Gurinder Gosal

Vancouver, Canada
gosal.gps@gmail.com
+1-778-325-8255

PROFESSIONAL PROFILE

- An experienced professional having rich experience in developing and deploying many **NLP, Data Science and Semantic Technology** projects. I have been involved in data integration, extraction, validation, analysis and presentation tasks for implementing data-driven solutions using advanced machine learning and NLP approaches.
- Led and managed many NLP, data science and semantic technology projects with involvement in all phases such as planning, stakeholder engagement, requirement analysis, design, development, integration and population, implementation and deployment, evaluation, evolution and versioning, and documentation thus influencing many strategic and operational decisions and adding business value.
- Practical experience in **Natural Language Processing and Understanding**. Developed and led many NLP tasks that were part of breakthrough natural language understanding technology used in **Bot Colony** (**http://botcolony.com)**.
- Identified opportunities in large, rich data sets and conceiving, designing and implementing **statistical / predictive** models and **machine learning algorithms** utilizing diverse sources of data to provide actionable insights and **data-driven solutions.**
- Have been using **advanced machine learning and statistical approaches** in model building, development and deployment in learning-based projects with reproducible and scalable data workflows.
- Reinforced **best data quality practices** by analyzing, researching and recommending the improvements in processes for data capture, storing, processing and dissemination enabling better system performance.
- Well versed in overall **ontology curation and development** process, especially in the bioinformatics domain. Developed many ontology-driven solutions using ontologies.
- Expertise in **Python, SQL, Java, SPARQL, OWL, Jena and R** having used these in the Ontological, NLP, Semantic and Data Analytic works in different domains.
- Involved in **writing research proposals**, preparing and publishing manuscripts and **participated** in conferences, workshops, scientific meetings and discussions and have **publications** in reputed journals such as Nature food journal, PLOS ONE, PLOS Computational Biology.
- Keen learner and always willing to learn latest techniques and new technologies in the domain. Therefore, I have taken many advanced data science/ ML/ NLP courses and got certifications to keep me up to date.

RELEVANT COURSEWORK AND CERTIFICATIONS

PhD in Computer Science
Feb. 2015- Dec. 2018
*Course Work: Concordia University, **Canada***
*University of Georgia, **USA***
*Thesis Work: Punjabi University, **India***

Master of Science in Computer Science
Aug. 2008- May 2010
*University of Georgia, **USA***

Relevant MS and PhD courses coursework:
(University of Georgia, USA)
Machine Learning,
Advanced Databases,
Software Engineering (all - **Grade A**),
Advanced Information Systems (**A-**),
(Concordia University, Canada)
Statistical Natural Language Processing,
Natural Language Understanding,
Semantic Computing

SKILL SET (CURRENT) AND CERTIFICATIONS

Data Science tools/resources	Scikit-learn, SciPy, NumPy, Pandas, Jupityr notebook, Matplotlib, Spacy, Tensorflow, Keras, MS-Excel
NLP tools/resources	Python, R, NLTK, OpenNLP, Spacy, GATE/Wikipedia, VerbNet, WordNet, SentiWordNet, Oxford corpora
Development languages/ tools	Python, R-Programming, JAVA, Eclipse IDE, PyCharm IDE, RStudio, SCRUM, GitHub, Freedcamp
Ontology/ Semantic Web tools	Apache Jena, OWL2, Protégé, SPARQL**,** RDF, ROBOT, Ontofox, OLS, RDFLib
Database technologies	SQL, Versant Object Database, MS-Access
NLP applications developed	(Lexicon, Rule-based, Statistical, ML-based and hybrid)

NLP Tasks - Text Mining - Information retrieval: Keyword and domain based document collection, Search (Keyword, Semantic), Information Extraction: Entity recognition -Biological and Other, Relation Extraction, Collocation extraction, Phrase Mining and Entity linking, **Text similarity**: Sentence similarity, phrase similarity, document similarity, **Template matching**: constraint based language based pattern matching)
Text classification: Text Clustering, Categorization and classification, Topic modelling, **Text Summarizing**: Ontological based, **Semantic annotation**: semantic labelling - Quality assignment, Domain assignment, Semantic Feature Ontology, **Sentiment analysis**: review analysis, twitter data analysis, **Core tasks**: Semantic feature support for World Sense Disambiguation, PP disambiguation, Question answering Ontology driven text mining and annotation in biomedical domain
NLP processes: Tokenizing, stemming, lemmatizing, chunk parsing, parsing POS tagging, Context free grammars, semantic labelling

DS Tasks: Predictive, Descriptive and Diagnostic Analytics using Clustering algorithms, Logistic regression, Decision Tree, Support Vector Machines, Bayesian Networks, Maximum Entropy, Conditional Random Field, Neural Networks/Deep Learning - PCA, Word-to-vec, RNN, CNN, Latent Semantic Indexing (LSI)

Relevant Certifications:

- Data Science Certification — **John Hopkins University, USA**
- The data science course: complete data science bootcamp — **Udemy**

PROFESSIONAL APPOINTMENTS

NLP and Data Science PDF Researcher **April 2017 – till date**
University of British Columbia (UBC) at BC Centre for Disease Control, Vancouver, Canada

- Working as a key researcher and strategist in the areas of data science, NLP and semantic technology to develop machine learning and NLP competencies in healthcare domain working closely with an interdisciplinary team of public health researchers, epidemiologists, medical microbiologists, laboratory technologists and software developers involved in developing open source AI-driven tools and knowledge frameworks that improve data sharing and re-use in public health.
- Leading and working on many data-driven projects, involving Canadian and international collaborators, that have a very important role defined in our Laboratory's current (The US FDA's GenomeTrakr, FDA's Resistome Tracker, The USDA Food Central database, EnteroBase, CINECA) and future collaborations.
- Involved in providing metadata harmonization tools and resources in The Canadian COVID Genomics Network (CanCOGeN), a newly formed initiative backed by $40 million in federal funding, is led by Genome Canada, in partnership with the six regional Genome Centres, national and provincial public health labs, sequencing centres and academic institutions. **CanCOGeN** will coordinate and scale up existing genomics-based **COVID-19 research in Canada and internationally** in order to accelerate public impact.

- Involved in providing metadata harmonization tools and resources in **CORRE - The COVID-19 Rapid Evidence Reviews Group** is a collaborative effort between ISARIC and several other international partners that aims to provide high quality rapid evidence synthesis for use in COVID-19 research activities.
- Implemented the analyzed and designed data science, NLP and Semantic technology projects with extensive code design and production, majority of which is available as open source code. Involved in the delivery and deployment of the projects available on BioConda and GitHub **(https://github.com/Public-Health-Bioinformatics/).** On the semantic technology front, I am a key member of the teams developing and curating ontologies - **FoodOn (http://foodon.org)**, a consortium-driven project to build a comprehensive and easily accessible global farm-to-fork ontology about food.
- **Key projects (as a team leader): Data Science**- Classification of NCBI bio-samples (team leader), Data harmonization of public health laboratory test result codes (team leader); **NLP**- LexMapr: A Text Mining Tool for Translating Short Biomedical Specimen Descriptions into Ontology Terms (team leader), **Semantic Technologies**- Term recommendation system from free-text for food ontologies (team leader).
- Involved in writing (and helping) research proposals (such as NSERC, MSFHR, GenomeBC, GenomeCanada, CINECA). Preparing and publishing manuscripts. Participated in conferences, workshops, scientific meetings and discussions (ISMB/ECCB, IC-Foods, ICBO). Publications in reputed journals such as **Nature** food journal, **PLOS** Computational Biology).

Senior Research Engineer (Data Science and NLP) **Nov 2012 to March 2017***
Research Engineer (Data Science and NLP) **April 2011 to Nov 2012**
* Intermediate (Leave for teaching/PhD):
Faculty position, Punjabi University Patiala Jan 2014 to Dec 2016
North Side Inc., Montreal, Canada

- Developed and led many NLP, data science and semantic technology tasks that were part of breakthrough natural language understanding technology used in **Bot Colony (http://botcolony.com/)**, a video game - the first application where a person can speak freely with a machine. Managed and participated in multiple modules of Bot Colony game in different roles as well assisted teams working on multiple projects involving complex NLP, speech, dialogue research and development tasks in North Side Inc.

Software Research Professional II (NLP and Ontologies) **June 2010 to April 2011**
University of Georgia, Athens, Georgia, USA

- As a key NLP and semantic technology developer and researcher, I analyzed, designed, developed, tested, implemented and evaluated a very large and comprehensive ontology, **ProKinO (https://ieeexplore.ieee.org/abstract/document/6120500)** that serves as a useful and efficient representation of the integrated knowledge about the complex proteins, called protein kinases, which are intimately involved in the genesis and behavior of cancer cells (**http://vulcan.cs.uga.edu/prokino/about/prokino**).
- Performed a large-scale integrative analysis of protein kinase data integrated from multiple disparate sources to investigate its role in cancer and published the results in reputed journal **Plos One (https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028782)**.

PROFESSIONAL DEVELOPMENT / PROJECTS

These projects have been grouped together as per their category (**for brevity, restricted to 6-7 projects described briefly for each category**) and might not be necessarily following the overall chronological order.

A) AI-DRIVEN (MACHINE LEARNING / DATA SCIENCE) PROJECTS

Technologies: For the following data science related projects, the technologies used are Python, Scikit-learn, SciPy, NumPy, Pandas, Jupyter notebook, Matplotlib, Spacy, Tenserflow, Keras, MS-Excel, NLTK, Java, R-Programming, MYSQL, PyCharm IDE and other NLP tools.

1. **Deep-learning based classification of COVID-19 literature using doc2vec model**
(BC Centre for Disease Control (UBC), Canada) Mar' 20 – June'20
    - As part of the **The Canadian COVID Genomics Network (CanCOGeN)** initiative, the deep learned doc2vec model is used to classify the COVID-19 literature using the titles and abstracts of

the papers into multiple classes such as diagnosis, treatment, clinical manifestation, case report, comorbidities and so on.

2. **Intelligent document retrieval system based on semantically enriched research questions related to COVID-19**
   (BC Centre for Disease Control (UBC), Canada) Mar' 20 – June'20
   - This project is part of providing the metadata harmonization tools and resources in **CORRE - The COVID-19 Rapid Evidence Reviews Group**. The research question related to COVID research are semantically enriched with knowledge from knowledge resources and intelligent query is used to search the literature that are also semantically processed. Based on the semantic similarity of text, the relevant and pertinent literature id retrieved from the sources.

3. **Clustering public health laboratory test result data**
   (BC Centre for Disease Control (UBC), Canada) Mar' 19 – Feb'20
   - The developed tool performs clustering of BCCDC Public Health Laboratory test result codes on selected entities, such as pathogens, test types, and disease names. The tool is helping **BCCDC Public Health Laboratory** in clustering the inconsistent, non-standard local codes to map and report out using standard terminologies (UMLS, SNOMED CT, LOINC) enabling the smooth transition of laboratory data towards standardization and interoperability.

4. **A hybrid system for ontology-driven classification of biosamples**
   (BC Centre for Disease Control (UBC), Canada) Jan' 19 – Oct'19
   - This system performs an ontology-driven classification of biosamples customizable to different third party classification schemas using hybrid rule-based and learning algorithms trained on GenomeTrakr corpus using different classes provided by the given Extended Interagency Food Safety Analytics Collaboration (IFSAC+) classification schema. The classification by our tool is being currently incorporated by US FDA's GenomeTrakr in NCBI samples. In these NCBI samples (e.g. in sample - **https://www.ncbi.nlm.nih.gov/biosample/12609651**), the last two fields IFSAC+ classification and FoodOn ontology term in attributes section have been provided by our tool.

5. **Clustering of food products using food-indexing features**
   (BC Centre for Disease Control (UBC), Canada) Apr' 18 – Oct'18
   - This project performed a learning based clustering of LanguaL-indexed US FDA's SIREN database on selected features such as food sources and processes. The tool has helped Hsiao Laboratory's FoodOn ontology development project (**http://foodon.org**), which is a consortium-driven project to build a comprehensive food ontology, to cluster and structure more than 9300 food products from SIREN database. The inaugural paper on FoodON has been published in prestigious nature publishing journal Science of Food (**https://www.nature.com/articles/s41538-018-0032-6**).

6. **Harvesting food collocations using statistical and learning approaches**
   (University of British Columbia at BC Centre for Disease Control, Canada) Jun' 18 – Nov'19
   - Food collocations, such as "extra virgin" in "extra virgin olive oil", comprising of bigrams (digrams) have been collected using corpora from EnteroBase, GenomeTrakr, USDA Food Data Central, Nutrient Database. Python based software implements methods ranging from raw frequency count to the sophisticated statistical association measures to learning methods for collocation identification and extraction. A large **dictionary resource of food collocations** has been harvested that is being used in multiple tasks related to fine tune the data-driven analysis and analytics in food domain.

7. **Prediction of prospective buyers from player data**
   (North Side Inc., Montreal, Canada) Apr' 14 – Nov' 14
   - The objective of the project has been to do the predictive analytic using neural network approach to identify the prospective buyers of Bot Colony game of North Side Inc. using the selected set of features in the large data collected from the free trial version used by players. The tool helped NorthSide Inc. to make a huge list of prospective buyers of the game using it further to aggressively market the game and it resulted in a huge financial benefit for the company.

8. **Classification of Noxiousness Levels of Pest Plants using multiple classifiers**
   (University of Georgia, USA) Aug' 10 – Oct'10
   - The objective of the project has been to do a high level of prediction for three classes of pest plants - weeds, state noxious, and federally noxious on the basis of traits data. We applied different machine learning classifiers using Weka Machine Learning software - Artificial Neural Network (ANN), A Multilayer Perceptron, A Bayes classifier, Support vector machines (SVMs), Decision tree learning.

B) NLP RELATED PROJECTS

Most of these NLP projects have been used in the language component in NLP pipeline used for breakthrough natural language understanding technology in **Bot Colony** video game - the first application where a person can speak freely with a machine. In the projects, **technologies** used were Jena, Java, Python, NLTK, OpenNLP, SPARQL, IBM Rational Modeler as UML modeling tool, Eclipse IDE, PyCharm IDE, XML/HTML Parser/Processor, CSV Parser, other NLP tools.

9. LexMapr: A Rule-Based Tool for Translating Short Biomedical Specimens into Ontology Terms
(University of British Columbia at BC Centre for Disease Control, Canada) Aug' 17 – till date

- A hybrid text mining tool LexMapr (**github.com/Public-Health-Bioinformatics/LexMapr**) that parses the short free-text sample metadata and maps the identified entities to terms from selected domain ontologies.
- I am the team leader of the LexMapr project and I represented Hsiao Laboratory to conduct a successful workshop at ICBO 2018 (**http://icbo2018.cgrb.oregonstate.edu/W04**) to demonstrate and test the LexMapr tool amongst the biomedical community and also presented it at ISMB/ECCB conference 2019 (**https://www.iscb.org/cms_addon/conferences/ismbeccb2019**). LexMapr, since its inception, has progressed into a very useful tool in which many research institutions and organizations have shown keen interest in collaborating further.

10. Ontology-driven entity linking and granular classification system for COVId-19 literature
(University of British Columbia at BC Centre for Disease Control, Canada) Feb' 20 – till date

- The entities are recognized from the COVID-related text using NER and the entities are further linked to standard ontology terms in the domain. The linked entities are further used to perform ontology-driven fine grained classification of the literature into multiple classes.

11. PP disambiguation using BNC nouns, verbs and adjectives
(North Side Inc., Montreal, Canada) Apr' 13 – Aug' 14

- The project used an approach of applying locally built constraints language which is generalized for different senses of a specific preposition to the sentences accumulated from British National Corpus (BNC) for the preposition under consideration. This project helped North Side Inc. to enrich the PP disambiguation module as part of the natural language understanding pipeline implemented in Bot Colony game.

12. Domain model for nouns using WordFinder
(North Side Inc., Montreal, Canada) Apr' 13 – Aug' 13

- This NLP-based project modeled domains for nouns based on processing categories from WordFinder. These domains are organized in the hierarchy of 5 levels. This project helped North Side Inc. to use this domain model in many modules of the NLU pipeline implemented in Bot Colony game.

13. Domain classification system for nouns
(North Side Inc., Montreal, Canada) Aug'13 – Nov' 13

- This project of assigning domains to nouns of oxford dictionary was based on classifying nouns into domains from WordFinder based domain model developed earlier. I developed a GUI based tool for assigning domains in batches to nouns of oxford dictionary. This tool was used by linguists for assigning domains in batches using SQL queries in constrained environment and then getting it checked manually. This domain assignment helped North Side Inc. to enrich the sense disambiguation module in NLU pipeline implemented in Bot Colony game.

14. NLP based ontology-driven quality model for nouns
(North Side Inc., Montreal, Canada) Nov'12– Mar' 13

- This project focused on inheriting the qualities (of types Quantitative, Ordered, Narrative and Enumerated) based on manually built quality-quales model for top genera of taxonomy of nouns into the noun taxonomy of all nouns in Oxford dictionary. The quality values of nouns have been used in different enabling modules of the natural language understanding pipeline implemented in Bot Colony game.

15. Sense disambiguation based on Roget's categories
(North Side Inc., Montreal, Canada) Aug'12 – Nov' 12

- This novel sense disambiguation NLP project based on Roget's categories primarily built on heuristic that if the Roget's category name is present in the definitions of one/more polysemous senses of the concept under consideration then those sense/senses are chosen as the disambiguated sense/senses of that concept. This project greatly helped North Side Inc. to enrich the sense disambiguation module as part of the natural language understanding pipeline implemented in Bot Colony game.

C) ONTOLOGY RELATED PROJECTS

For the ontology related projects, the **technologies** used were Protégé, OntoFox, OWL, RDF, Jena, Java, Python, SPARQL, IBM Rational Modeler as UML modeling tool, Eclipse IDE, PyCharm IDE, XML/HTML Parser/Processor, CSV Parser, NLP tools.

16. **Design and development of COVID Literature Classification Ontology (COVLICO) for data harmonization of COVID Literature and Metadata**
    (University of British Columbia at BC Centre for Disease Control, Canada) Feb' 20 – till date
    - The corpus of domain related terms of COVID-19 literature and metadata have been automatically gathered from multiple sources using text mining techniques and the semi-automatic population of the ontology serving the COVID-19 domain. The ontology is being used for a couple of ontology-driven applications and has the potential of numerous applications when released in public domain.
17. **FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration,**
    (University of British Columbia at BC Centre for Disease Control, Canada) Apr' 17 – To date
    - FoodOn (http://foodon.org) is a consortium-driven project to build a comprehensive and easily accessible global farm-to-fork ontology about food, that accurately and consistently describes foods commonly known in cultures from around the world. Much of FoodOn's vocabulary comes from transforming LanguaL, a mature and popular food indexing thesaurus, into a World Wide Web Consortium (W3C) OWL Web Ontology Language-formatted vocabulary that provides system interoperability, quality control, and software-driven intelligence. Apart from curation I clustered and organized more than 9300 food products from SIREN database.
    - FoodOn define the complexity of food, diet, biochemical interactions and phenotypic health, provide the necessary instrument for linking and sharing the food roles for a wide range of analyses.
18. **Genomic Epidemiology Ontology (GenEpiO)**
    (University of British Columbia at BC Centre for Disease Control, Canada) Apr' 17 – To date
    - The Genomic Epidemiology Ontology **(https://genepio.org/)"** has been developed as a comprehensive controlled vocabulary for infectious disease surveillance and outbreak investigations.
19. **Ontology Based Semantic Annotation System (OBSAS) for Protein Kinases in Cancer -PhD Thesis**
    (University of Georgia, USA and PU, India) Dec' 14 – Mar' 18
    - OBSAS extracts the important protein kinase and cancer information from the bi omedical text and further semantically annotates them with classification, disease, and functional information using ontological knowledge. The linkages to the external sources of knowledge provided by the system for these semantic annotations further facilitate the extensive literature-navigation.
20. **Semantic Feature Ontology of Nouns**
    (North Side Inc., Montreal, Montreal, Canada) Apr' 11 – Mar' 17
    - The project was an extensive work of building ontology of nouns using inheritance of semantic features through genera of nouns extracted from definitions of nouns in Oxford Advanced Learners Dictionary. The nouns were associated with their genera using relationships of aggregative and partitive types extracted from definitions. More than 140 categories of Semantic Feature Ontology were assigned to the entire set of nouns using inheritance program while special cases, such as, *something, someone, pronouns etc.* treated with a special handling.
    - This project helped North Side Inc. to enrich the sense disambiguation module and built **a twenty-twenty question game** component implemented in Bot Colony game.
21. **ProKinO Ontology Browsing and Visualization Tool,**
    (University of Georgia, Athens, GA, USA Jul' 10 – Nov' 11

- For ProKinO, a standard web browsing tool has been also made available for protein kinase community (http://vulcan.cs.uga.edu/prokino/).
- Overall ProKinO project helped Kannan lab got a grant worth $720,000 by American Cancer Society for a research project of which ProKinO was one of the key objectives.

22. Protein Kinase Ontology (ProKinO),
(University of Georgia, Athens, GA, USA) Jun' 09 – Nov'11

- Analyzed, designed, developed, tested, implemented and evaluated a very large and comprehensive ontology, ProKinO that serves as a useful and efficient representation of the integrated knowledge about the complex proteins, called protein kinases, which are intimately involved in the genesis and behavior of cancer cells (**http://vulcan.cs.uga.edu/prokino/about/prokino**). I led the team involved in the development of this large scale ontology, ProKinO and I was involved in all phases of ProKinO development process.
- Once the ProKinO ontology was developed, I performed several integrative analyses of ProKinO data and used ontology-based data analysis to demonstrate the generation of testable hypotheses regarding cancer mutations. The significant work for ProKinO has resulted in two publications describing integrative analysis work and ontology development framework respectively.
- With the establishment of ProKinO ontology as a rich biological knowledge resource, Kannan group collaborated with many individuals and research groups belonging to the University of Maryland, GenenTech, Monash University, Mount Sinai hospital and other departments at the University of Georgia. Since its inception, many graduate students, including a PhD student, have worked on multiple projects directly related to ProKinO or associated ontology-driven projects.

SELECTED PUBLICATIONS

Link for selected publications list:
https://github.com/gosalpers/publications/blob/master/SELECTED%20PUBLICATIONS.pdf