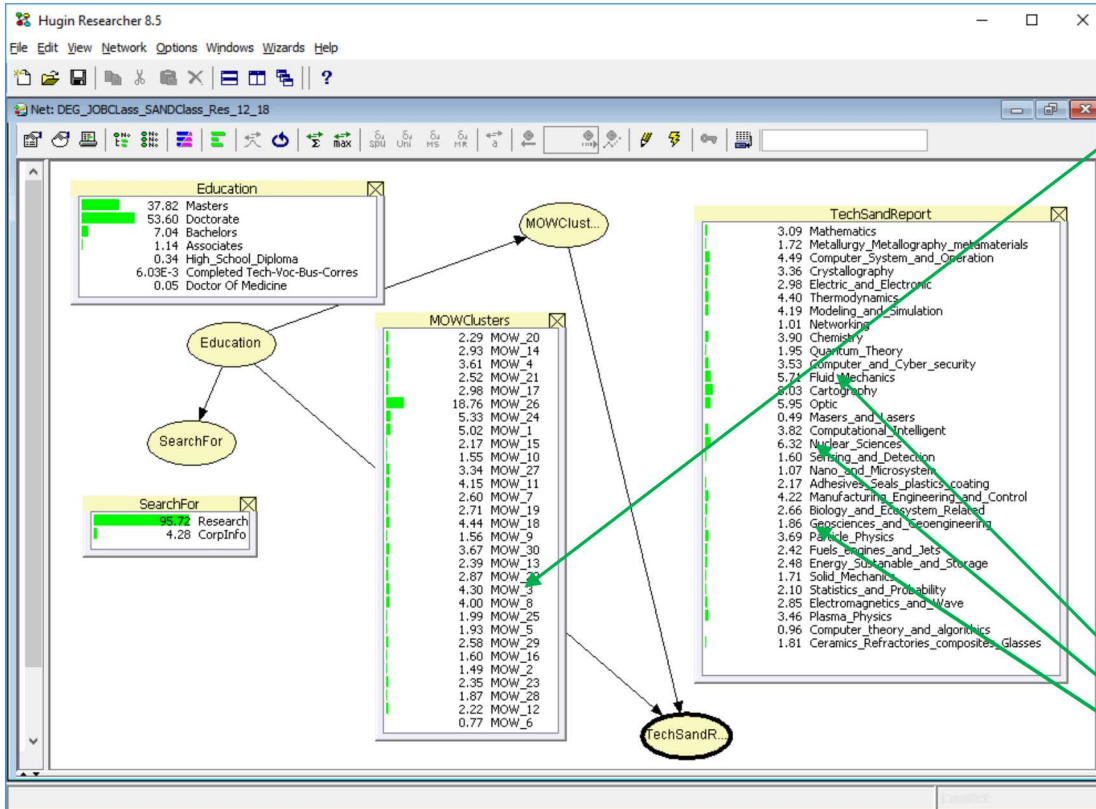
$$P(x|E) = \frac{P(x) * P(E|x)}{P(E)}$$

$P(\text{computer})$  points to  $P(x)$   
 $P(\text{webDev}/\text{computer})$  points to  $P(E|x)$   
 $P(\text{computer} | \text{webDev})$  points to  $P(x|E)$   
 $P(\text{webDev})$  points to  $P(E)$

# Distribution of Sandia R&D S&E Clusters and Usage of SAND Reports

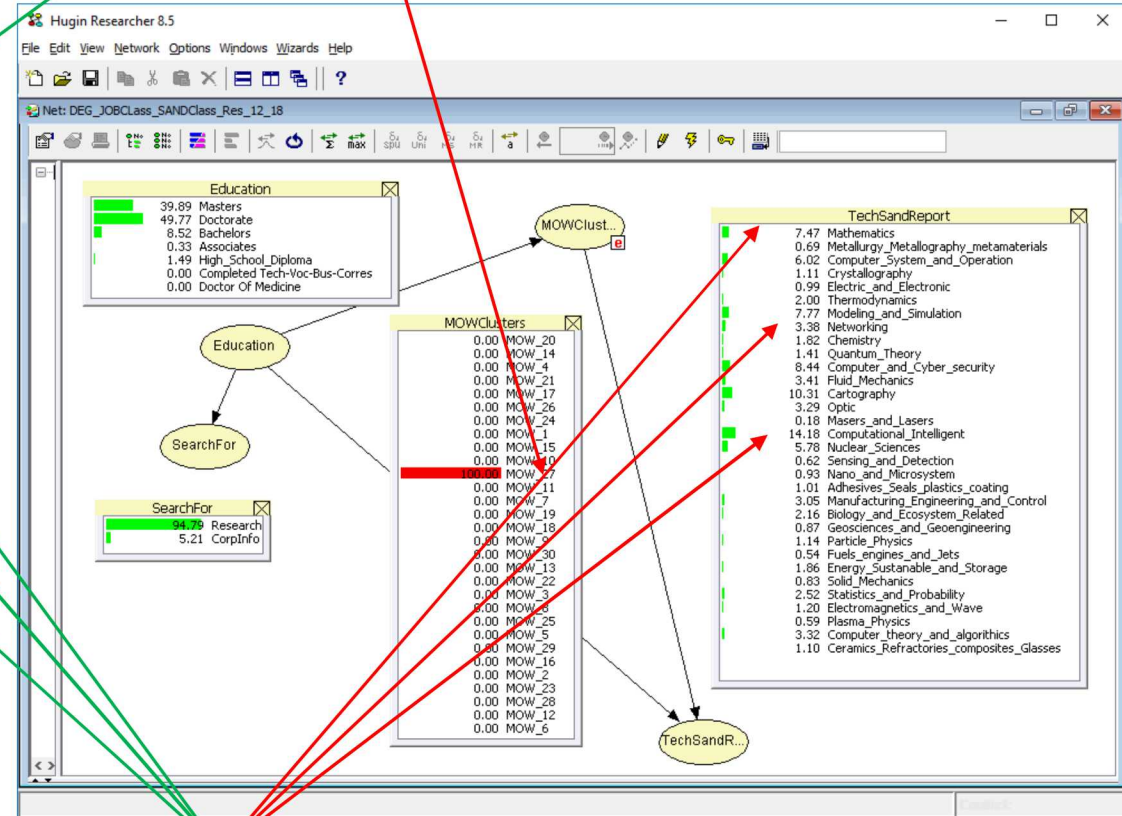
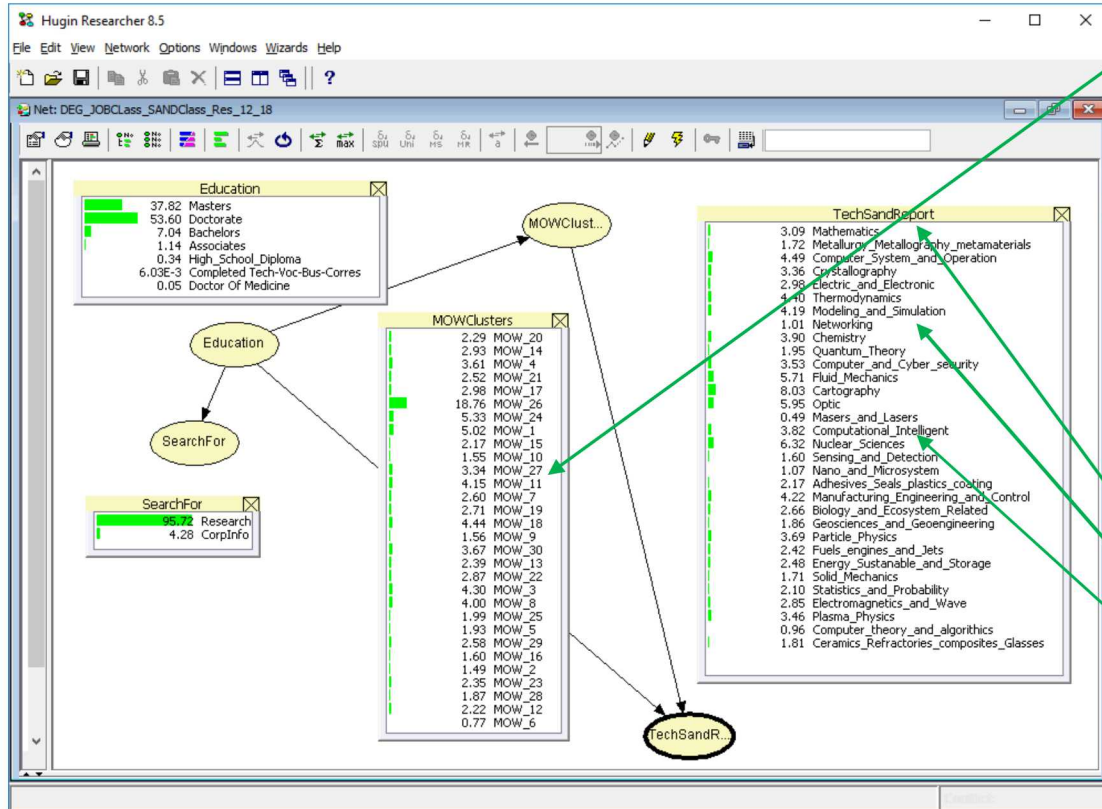


# Comparison of SAND Reports Checked by Nuclear-Waste Management and Other Groups



Geoscience	1.86% to 22.14%
Nuclear science	6.32% to 14.04%
Fluid mechanics	5.71% to 14.18%

# Comparison of SAND Reports Checked by Machine-learning and Other Groups



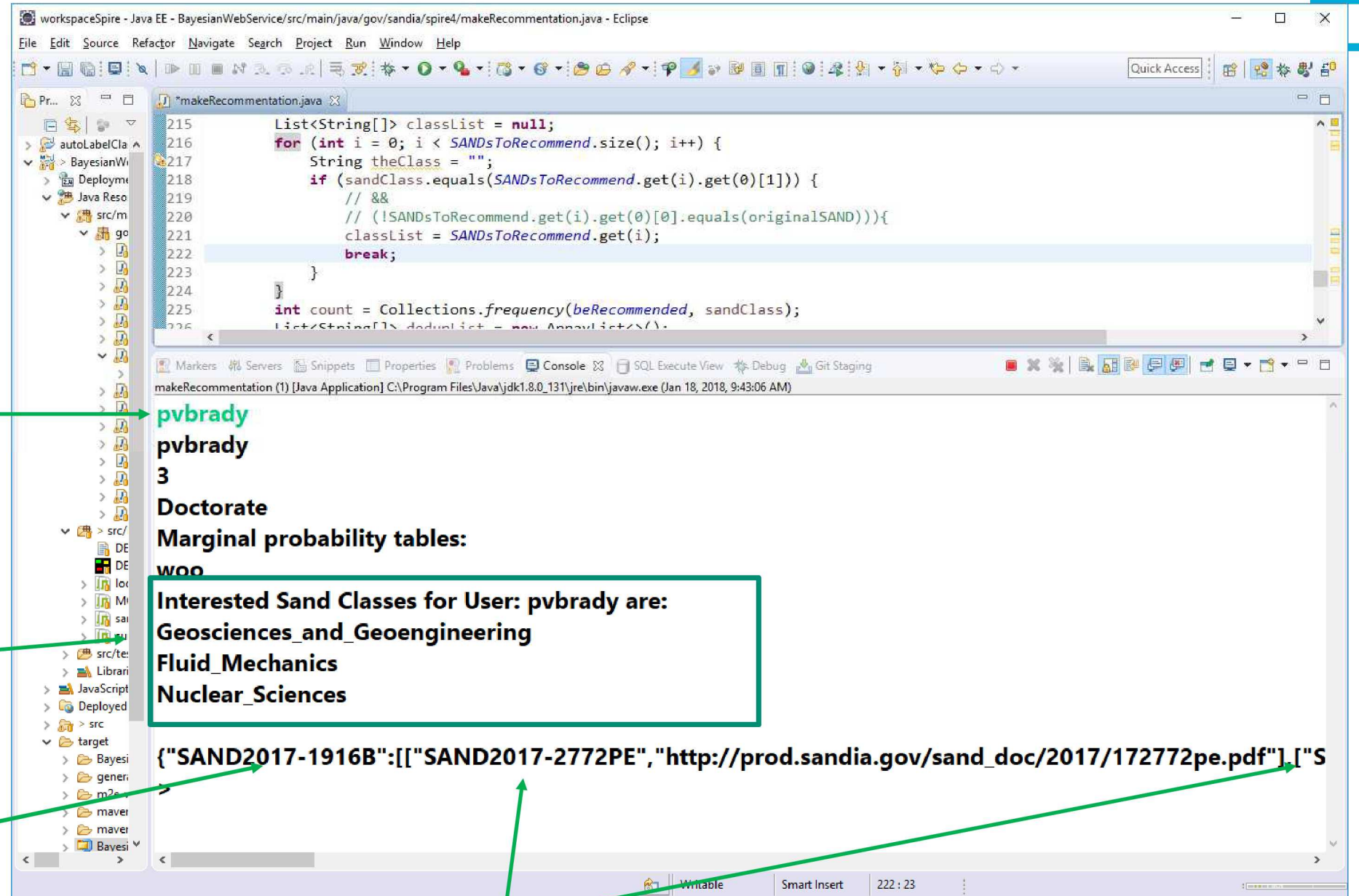
Computer intelligence 3.82% to 14.18%  
 Mathematics 3.09% to 7.47%  
 Modeling simulation 4.19% to 7.77%

How personalized recommendations are determined

The User

Sand classes interested

Queried SAND



```

workspaceSpire - Java EE - BayesianWebService/src/main/java/gov/sandia/spire4/makeRecommendation.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Pr...
autoLabelCla
BayesianWi
Deployment
Java Reso
src/m
go
src/
DE
DE
loc
M
sa
src/te
Librari
JavaScript
Deployed
src
target
Bayesi
gener
m2e
maver
maver
Bavesi

```

```

215 List<String[]> classList = null;
216 for (int i = 0; i < SANDsToRecommend.size(); i++) {
217     String theClass = "";
218     if (sandClass.equals(SANDsToRecommend.get(i).get(0)[1])) {
219         // &&
220         // (!SANDsToRecommend.get(i).get(0)[0].equals(originalSAND))){
221         classList = SANDsToRecommend.get(i);
222         break;
223     }
224 }
225 int count = Collections.frequency(beRecommended, sandClass);
226 List<String[]> dedupList = new ArrayList<>();

```

Markers Servers Snippets Properties Problems Console SQL Execute View Debug Git Staging

makeRecommendation (1) [Java Application] C:\Program Files\Java\jdk1.8.0\_131\jre\bin\javaw.exe (Jan 18, 2018, 9:43:06 AM)

pvbrady  
pvbrady  
3  
Doctorate  
Marginal probability tables:  
woo

**Interested Sand Classes for User: pvbrady are:**  
Geosciences\_and\_Geoengineering  
Fluid\_Mechanics  
Nuclear\_Sciences

{\"SAND2017-1916B\": [[\"SAND2017-2772PE\", \"http://prod.sandia.gov/sand\_doc/2017/172772pe.pdf\"], [\"S

Writable Smart Insert 222 : 23

Recommended SANDS

How personalized recommendations are determined

The User

Sand classes interested

Queried SAND

```

workspaceSpire - Java EE - BayesianWebService/src/main/java/gov/sandia/spire4/makeRecommendation.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
makeRecommendation.java
89     getSANDsToRecommend(docUrls); // build SANDsToRecommend list for
90     // this user
91     List<String[]> rList = recommendation(docUrls);
92
93     int count = 0;
94     List<String[]> arr = null;
95
96     String origSAND = null;
97     String origSAND_ID = null;
    
```

Markers Servers Snippets Properties Problems Console SQL Execute View Debug Git Staging

makeRecommendation (1) [Java Application] C:\Program Files\Java\jdk1.8.0\_131\jre\bin\javaw.exe (Jan 24, 2018, 1:46:52 PM)

**Number of SAND Reports: 32**  
**Enter a username, or 'q' to quit**  
 > **jherzer**  
**jherzer**  
**27**  
**Masters**  
**Marginal probability tables:**  
**woo**  
**Interested Sand Classes for User: jherzer are:**  
**Computational\_Intelligent**  
**Computer\_theory\_and\_algorithics**  
**Networking**

**{"SAND2017-7067C": [{"SAND2017-2553J", "http://prod.sandia.gov/sand\_doc/2017/172553j.pdf"}, {"SAN**

Writable Smart Insert 51 : 55

Recommended SANDS



[Quantum computing is and is not amazing](#)

SAND2017-7067C

Searched SAND Report by jherzer

[Heuristic approach to Satellite Range Scheduling with Bounds using Lagrangian Relaxation](#)

SAND2017-2553J

[Quantum Approximation Algorithms](#)

SAND2017-7463C

[Approximate Constraint Satisfaction in the Quantum Setting](#)

SAND2017-9140C

[The Trilinos Project Exascale Roadmap](#)

[http://prod.sandia.gov/sand\\_doc/2017/178287pe.pdf](http://prod.sandia.gov/sand_doc/2017/178287pe.pdf)

. Contributors: Mark Hoemmen, Siva Rajamanickam, Tobias Wiesner, Lois McInnes trilinos.github.io SAND2017-8287PE

[A Hierarchical Low-rank Solver for Sparse Linear Systems](#)

SAND2017-2694C

Recommend to jherzer  
the SANDs related to  
his search by SPIRE

- Analytics provide the foundation to build up an intelligent information retrieval environment:
  - Understanding the similarities of employees through clustering
  - Collecting labeled documents for Machine Learning guided by Taxonomy
  - Classifying text documents with trained classifiers
- Probabilistic graphic model associates users and documents:
  - The model reasons and infers information needed by users
  - The model will be used to predict users' information needs
- More work needed to build a better taxonomy to select labeled documents for learning
  - This is always the issue for supervised learnings
- Expanding the Personalized Information Pull and Push mechanisms to cover more MOWs and other document types. Ideally, to all employees and to all data repositories

# Questions