

Enhancing search results relevance with machine learning

Pengchu Zhang
John Herzer

June 2016

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.





The Holy Grail: Conceptual search

- ◆ Go beyond keyword search to actually search on the concept behind the customer's query
- ◆ Problems due to the imprecise nature of language
 - **Synonymy**
 - Query is for "dog" but desired document contains "canine"
 - **Polysemy**
 - Does query for "lead" refer to "leading a team" or the chemical element "Pb"
 - **Stemming**
 - Searching for "strike" vs "striking" vs "struck"
- ◆ How do we implement conceptual search?



Acquiring a corporate dictionary

- ◆ Commercial / open source dictionaries don't have words unique to our organization
- ◆ Efforts to build a corporate ontology/taxonomy/dictionary tend to fizzle
- ◆ What we need is a way to build the dictionary from our own corpus in an automated way
- ◆ Word2Vec is an unsupervised machine learning approach that lets us identify related words from the corpus



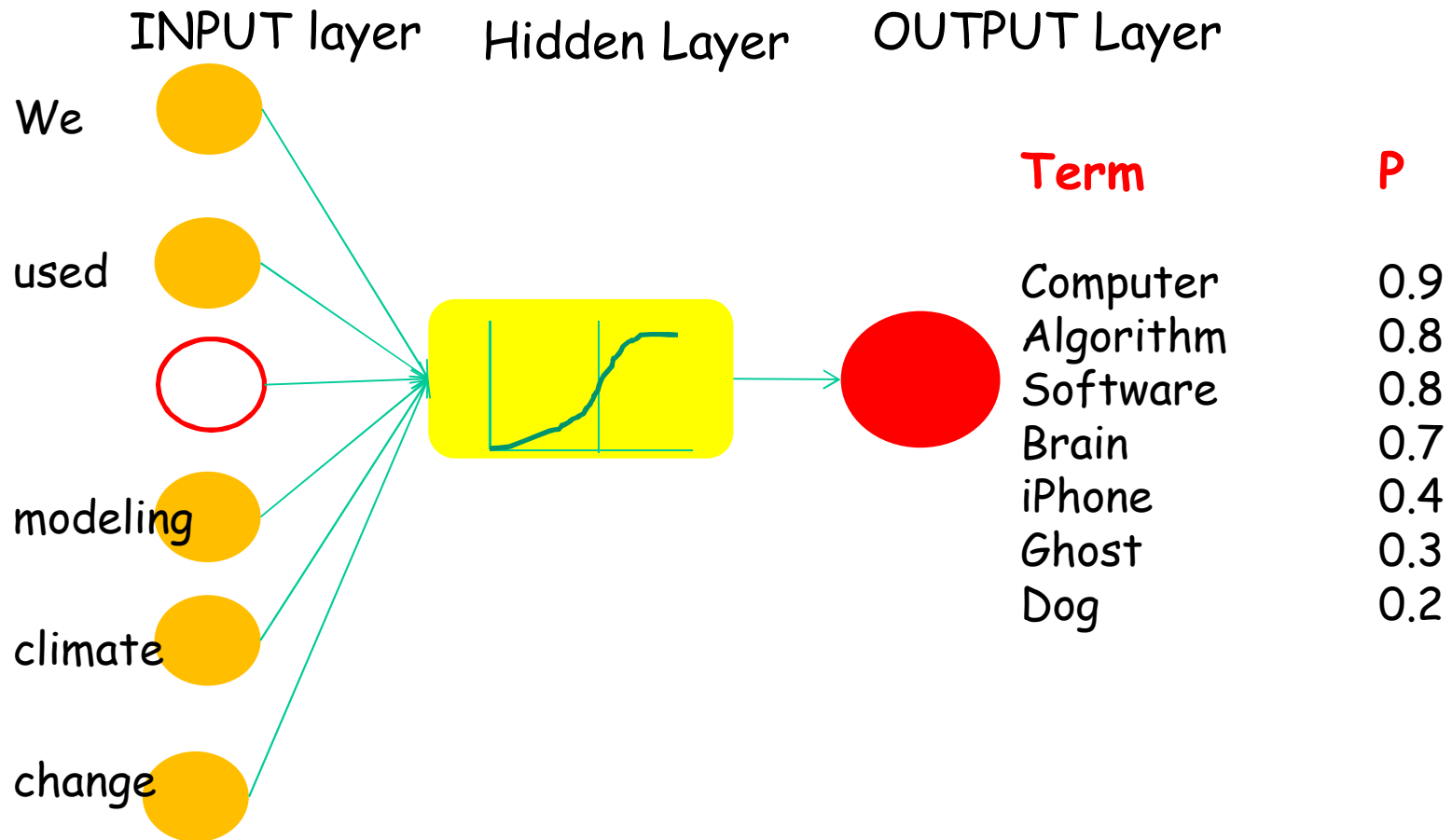


To Improve Search Results with Word2Vec

- ◆ Query with a single term or phrase
 - *"retirement"* Search engine will return documents that contain the term and rank the documents based on the frequency of "retirement";
- ◆ Word2Vec expands the query term into a set of RELATED terms or phrases
 - "retirement
Pension
Savings
Eligible
pension_fund
income_plan
Pension_plans"*
 - Search engine will return documents that contain all or some of the terms or phrases and rank the documents based on the frequencies of the set of terms or phrases, the set of terms/phrases represents as "Concept"*
- ◆ How to expand a set of RELATED terms from a single query term?



Concept of Neural Network Language Model





Problems of Word Representation in Traditional Language Models

◆ One-Hot Representations

- Simple way to encode discrete concepts, such as words
 - *"we used computer modeling climate change"*

- We =	[1 0 0 0 0 0]
- used =	[0 1 0 0 0 0]
- computer =	[0 0 1 0 0 0]
- modeling =	[0 0 0 1 0 0]
- climate =	[0 0 0 0 1 0]
- change =	[0 0 0 0 0 1]
- A one-hot encoding makes no assumption about word similarity
 - *All words are equally different from each other*
- This representation is very high in dimensions
 - *The dimensionality is the size of the vocabulary*
 - *A typical vocabulary size is 100,000*





Word2Vec in Natural Language Applications

Collobert, R., *et al.* Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* 12, 2493-2537 (2011).

Socher, R., Lin, C. C-Y., Manning, C. & Ng, A. Y. Parsing natural scenes and natural language with recursive neural networks. In *Proc. International Conference on Machine Learning* 129-136 (2011).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed representations of words and phrases and their compositionality. In *Proc. Advances in Neural Information Processing Systems* 26 3111-3119 (2013).

Sutskever, I. Vinyals, O. & Le. Q. V. Sequence to sequence learning with neural networks. In *Proc. Advances in Neural Information Processing Systems* 27 3104-3112 (2014).

Cho, K. *et al.* Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proc. Conference on Empirical Methods in Natural Language Processing* 1724-1734 (2014).

Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. In *Proc. International Conference on Learning Representations* <http://arxiv.org/abs/1409.0473> (2015).

Lecun, Y, Bengio, Y and Hinton G. Deep Learning. *Nature* 521, 436-444 (28 May 2015)



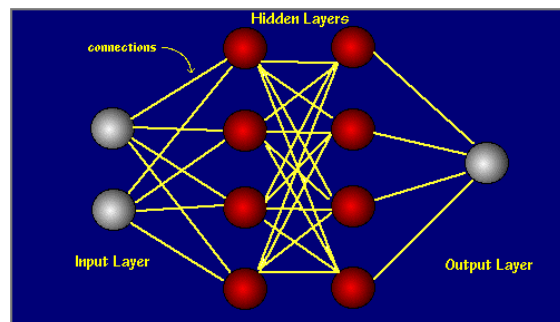
Representing Words as Vectors and Adjusting the Vector based on How Words are Used in Writing and Talking Word2Vec (Google 2013)

Initial vector

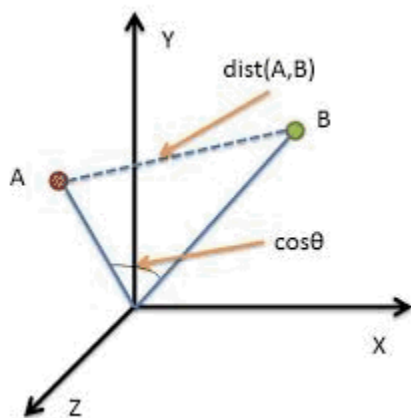
New vector

The
dog (cat)
is
walking
in
the
room

(0.12, 0.23, 0.22)
(0.32, 0.27, 0.94)
(0.18, 0.88, 0.45)
(0.23, 0.92, 0.23)
(0.77, 0.25, 0.11)
(0.12, 0.23, 0.22)
(0.41, 0.13, 0.29)



(0.12, 0.23, 0.22)
(0.62, 0.99, 0.14)
(0.18, 0.88, 0.45)
(0.23, 0.92, 0.23)
(0.77, 0.25, 0.11)
(0.12, 0.23, 0.22)
(0.41, 0.13, 0.29)

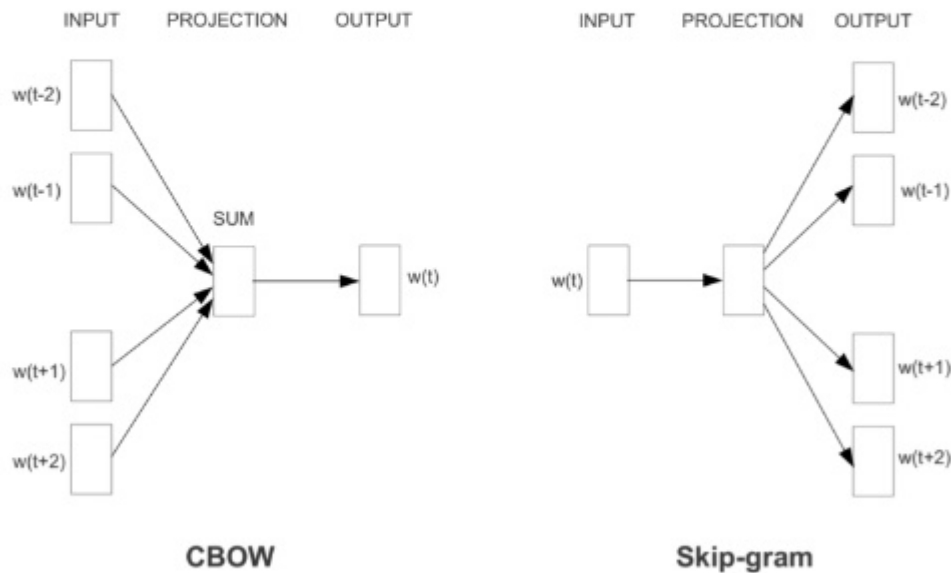


$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}.$$

Two Approaches to Build Word2Vec Models

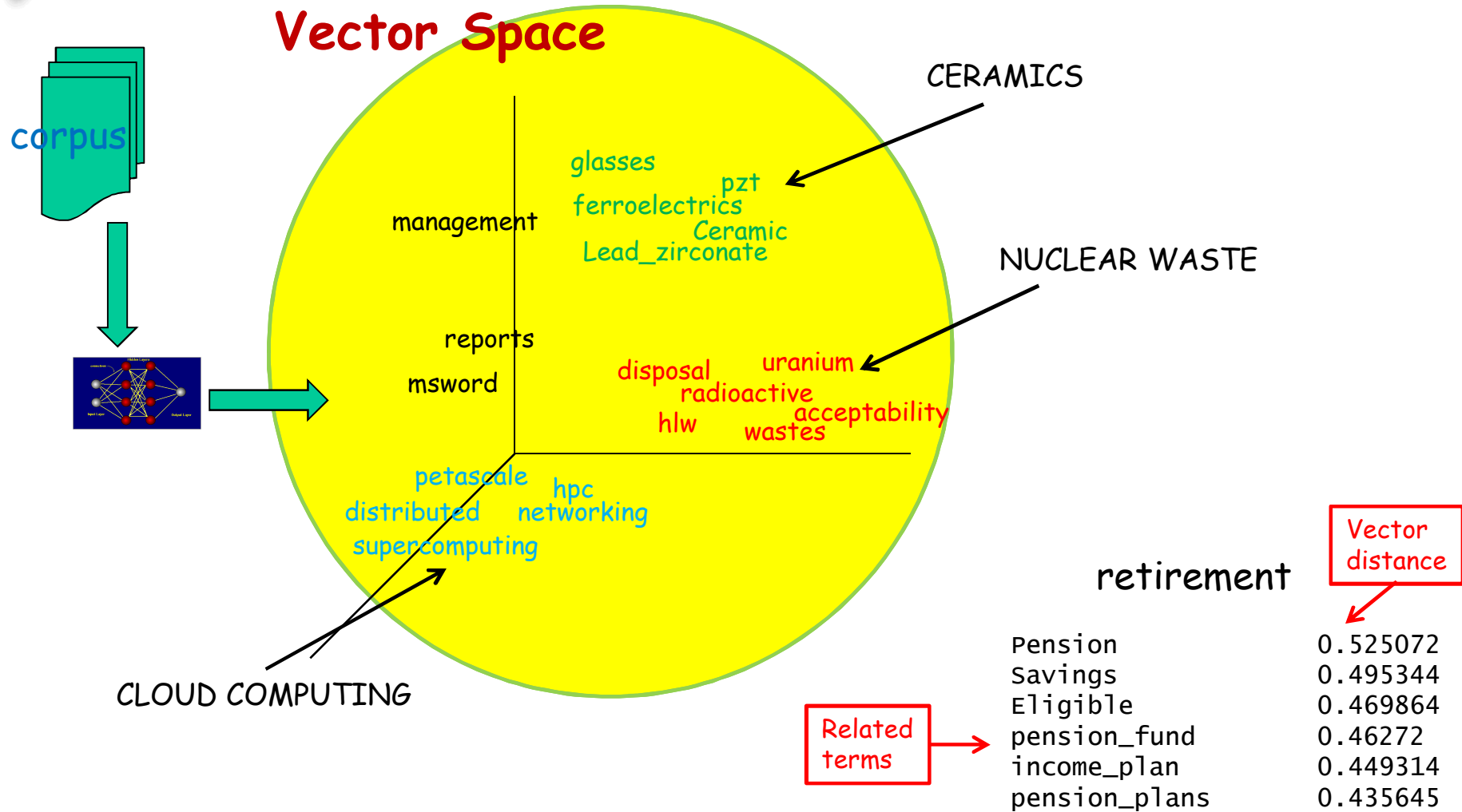
Word2Vec



Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jerrey Dean. Distributed Representations of Words and Phrases and their Compositionality. NIPS, 2013.



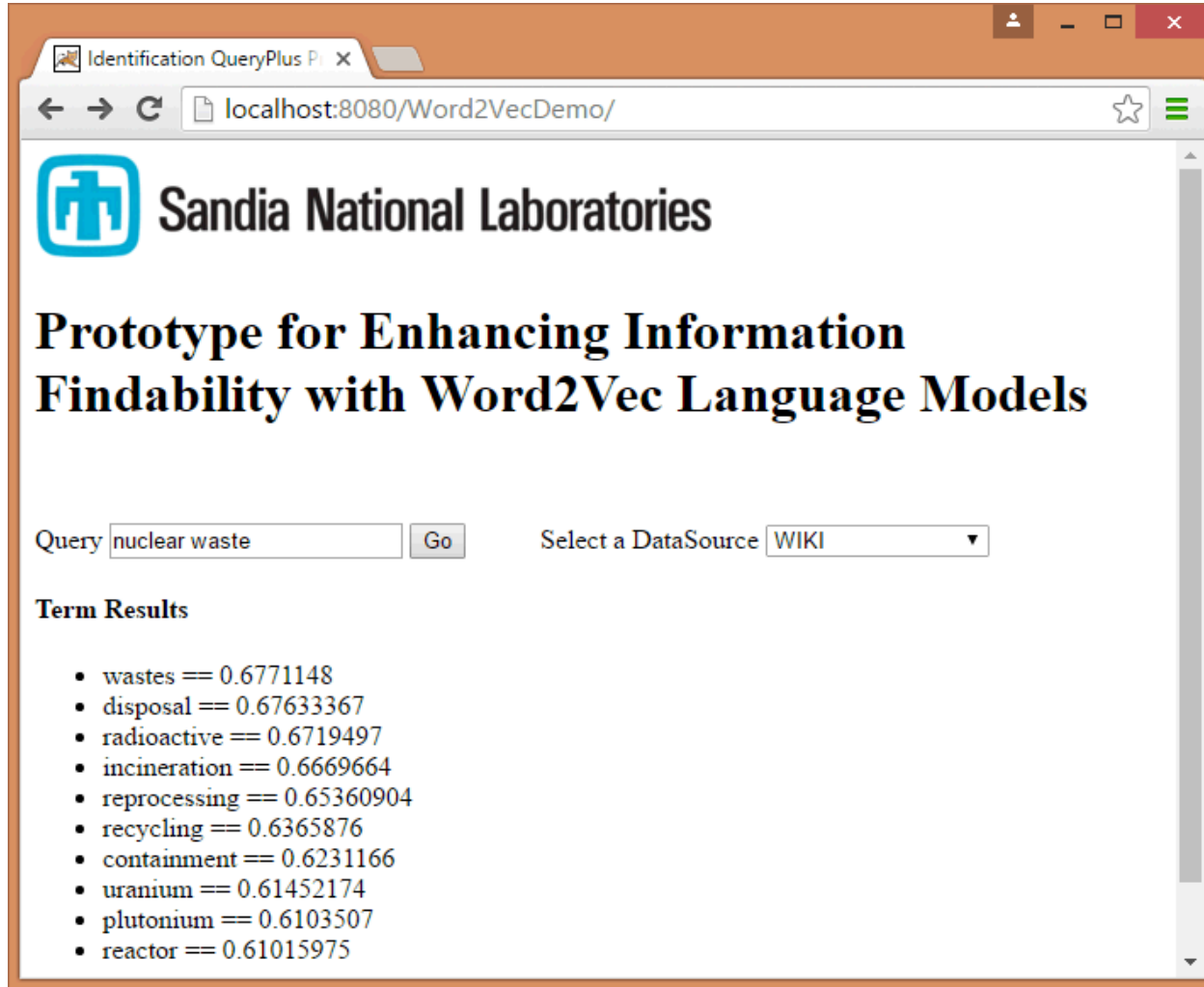
Conceptual Vector Space distribution of terms by Word2Vec



$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}.$$



Example of Term Expansion



The screenshot shows a web browser window with the address bar displaying 'localhost:8080/Word2VecDemo/'. The page features the Sandia National Laboratories logo and title. Below the title is a section for a query and data source selection. The query 'nuclear waste' is entered in a text box, and 'WIKI' is selected in a dropdown menu. A 'Go' button is next to the query box. Below this, the 'Term Results' section displays a list of terms and their associated scores.

Identification QueryPlus P x

localhost:8080/Word2VecDemo/

 Sandia National Laboratories

Prototype for Enhancing Information Findability with Word2Vec Language Models

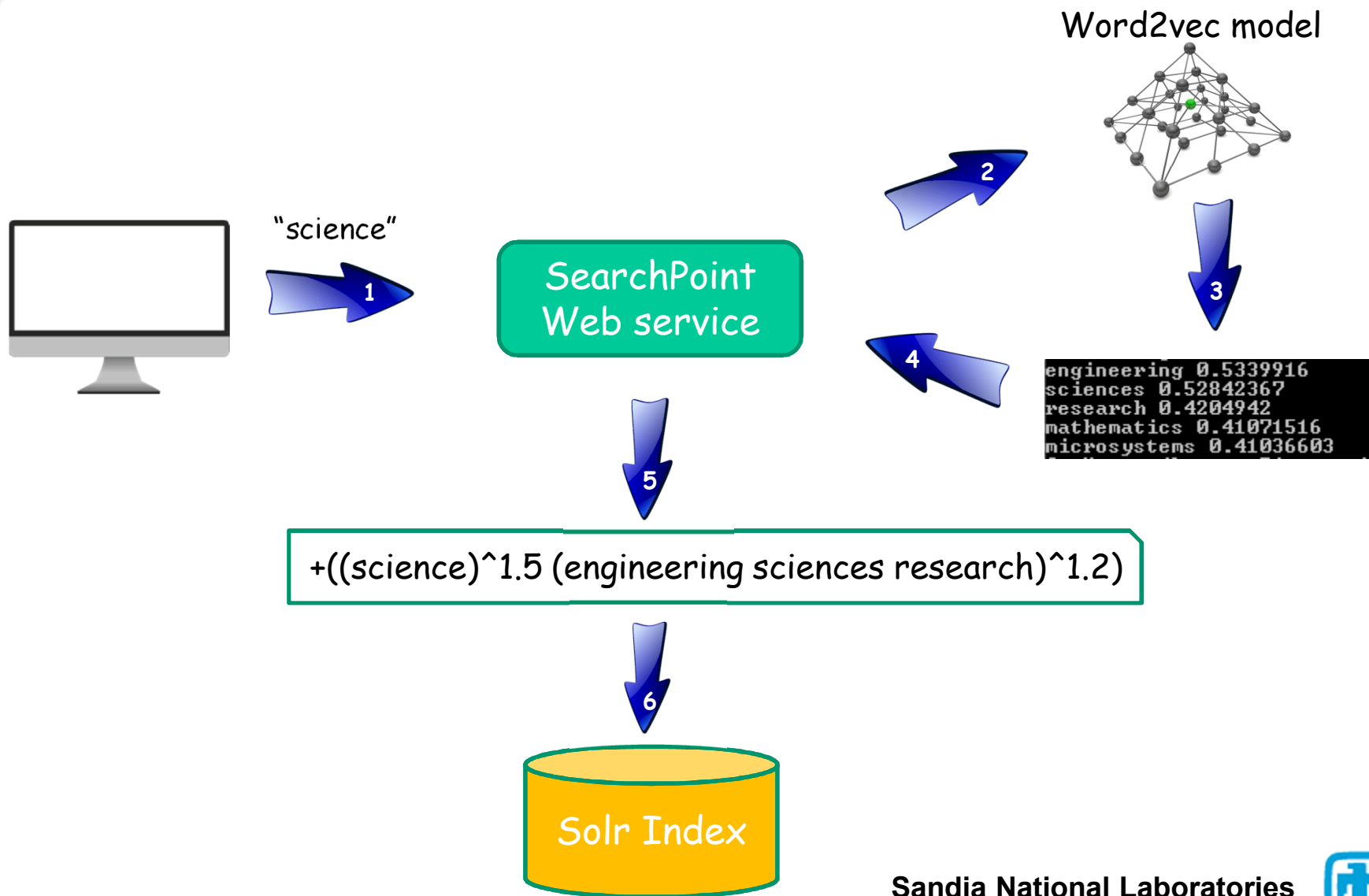
Query Select a DataSource

Term Results

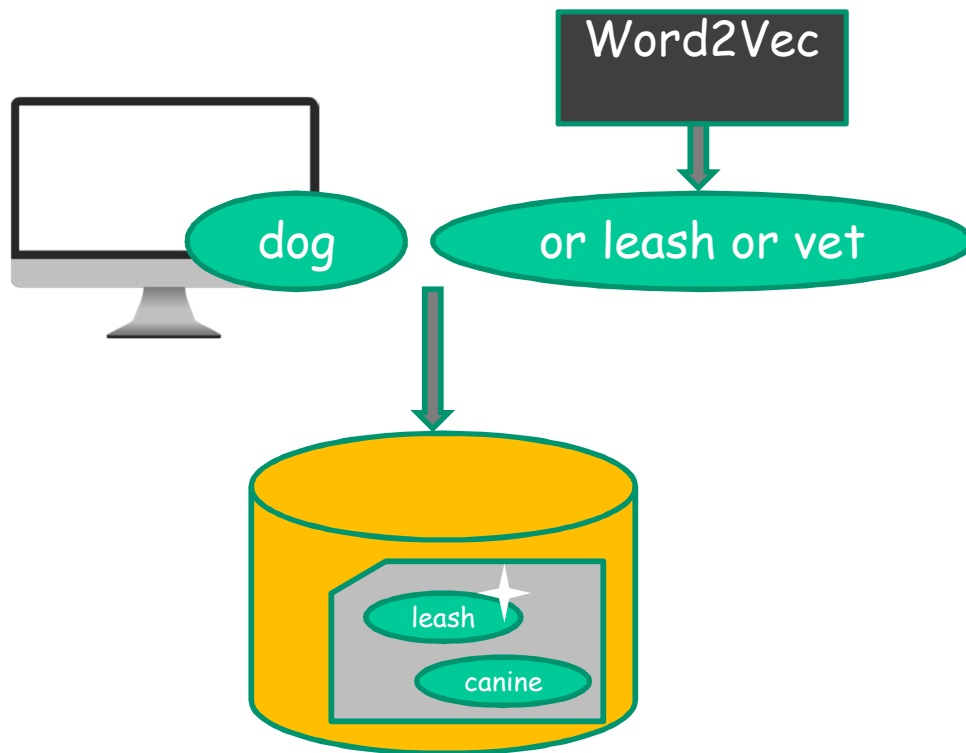
- wastes == 0.6771148
- disposal == 0.67633367
- radioactive == 0.6719497
- incineration == 0.6669664
- reprocessing == 0.65360904
- recycling == 0.6365876
- containment == 0.6231166
- uranium == 0.61452174
- plutonium == 0.6103507
- reactor == 0.61015975



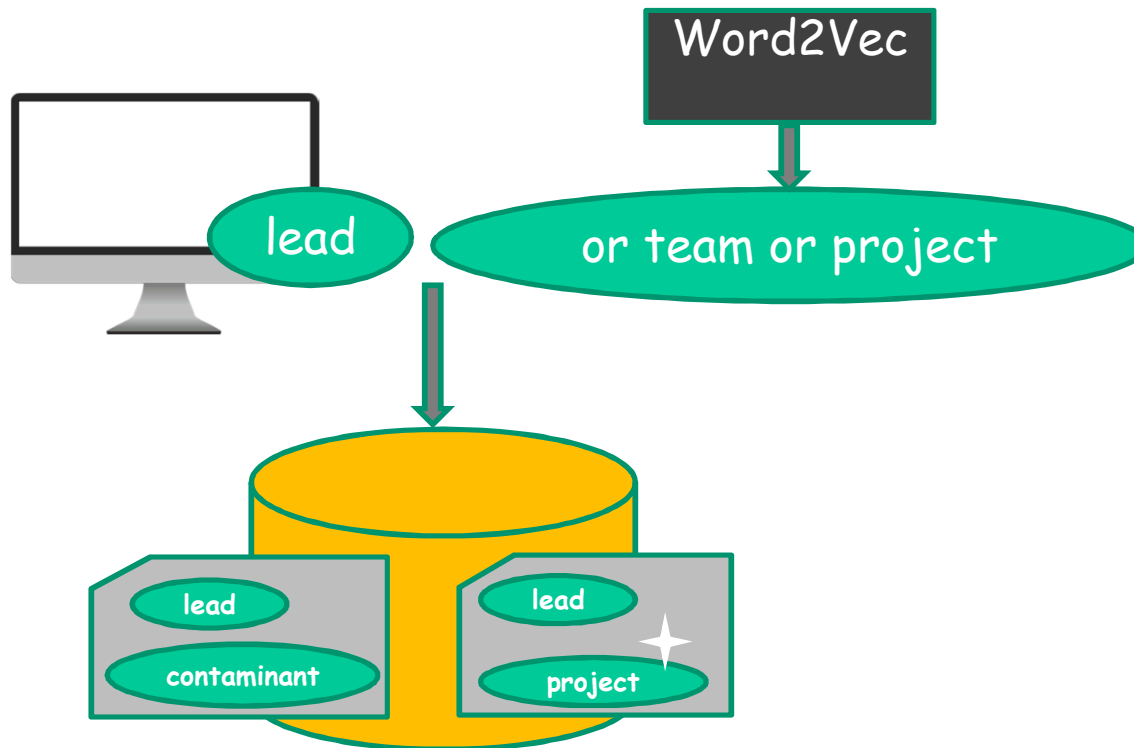
Enhancing queries with Word2Vec



Handling synonyms with Word2Vec



Handling homonyms with Word2Vec



Results without query expansion

Inside  Sandia National Laboratories

Techweb

SMM

Policies

Orgs

News

SearchPoint

SON/SRN Web

Click below to change search content (Non-UUR results possible).

SON/SRN Web

Corporate Policy System

Technical Library

FileNet

uckyballs

Search

Look Up Acronyms

[Search
Improvements
Video \(2:30\)](#)

[Provide
Feedback!](#)

Showing 1 to 25 of 62 from SON/SRN Web [Next](#)

1. [Sandia National Labs: News: Publications: Sandia Lab News: November 23, 2007 *rank = 2](#)
Sandia, NNSA, DOE, DHS, national security, **uckyballs**, yucca mountain, fit business,
augmented
<http://www.sandia.gov/LabNews/071123.html>

2. [Morning Media Report - October 31, 2007 *rank = 3](#)
arranged in interlocking pentagons "fullerenes," (and sometimes "**uckyballs**") to celebrate
<http://info.sandia.gov/corpdata/media-reports/media-daily/past-issues/2007/october/20071031.html>

3. [Morning Media Report - October 29, 2007 *rank = 7](#)
/ NNSA-DOE News Stories Sandia News Stories Experiments, Simulations Reveal Birth Secret
Of **Buckyballs**
<http://info.sandia.gov/corpdata/media-reports/media-daily/past-issues/2007/october/20071029.html>

4. [Buckyball birth observed by Sandia nanotech researcher - November 21, 2007 line=4](#)
, 2007 **Buckyball** birth observed by Sandia nanotech researcher Work confirms hypothesis of
Nobel laureate
<https://share.sandia.gov/news/resources/releases/2007/buckyball.html>

5. [abstract.5265.html line=5](#)
Molecular dynamics study of a CNT-**uckyball**-enabled energy absorptionsystem
<http://lammps.sandia.gov/abstracts/abstract.5265.html>

6. [abstract.3079.html line=6](#)
Molecular dynamics simulation of impact response of **uckyballs**
<http://lammps.sandia.gov/abstracts/abstract.3079.html>



Results with query expansion

Inside  Sandia National Laboratories

Techweb

SMM

Policies

Orgs

News

SearchPoint

SON/SRN Web

Click below to change search content (Non-UUR results possible).

SON/SRN Web

Corporate Policy System

Technical Library

FileNet

uckyballs

Search

Look Up Acronyms

[Search
Improvements
Video \(2:30\)](#)

[Provide
Feedback!](#)

Showing 1 to 25 of 381 from SON/SRN Web [Next](#)

1. [Sandia National Labs: News: Publications: Sandia Lab News: November 23, 2007 *rank = 2](#)

Sandia, NNSA, DOE, DHS, national security, **uckyballs**, yucca mountain, fit business, augmented

<http://www.sandia.gov/LabNews/071123.html>

2. [Morning Media Report - October 31, 2007 *rank = 3](#)

arranged in interlocking pentagons "**fullerenes**," (and sometimes "**uckyballs**") to celebrate

<http://info.sandia.gov/corpdata/media-reports/media-daily/past-issues/2007/october/20071031.html>

3. [Morning Media Report - October 29, 2007 *rank = 7](#)

/ NNSA-DOE News Stories Sandia News Stories Experiments, Simulations Reveal Birth Secret Of **Buckyballs**

<http://info.sandia.gov/corpdata/media-reports/media-daily/past-issues/2007/october/20071029.html>

4. [Buckyball birth observed by Sandia nanotech researcher - November 21, 2007 line=4](#)

, 12 to a **uckyball**. Departing atoms leave the same number of pentagons until the **fullerene** shrinks

<https://share.sandia.gov/news/resources/releases/2007/buckyball.html>

5. [abstract.3079.html line=5](#)

Molecular dynamics simulation of impact response of **uckyballs**.

<http://lammps.sandia.gov/abstracts/abstract.3079.html>

6. [abstract.1064.html line=6](#)

Coarse-Grained Potential Models for Phenyl-Based Molecules: II.Application to **Fullerenes** CC

<http://lammps.sandia.gov/abstracts/abstract.1064.html>



Acronyms without query expansion

Inside  Sandia National Laboratories

Techweb SMM Policies Orgs News

SearchPoint

SON/SRN Web

Click below to change search content (Non-UUR results possible).

SON/SRN Web Corporate Policy System Technical Library FileNet

sar

Search

Look Up Acronyms

[Search Improvements Video \(2:30\)](#)

[Provide Feedback!](#)

Showing 1 to 25 of 2270 from SON/SRN Web [Next](#)

1. [Synthetic Aperture Radar \(SAR\) External Homepage *rank = 1 **BESTBET](#)
Synthetic Aperture Radar (SAR), SAR, Aircrafts, Interferometry, Imagery, Surveillance, Intelligence
<http://www.sandia.gov/radar/sar.html>
2. [SAND2013-8945A *rank = 2](#)
SAND2013-8945 Superpixel segmentation using multiple SAR image products
<http://prod.sandia.gov/techlib/access-control.cgi/2013/138945a.pdf>
3. [Talent - Home - New to SAR *rank = 3](#)
Talent - Home New to SAR BrowseTab 1 of 2. PageTab 2 of 2. SearchPoint Talent - Home New to SAR I
<https://sharepoint.sandia.gov:443/sites/talent/SitePages/New%20to%20SAR.aspx>
4. [Sandia National Laboratories: Synthetic Aperture Radar \(SAR\) Imagery *rank = 4](#)
keywordsSynthetic Aperture Radar (SAR) Imagery Sandia National Laboratories Exceptional service
<http://www.sandia.gov/radar/imagery/index.html>
5. [SAR Lab News Interactive *rank = 5](#)
Archives: SAR Sandia team tests SAR at Moriarty airport July 30, 2009 à 3:40 pm A Sandia team recently
<https://info.sandia.gov/newscenter/interactive/index.php/tag/sar>
6. [SAR Development line=6](#)
Aperture Radar (SAR) OverviewSAR ProcessingImagesExample ApplicationsAircraftTechnical FirstsRequired
<https://sharepoint.sandia.gov:443/sites/talent/PreRecruiting%20Support/UUR%20Pres%202013.ppt>

Sandia National Laboratories



Acronyms with query expansion

Inside  Sandia National Laboratories

Techweb SMM Policies Orgs News

SearchPoint

SON/SRN Web

Click below to change search content (Non-UUR results possible).

SON/SRN Web	Corporate Policy System	Technical Library	FileNet
sar			
Search			
Look Up Acronyms			
Search Improvements Video (2:30) Provide Feedback!			

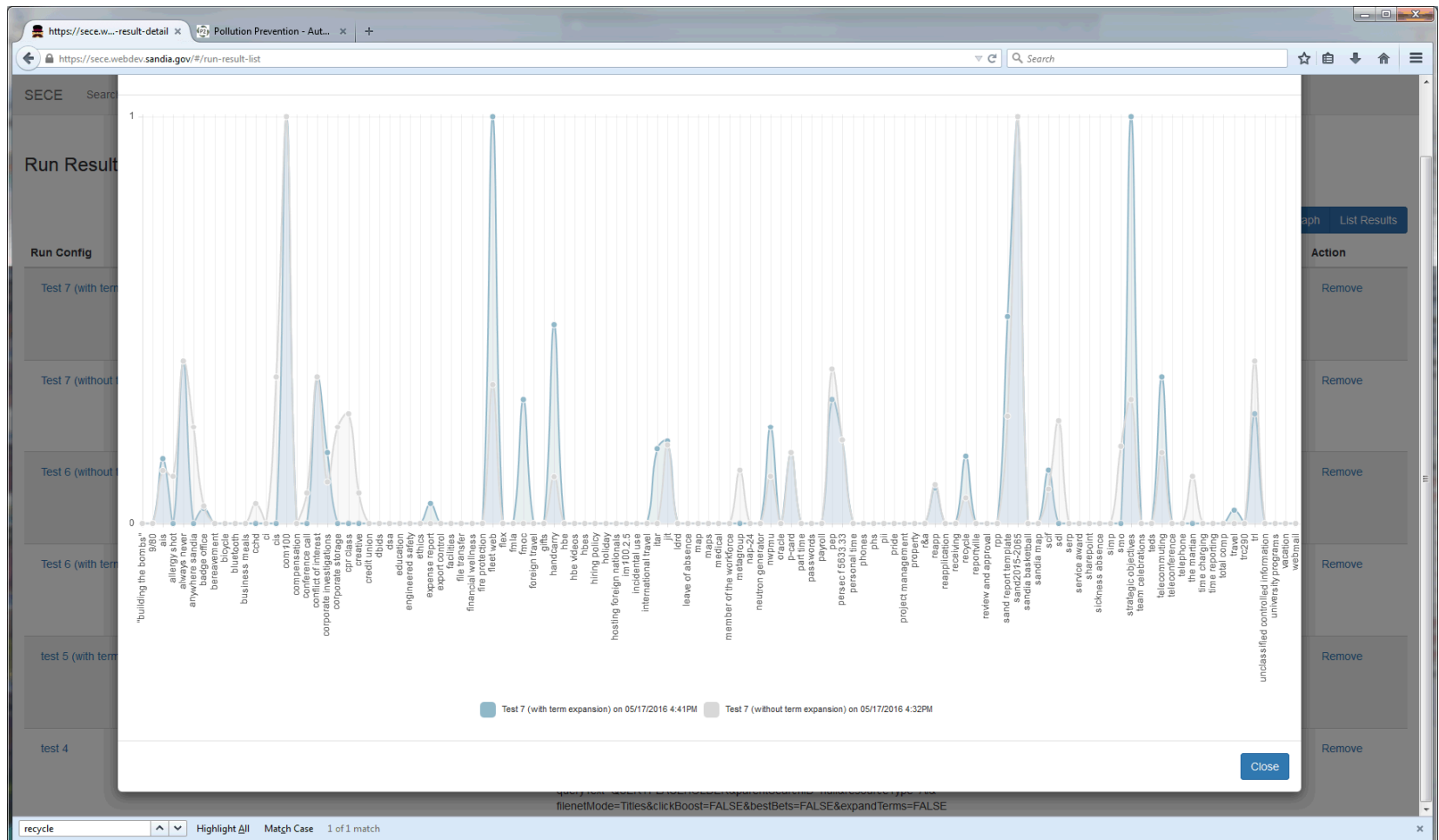
Showing 1 to 25 of 9817 from SON/SRN Web [Next](#)

1. [Synthetic Aperture Radar \(SAR\) External Homepage](#) *rank = 1 **BESTBET
Synthetic Aperture [Radar \(SAR\)](#), [SAR](#), Aircrafts, Interferometry, [Imagery](#), Surveillance, Intelligence
<http://www.sandia.gov/radar/sar.html>
2. [Talent - Home - New to SAR](#) *rank = 3
would like to welcome you to , Airborne ISR. Are you new to Synthetic Aperture [Radar \(SAR\)](#)
<https://sharepoint.sandia.gov:443/sites/talent/SitePages/New%20to%20SAR.aspx>
3. [Sandia National Laboratories: Synthetic Aperture Radar \(SAR\) Imagery](#) *rank = 4
keywordsSynthetic Aperture [Radar \(SAR\) Imagery](#) Sandia National Laboratories Exceptional service
<http://www.sandia.gov/radar/imagery/index.html>
4. [SAR Lab News Interactive](#) *rank = 5
Archives: [SAR](#) Sandia team tests [SAR](#) at Moriarty airport July 30, 2009 â 3:40 pm A Sandia team recently
<https://info.sandia.gov/newscenter/interactive/index.php/tag/sar>
5. [SAR Development](#) *rank = 8
Aperture [Radar \(SAR\)](#) OverviewSAR ProcessingImagesExample ApplicationsAircraftTechnical FirstsRequired
<https://sharepoint.sandia.gov/sites/talent/PreRecruiting%20Support/UUR%20ISR%20Pres%202013.ppt>
6. [SAR Development line=6](#)
Aperture [Radar \(SAR\)](#) OverviewSAR ProcessingImagesExample ApplicationsAircraftTechnical FirstsRequired
<https://sharepoint.sandia.gov:443/sites/talent/PreRecruiting%20Support/UUR%20Pres%202013.ppt>

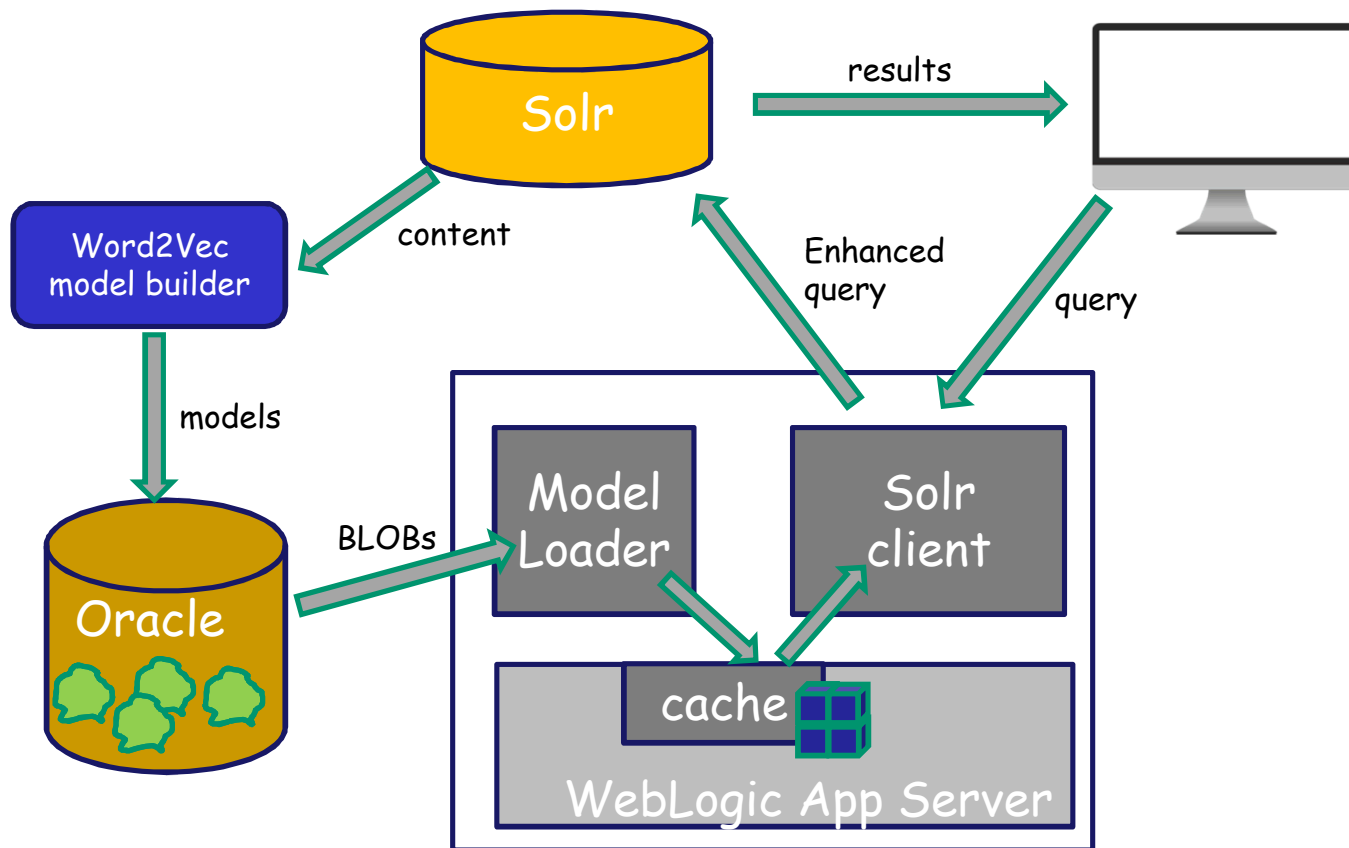




Evaluation



Integrating Word2Vec with the SearchPoint App



Performance considerations

- ◆ SearchPoint is a J2EE app deployed to Weblogic
- ◆ Model File sizes
 - solrTerms.ser 479Mb
 - solrPhrases.ser 825 Mb
- ◆ Models loaded during Servlet init
 - 69 seconds to load terms
 - Java memory usage: -Xms256m -Xmx2g
- ◆ Time to access terms for query: < 150ms





Future work

- ◆ Tackle the query disambiguation problem (polysemy)
 - Predict most likely meaning of query term based on corpus usage
 - Predict personalized meaning from customer's prior usage
- ◆ Identify phrases within the query
 - Can eliminate many non-relevant results
 - Given an expertise query "server admin java"
"server admin" or java will provide better results than server or admin or java

