SAND2017-3595C

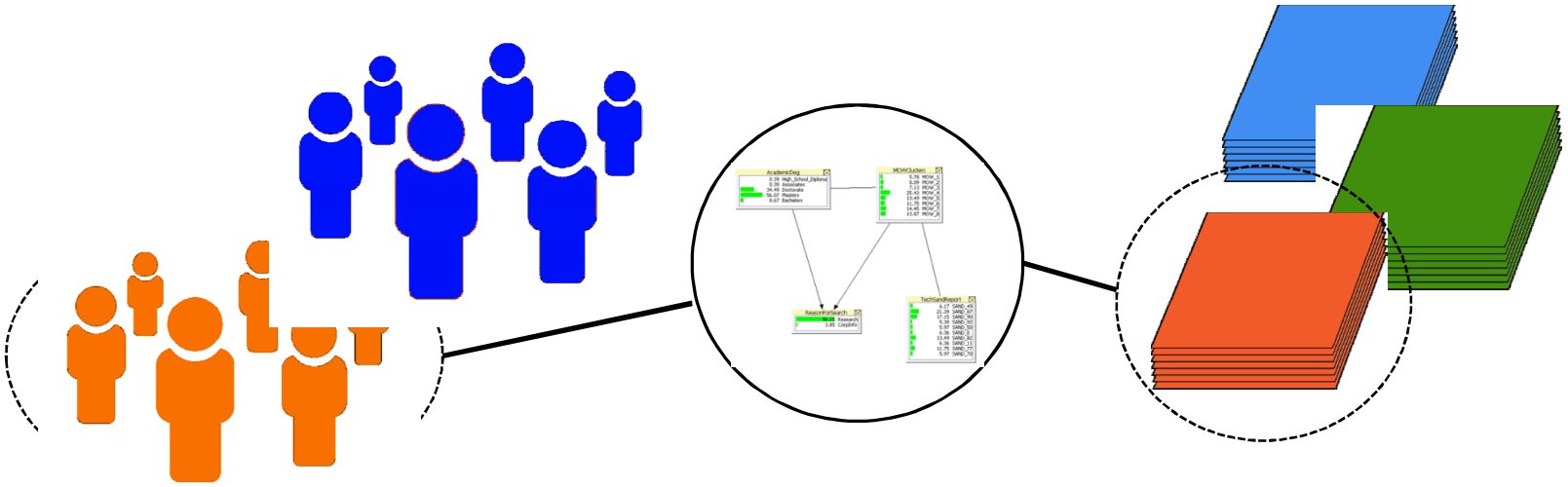
Classification of Technical Documents with Deep Learning

John Herzer
Pengchu Zhang

NLIT Summit 2017

The SPIRE project

- The Sandia Personalized Information Retrieval Environment
- Matches customers with relevant content based on their personal attributes
- To accomplish this we need to cluster customers as well as documents
- We're starting with SAND documents

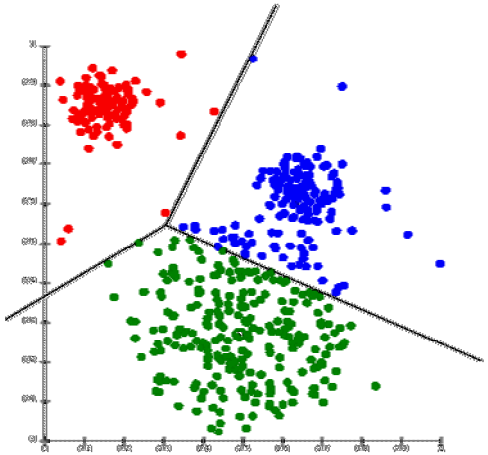


Managing 70 years worth of technical documents

- SAND Documents are official Sandia reports
- Maintain both classified and unclassified collections
- Go back to the 1950s
- Full text indexing of over 140,000 SAND Documents
- Metadata on these reports is available via online catalog
- Over 30,000 documents have not been cataloged yet
- Cataloging technical content is a very labor intensive, costly effort

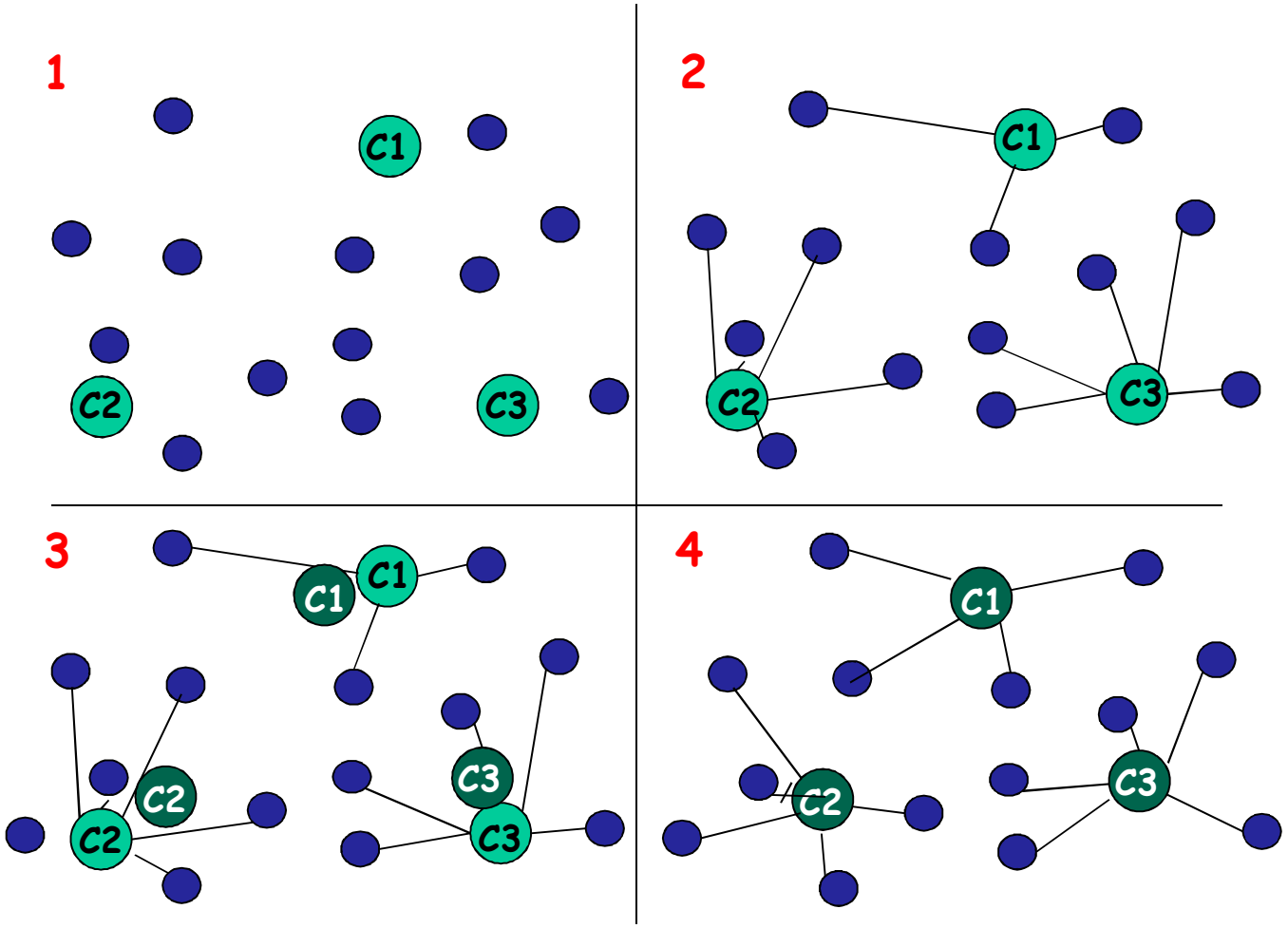
Topical Clustering

- Grouping documents such that similar ones are in the same group
- Document groupings can serve as subject categories or topics
- A type of unsupervised machine learning used in data mining
- Popular algorithms include K-means and Latent Dirichlet Allocation (LDA)



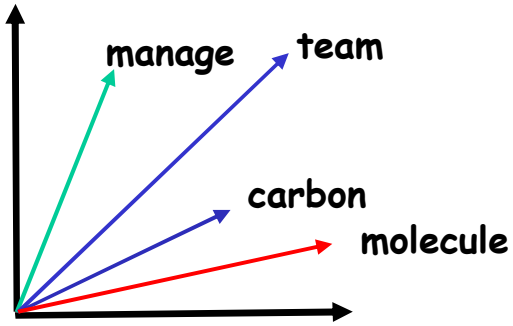
https://commons.wikimedia.org/wiki/File:KMeans-Gaussian-data.svg

Steps in K-Means clustering



1. Generate random points as centroids for each cluster
2. Assign each data point to the nearest centroid
3. Calculate the new centroid location for each cluster
4. Reassign any points that are now closer to a different centroid

Converting text into numerical vectors for clustering



manage = [-5, 0, 22, 1, 0 …]

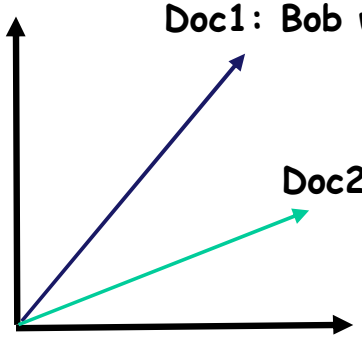
team = [3, 1, 0, 0, 15 …]

carbon = [0, 0, -8, 2, 10 …]

molecule = [2, -3, 0, 2, 20 …]

Word2vec creates a vector of numerical values for each word that represents its distribution across the corpus

Creating a numerical vector for each document in the corpus



Doc1: Bob will manage the team

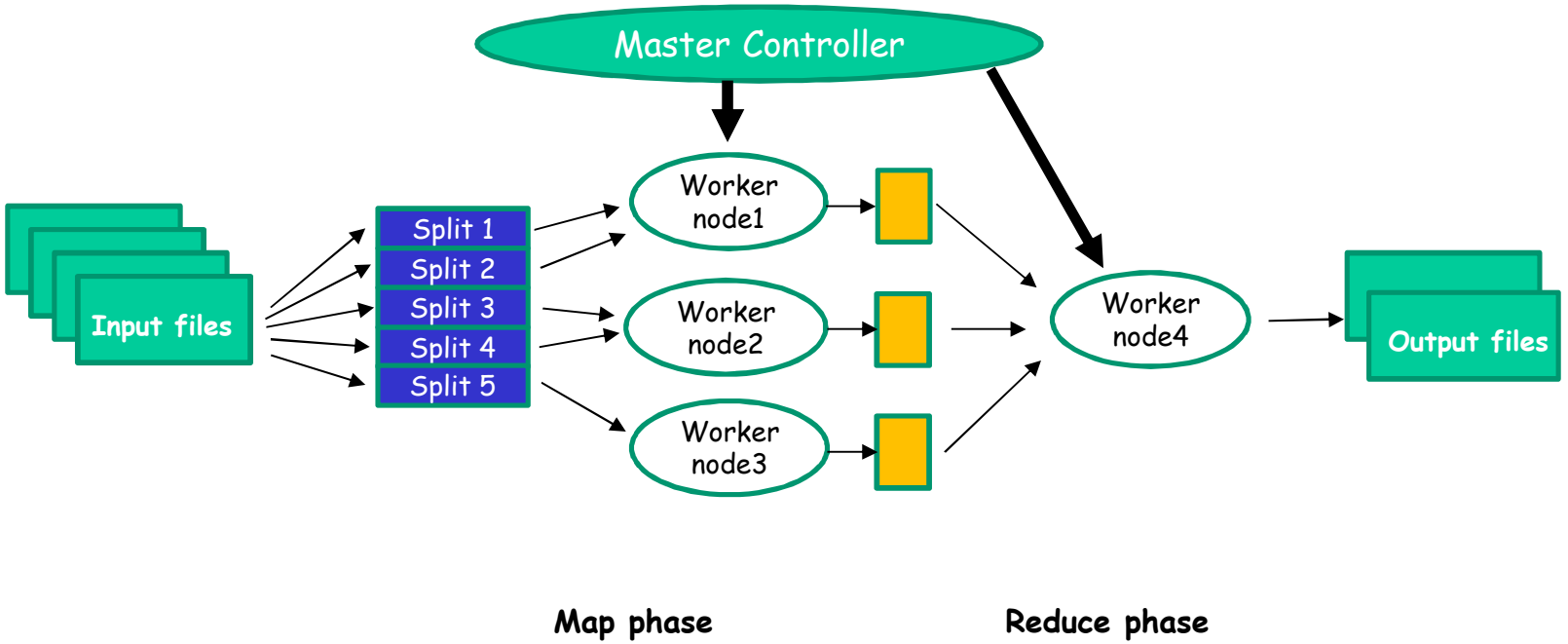
Doc1 = [2, 1, 11, 0, 7 …]

Doc2: We found the carbon molecule

Doc2 = [1, -1, -3, 2, 15 …]

By averaging the vectors for each word in a document, we can create a vector that represents that document

The Hadoop Map / Reduce Framework



Map/Reduce provides a powerful, distributed processing environment

A new approach to labeling clusters

- Generalization and Summarization are difficult tasks in AI
- Labeling clusters is particularly challenging
- In our case, we already know what the desired labels are
- We can use the Subject Category Guide to identify topics of interest
- We can take a shortcut by clustering the labels along with the documents

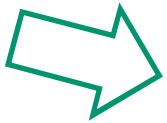
Merging Subject Category Guide Topics with SAND documents for clustering

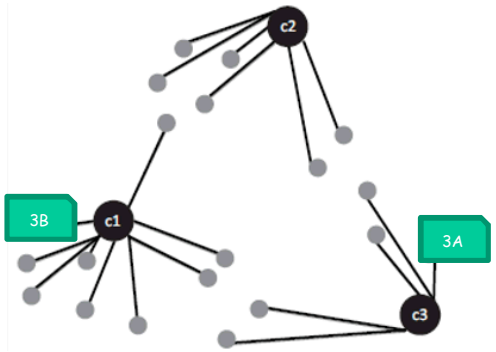
Subject Category Guide

03		Astronomy and Astrophysics	
	A	Astronomy	Observations of celestial bodies and their distances and position. Also includes Astronomical instruments.
	B	Astrophysics	Physical and chemical aspects of celestial bodies and their origin and evolution. Includes Astronomical spectroscopy; Stellar spectra; Planetary spectra.
	C	Celestial Mechanics	The motions of celestial bodies under the influence of gravity.

3B Astrophysics – Physical and chemical ...

3C Celestial Mechanics – The motions of ...





Example of an automatically labeled cluster

http://prod.sandia.gov/techlib/access-control.cgi/2016/165553pe.pdf	SAND2016-5553	Used Fuel Disposition Campaign Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, '	18
http://prod.sandia.gov/techlib/access-control.cgi/2016/165763pe.pdf	SAND2016-5763	Used Fuel Disposition Campaign KOSINA Collaboration UFD Working Group University of Nevada/Las Vegas June 7-9, 2016 San'	18
http://prod.sandia.gov/techlib/access-control.cgi/2016/165878pe.pdf	SAND2016-5878	FCR&D FY 2017 Planning Package Review Used Fuel Disposition R&D Campaign DOE-Managed HLW and SNF Research 1.02.08	18
http://prod.sandia.gov/techlib/access-control.cgi/2016/165934pe.pdf	SAND2016-5934	SNL National Laboratories is a multi-program laboratory managed and operated by SNL Corporation, a wholly owned subsidiary of Lockheed M'	18
http://prod.sandia.gov/techlib/access-control.cgi/2016/166199pe.pdf	SAND2016-6199	Joint Fuel Cycle Studies (JFCS) Fuel Cycle Alternative Working Group (FCAWG) How did we get to where we are now?	18
http://prod.sandia.gov/techlib/access-control.cgi/2016/167111pe.pdf	SAND2016-7111	Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lock'	18
http://prod.sandia.gov/techlib/access-control.cgi/2016/167941pe.pdf	SAND2016-7941	Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lo'	18
http://prod.sandia.gov/techlib/access-control.cgi/2016/168229r.pdf	SAND2016-8229	International Approaches for Nuclear Waste Disposal in Geological Formations: Report on Fifth Worldwide Review Prep	18
http://SCG-Url	8G Geology and Mineralogy	Structures, properties, and classification of rocks Rock formations and Rock constituents Mineralogy Paleontology Stratigraphy'	18
http://SCG-Url	18F Radioactivity	Radioactive decay Natural and Induced Radioactivity Interaction of Charged Particles and Radiation with Matter Radioactive fallout Fission Criticalit	18
http://SCG-Url	24C Solid Wastes Pollution and Control	Pollution by solid wastes ncluding Garbage, Scrap, Junked automobiles, Spoil, Sludge, Containers Disposal methods such as Composting, Injection well'	18
https://prod.sandia.gov/techlib/auth-required.cgi/2000/000194c.pdf	SAND2000-0194	Gas Generation in the WIPP	18
https://prod.sandia.gov/techlib/auth-required.cgi/2000/000195c.pdf	SAND2000-0195	Retardation of Dissolved Actinide Elements in the Culebra Dolomite	18
https://prod.sandia.gov/techlib/auth-required.cgi/2000/000196c.pdf	SAND2000-0196	Solubilities of Actinide Elements in WIPP Brines	18
https://prod.sandia.gov/techlib/auth-required.cgi/2000/001027c.pdf	SAND2000-1027	The WIPP Chemistry Program: Some Interesting aspects and Lessons Learned	18
https://prod.sandia.gov/techlib/auth-required.cgi/2000/001182j.pdf	SAND2000-1182	Guest Editorial: The 1996 Performance Assessment for the Waste Isolation Pilot Plant	18
https://prod.sandia.gov/techlib/auth-required.cgi/2000/002392c.pdf	SAND2000-2392	Application of Anthropogenic Analogs to the WIPP	18
https://prod.sandia.gov/techlib/auth-required.cgi/2000/002396c.pdf	SAND2000-2396	Nuclear Waste Management Workshop	18
https://prod.sandia.gov/techlib/auth-required.cgi/2000/002398c.pdf	SAND2000-2398	Success and Experiences of the Waste solation Pilot Plant (WIPP) Project	18
https://prod.sandia.gov/techlib/auth-required.cgi/2000/002953c.pdf	SAND2000-2953	Overview of Sandia Role in Nuclear Waste Disposal	18
https://prod.sandia.gov/techlib/auth-required.cgi/2001/010051c.pdf	SAND2001-0051	Overview of Components in Waste Form Degradation Model in TSPA-SR	18
https://prod.sandia.gov/techlib/auth-required.cgi/2001/010689p.pdf	SAND2001-0689	Total System Performance Assessment for Yucca Mountain	18
https://prod.sandia.gov/techlib/auth-required.cgi/2001/010829p.pdf	SAND2001-0829	Overview of WIPP Near-Field Chemistry	18
https://prod.sandia.gov/techlib/auth-required.cgi/2001/010830p.pdf	SAND2001-0830	How WIPP Solved the Gas Problem	18
https://prod.sandia.gov/techlib/auth-required.cgi/2001/010870p.pdf	SAND2001-0870	Chemical Retention Processes in the WIPP PA	18

Three labels were inserted into this cluster related to WIPP

- Geology and Mineralogy
- Radioactivity
- Solid Wastes Pollution and Control

Technical Report Classification

- Understand the document corpus
- Understand the users' interests in technical documents
- Attribute the documents in historical repositories and the newly generated documents into classes
- Ultimately, the classifiers will help to leverage the contents of technical documents to users' interests and provide the personalized document retrieval

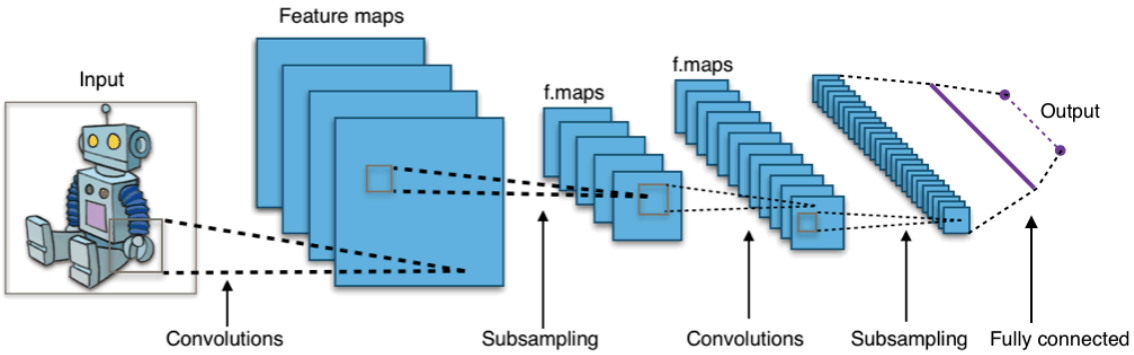
Techniques used for Text Classification

- Naïve Bayes Classifier
- TF-IDF
- Instantaneously training neural networks
- Support vector machines (SVM)
- Artificial neural networks
- K-nearest neighbor algorithms
- Decision trees
- …

Why Convolutional Neural Networks

- Based on a Deep Learning architecture
- Enable training on large data sets simultaneously
- Enable extraction of rich features from text documents with their flexible configuration
- Produce classifiers that robustly categorize documents

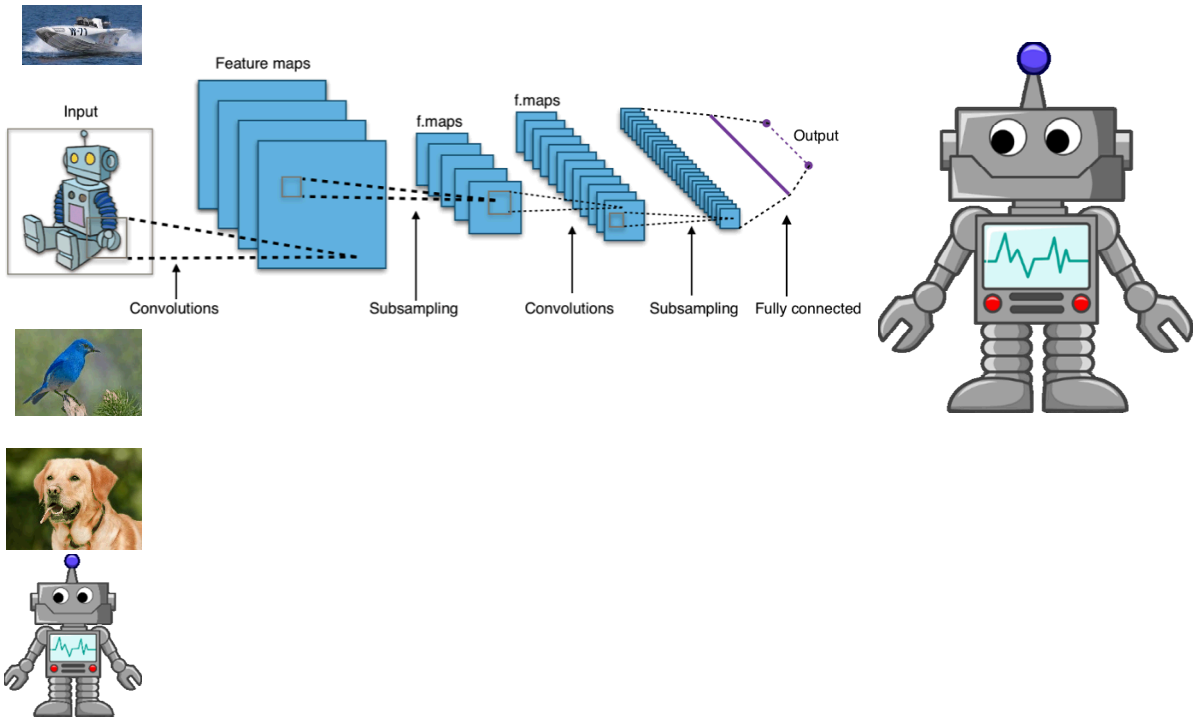
Convolutional Neural Network for Categorization



Note: some images are from Google Images with "Labeled for reuse" under "Usage rights"

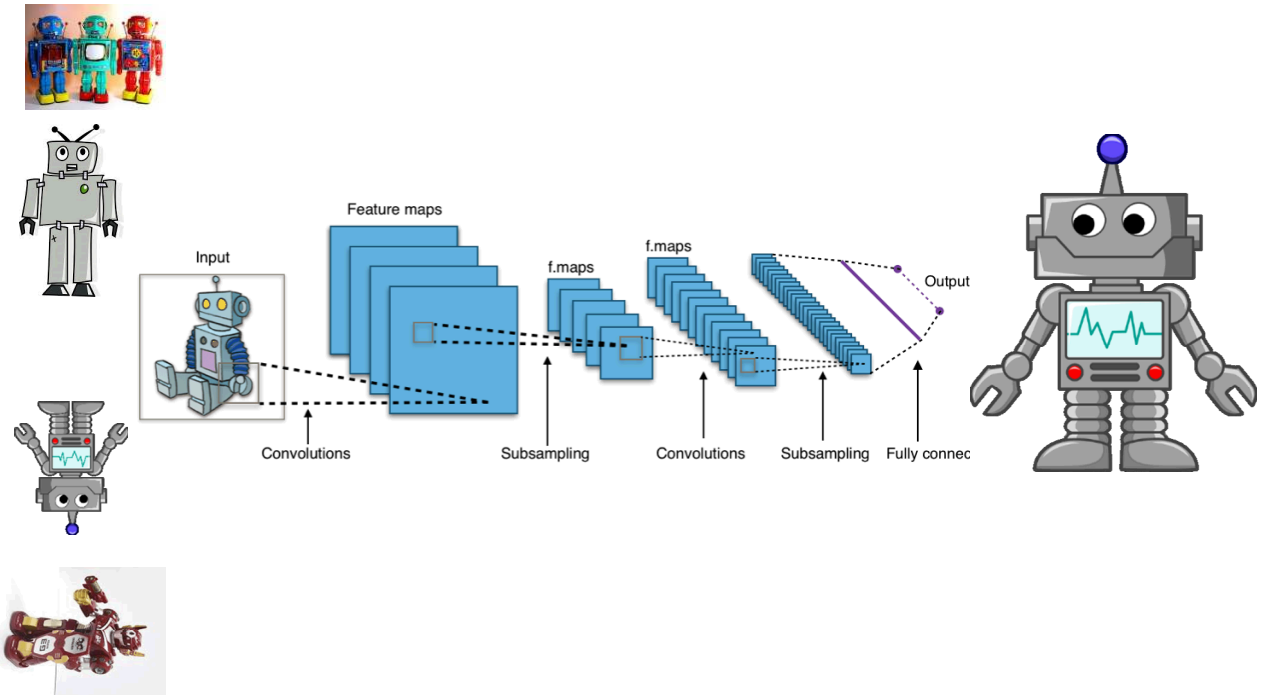
Convolutional Neural Network for Categorization

1. classify the unknown object into correct class

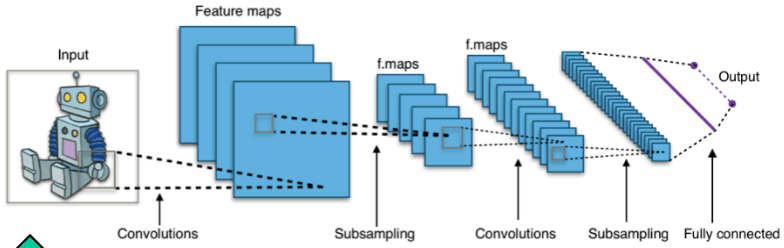


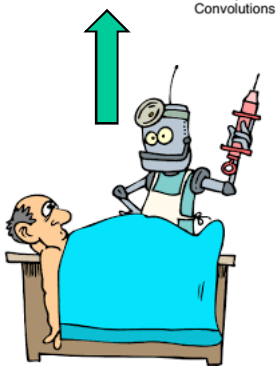
Convolutional Neural Network for Categorization

2. recognize and identify the object with multiple formats



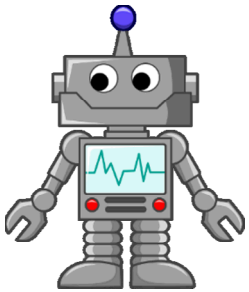
Convolutional Neural Network for Categorization
3. recognize possible classes





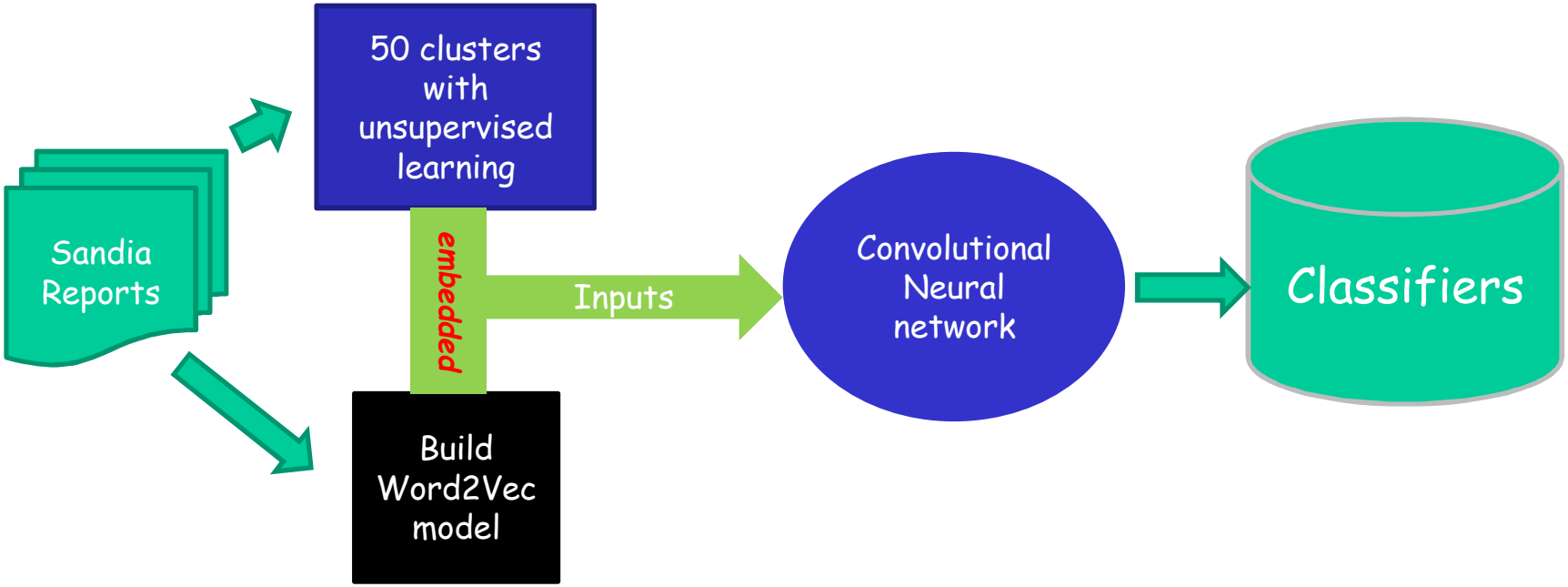
Doctor 25%

Patient 25%

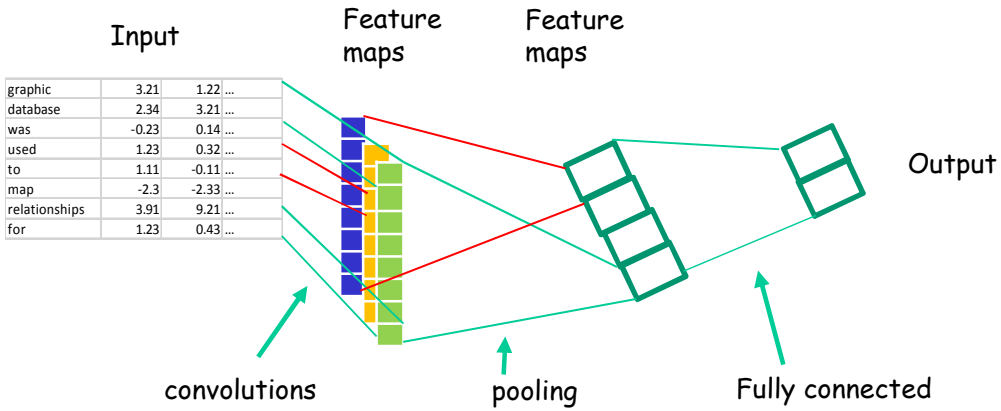


robot 15%

Build Sandia Report Classifiers with Convolutional Neural Networks



Convolutional Neural Network for Text Categorization



Data Preparation

- Collect Sandia Technical Reports, 28,000 documents
  - Include reports, final drafts for publication, conference abstracts, PowerPoint presentations
  - All in PDF format and were converted into plain text files
- Conduct data cleanup
  - Remove non alphabetic characters
  - Parse the concatenated strings into single terms
  - Remove extra white spaces
- Cluster the text files into 50 clusters
  - With Latent Dirichlet Allocation algorithm and the package Merritt
  - Save the clusters into 50 directories and the numerical directory names will be used as the labels:
    - *E.g., all files in directory "00" should be in cluster "0" and used for training Classifier "0"*

Build a Word2Vec model to Embed Words

- Concatenated all text files into one text file
- With the Google W2V algorithm to train a 200 dimension model
- Examples:

  density 0.224760 -0.128736 0.073169 0.062732 0.050471 -0.051928…

  chemical 0.128066 0.061602 0.100222 0.244458 0.037830 …

  initial 0.226996 -0.074917 0.037259 0.016864 -0.024816 -0.062900..

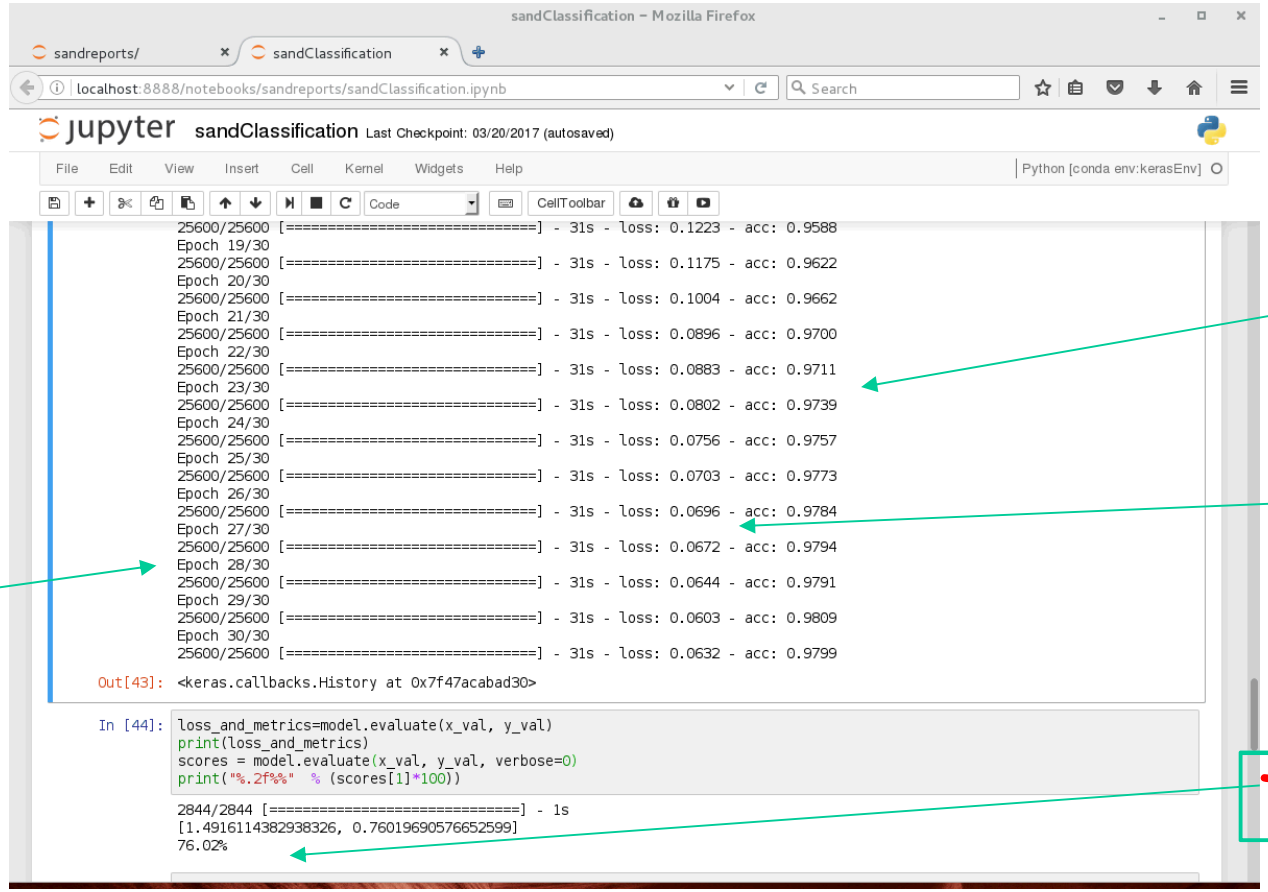
Build a Convolutional Neural Network

- Built the CNN with Theano, an open source deep learning package within Keras, a higher level wrapper
  - J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley and Y. Bengio. "Theano: A CPU and GPU Math Expression Compiler". *Proceedings of the Python for Scientific Computing Conference (SciPy) 2010. June 30 - July 3, Austin, TX* (BibTeX)
- Built 3 layers of one dimension convolutional layers with "Relu" as the activation function
- Built 3 layers of Maxim Pooling layers between the convolutional layers, also using "ReLu" as the activation function
- At the end, two fully connected layers were added, one used "Relu" and the other one used "Softmax" as the activation function
- A 0.2 drop out rate was set up to avoid "overfitting" during training

Procedures for Training the CNN

- Read in theWord2Vec model
- Read in all text files with the directory names as the labels
- Tokenized the terms in text files
- Divided the data into training and validation sets: 80% vs. 20%
- Embedded the tokens with the Word2Vec model to form a "not trainable" layer in the CNN
- Trained the CNN with 30 epochs
  - Based on the decrease in loss function and increase in accuracy

Results from Training



Epoch

accuracy

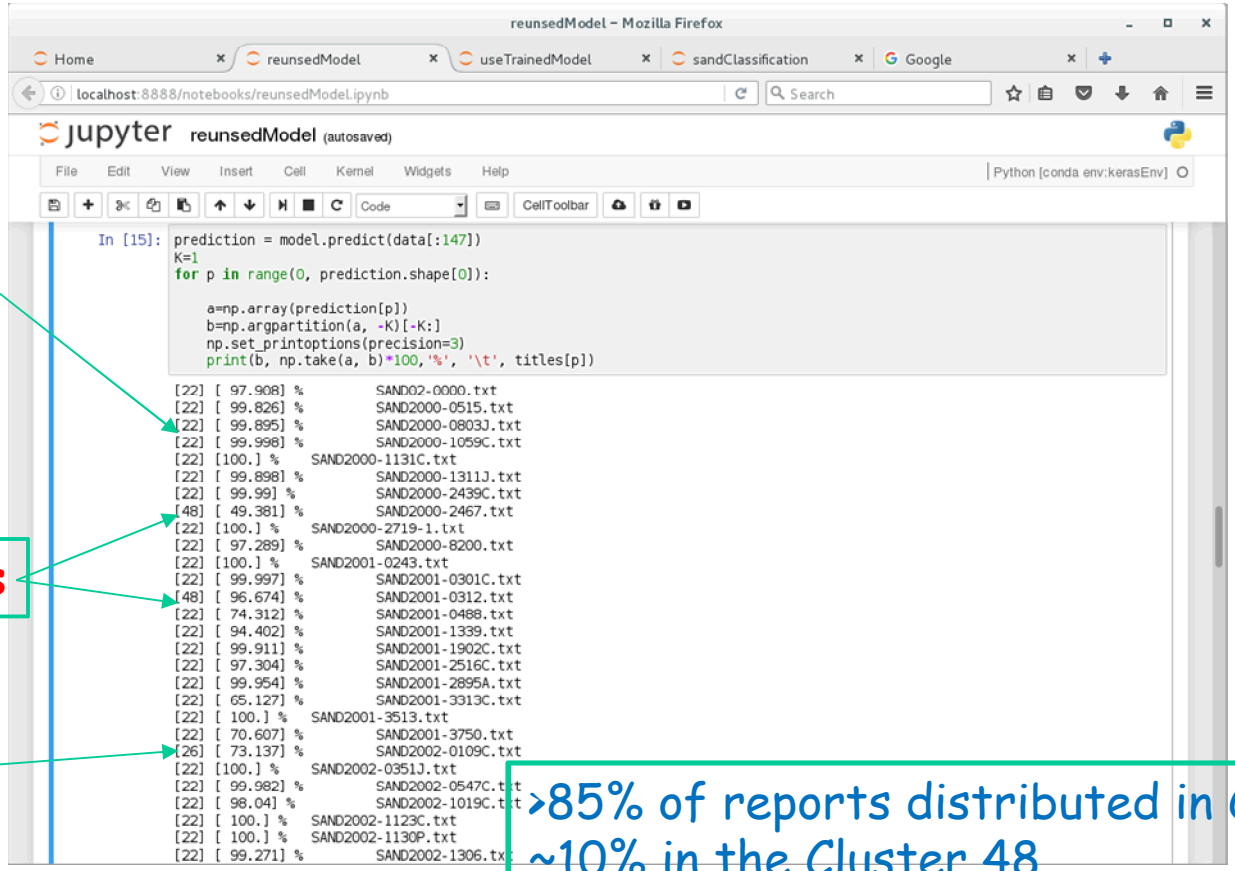
Loss

Training Score

Validate the Classifiers

- Shuffle all Sandia Technical Reports
- Cluster them into 100 clusters
- What we expect:
  - Majority of the reports in one cluster should be classified into one class
  - The secondary class should be consistent
    - *One report may be in the secondary class*
    - *One report may be assigned to the primary and secondary classes*

Classes for Reports in One Cluster



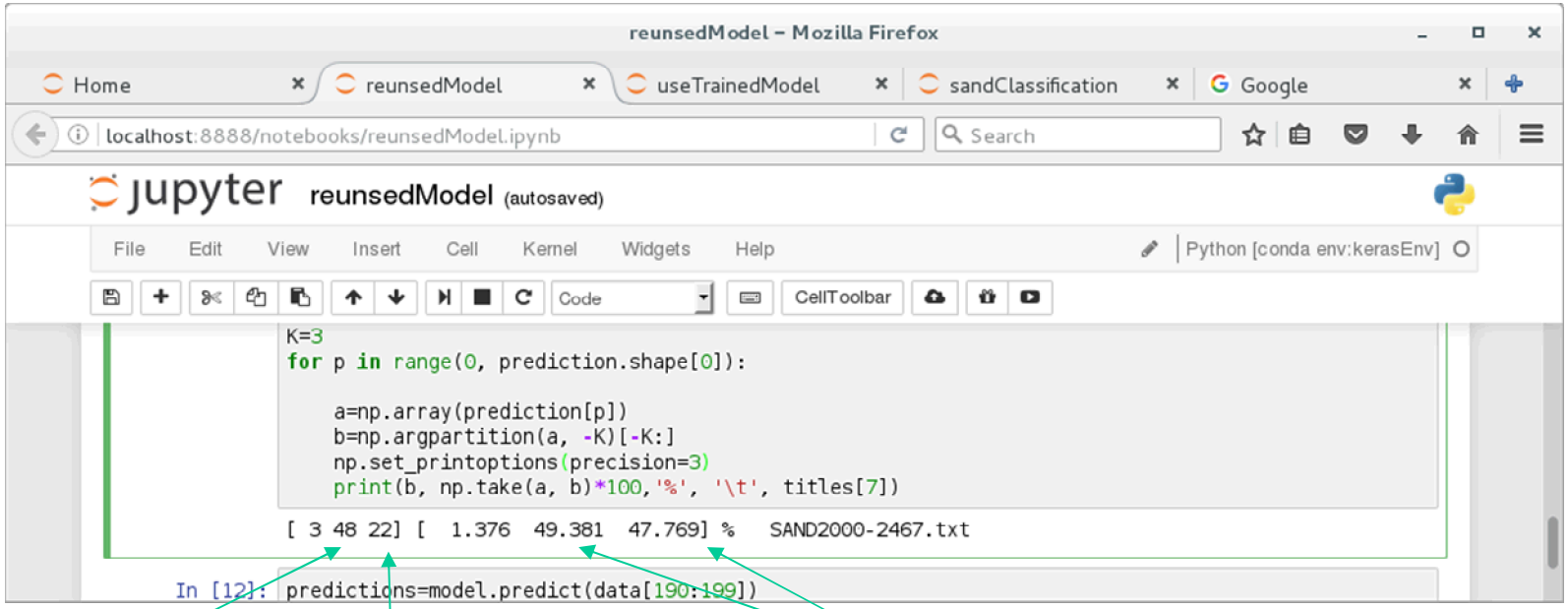
Primary class

Secondary class

Others

>85% of reports distributed in Class 22
~10% in the Cluster 48

Report is assigned to primary and secondary classes



Secondary class | **Primary class** | **Probabilities**

Conclusions

- CNN is capable of building text classifiers for large data collections
- Accuracy for training reached 98%
- Validation accuracies are above 80%
- Classifiers consistently categorize text documents
- Continue investigating the optimal number of classes for a data repository
- Automatically update the classifiers with new data added to the data collection
- Investigate the potential to automate class labeling

Future Applications

- Classifiers will be used to identify the users' interests in technical reports
- Classifiers will be used to classify the new documents into proper class(es)
- Recommendations will be made to the users for the technical reports in the classes most interesting to them when:
  - Users search for reports
  - New reports become available