

EDUCATION

Northwestern University (NU) , Evanston, IL	Sep. 2013 - June 2019 (Expected)
<ul style="list-style-type: none"> PhD candidate, Department of Materials Science and Engineering 	
Nanjing University (NJU) , Nanjing, China	Sep. 2009 - July 2013
<ul style="list-style-type: none"> Bachelor of Science in Physics. 	
University of California, Los Angeles (UCLA) , Los Angeles, CA	July 2012 – Sep. 2012
<ul style="list-style-type: none"> Exchange Program: Cross-disciplinary Scholars in Science and Technology (CSST) 	

TECHNICAL SKILLS

Python: 5+ years, including data mining and visualization, web scrapping and machine/deep learning.

Data analysis skills: Proficiency with Pandas, Scikit-learn, LightGBM, Keras, NLTK, Django, BeautifulSoup and TensorFlow; Strong experience with database management systems including MySQL; Some experience with R.

Other skills: Expert knowledge with C++, Git, Docker, AWS, web API and HTML.

RESEARCH EXPERIENCE

Department of Materials Science and Engineering , NU, IL	Sep. 2013 - June 2019 (Expected)
<i>Research Assistant, Advisor: Professor Chris Wolverton</i>	

- Materials informatics: large databases and machine learning for materials design and discovery

- Managing and maintaining a computational materials database (using MySQL) containing calculated physical and chemical properties for >600,000 compounds.
- Developed a simple and efficient web API based on REpresentational State Transfer (REST) principles to provide the community with an easy access to our database (using Django rest-framework).
- Trained three regression models (*LASSO*, *SVR* and *random-forest*) using our materials dataset and predicted materials band gaps with ~20% relative RMSE. Constructed our feature space using elemental-property-based attributes and perform univariate feature selection to reduce feature dimensions.

DATA SCIENCE EXPERIENCE

- Home Credit Default Risk: predict how capable each applicant is of repaying a loan	May 2018
<ul style="list-style-type: none"> Applied one-hot encoding for categorical features and preformed PCA to reduce dimensions. Conducted recursively feature elimination to further extract the most important features. Trained a large dataset with more than 300,000 observations using <i>gradient boost decision tree model</i> (using LightGBM) and our predicted area under the ROC curve is 0.796. Ensemble and cross validation have been utilized to overcome potential overfitting. Amazon Web Services (AWS) has been used to accelerate training process. 	
- Avito Demand Prediction Challenge: predict demand for an online classified advertisement	June 2018
<ul style="list-style-type: none"> Applied latent semantic analysis for natural language processing (using NLTK and polyglot) to extract test features from advertisement titles and descriptions. Trained <i>convolutional neural network</i> (using TensorFlow) to collect image features from images of items provided by each seller. Stored large datasets (more than 15 GB in total) using Hierarchical Data Format (HDF) and Feather format to facilitate data I/O. Performed multiprocessing to accelerate feature engineering process. Trained the dataset with more than 1,500,000 observations and 900 features using <i>gradient boost decision tree model</i> (using LightGBM) and our predicted RMSE is 0.2236 (ranked top 11%). 	

PUBLICATIONS

P. C. Chen, M. Liu, J. S. Du, B. Meckes, S. Z. Wang, H. X. Lin, V. P. Dravid, C. Wolverton, C. A. Mirkin, *Interface and heterostructure design in polyelemental nanoparticles*, **Science** 363, Issue 6430 (2019)

L. Huang, P. C. Chen, M. Liu, X. Fu, Y. Yu, P. Gordiichuk, Z. Xie, C. Wolverton, Y. Kang, C. A. Mirkin, *Catalyst design by DFT and scanning probe block copolymer lithography*, **PNAS** (2018)