Big Data for Material Science

Course Syllabus

Course Code: MSE 587
Credits: 3
Instructor: Dr. Burhan Beycan
Duration: 14 weeks
Prerequisites: Materials Science Fundamentals, Statistics, Basic Programming

---

Course Description

This course explores the application of big data analytics and computational methods to materials science research. Students will learn to handle large-scale materials databases, apply machine learning techniques for materials discovery, and develop data-driven approaches for understanding structure-property relationships in materials.

Learning Objectives

Upon completion of this course, students will be able to:

1. **Navigate and utilize** major materials databases (Materials Project, OQMD, AFLOW)
2. **Apply statistical methods** to analyze large materials datasets
3. **Implement machine learning algorithms** for materials property prediction
4. **Design high-throughput computational workflows** for materials screening
5. **Visualize and interpret** complex materials data using advanced techniques
6. **Develop predictive models** for materials discovery and optimization

Course Outline

Week 1-2: Introduction to Materials Informatics

- Overview of materials databases and repositories
- Data formats in materials science (CIF, POSCAR, JSON)
- Introduction to the Materials Project ecosystem
- Data quality assessment and validation

Week 3-4: Database Management and Querying

- SQL fundamentals for materials data
- NoSQL databases (MongoDB) for unstructured data
- API usage for materials databases
- Data extraction and preprocessing techniques

Week 5-6: Statistical Analysis of Materials Data

- Descriptive statistics for materials properties
- Correlation analysis and feature selection
- Principal component analysis (PCA)
- Clustering techniques for materials classification

Week 7-8: Machine Learning Fundamentals

- Supervised learning for property prediction
- Regression models for continuous properties
- Classification for materials categorization
- Cross-validation and model evaluation

Week 9-10: Advanced Machine Learning Applications

- Neural networks for materials science

- Deep learning for crystal structure prediction
- Ensemble methods and model optimization
- Uncertainty quantification in predictions

Week 11-12: High-Throughput Computational Methods

- Automated DFT calculations
- Workflow management systems (FireWorks, AiiDA)
- Cloud computing for materials research
- Parallel processing and optimization

Week 13-14: Materials Discovery and Design

- Inverse design methodologies
- Multi-objective optimization
- Active learning strategies
- Case studies in accelerated materials discovery

Assessment Methods

Component	Weight	Description
Data Analysis Projects	35%	Hands-on analysis of real materials datasets
Machine Learning Assignment	30%	Develop predictive models for materials properties
Research Project	25%	Independent materials informatics research
Participation & Quizzes	10%	Class engagement and knowledge checks

Required Tools and Software

Programming Environment

- **Python 3.8+** with scientific computing stack
- **Jupyter Notebook** for interactive analysis
- **Pandas, NumPy, SciPy** for data manipulation
- **Scikit-learn** for machine learning

Materials Science Tools

- **Pymatgen** for materials analysis
- **ASE (Atomic Simulation Environment)**
- **Materials Project API** access
- **VESTA** for crystal structure visualization

Big Data Tools

- **Apache Spark** for distributed computing
- **MongoDB** for database management
- **Plotly/Bokeh** for interactive visualization

Textbooks and Resources

Primary References

- **"Materials Informatics"** by Krishna Rajan
- **"Data-Driven Materials Science"** by Stefan Curtarolo et al.
- **"Python for Data Analysis"** by Wes McKinney

Online Resources

- Materials Project Documentation

- NOMAD Laboratory tutorials
- Kaggle materials science datasets
- GitHub repositories for materials informatics

Major Projects

Project 1: Materials Database Analysis

Comprehensive analysis of a large materials database, including data cleaning, statistical analysis, and identification of trends in materials properties.

Project 2: Property Prediction Model

Development of machine learning models to predict specific materials properties (e.g., band gap, formation energy, elastic moduli) using compositional and structural descriptors.

Project 3: High-Throughput Screening

Design and implementation of a computational workflow for high-throughput screening of materials for specific applications (e.g., photovoltaics, batteries, catalysts).

Final Project: Materials Discovery Challenge

Independent research project addressing a real materials discovery challenge using big data approaches and machine learning techniques.

Laboratory Sessions

Lab 1: Database Exploration and API Usage

Hands-on experience with major materials databases, learning to query and extract relevant data for analysis.

Lab 2: Feature Engineering for Materials

Creating meaningful descriptors from crystal structures and compositions for machine learning applications.

Lab 3: Predictive Modeling Workshop

Building and evaluating various machine learning models for materials property prediction.

Lab 4: Visualization and Interpretation

Advanced techniques for visualizing high-dimensional materials data and interpreting model results.

Grading Scale

Grade	Percentage	Description
A	90-100%	Exceptional performance and insight
B	80-89%	Strong understanding with good application
C	70-79%	Adequate performance with room for improvement
D	60-69%	Below expectations, significant gaps
F	<60%	Unsatisfactory performance

Course Policies

Computational Resources

Students will have access to high-performance computing resources for large-scale calculations and data analysis.

Collaboration Policy

Collaboration is encouraged for learning concepts, but all submitted work must be individually completed and properly attributed.

Data Ethics

Students must adhere to ethical guidelines for data usage and respect intellectual property rights of database providers.

Software Licensing

All software used in the course must comply with institutional licensing agreements and open-source requirements.

Industry Connections

Guest Lectures

- Materials scientists from national laboratories
- Industry experts in materials informatics
- Database developers and maintainers

Real-World Applications

- Case studies from automotive industry
- Aerospace materials development
- Energy storage and conversion materials
- Pharmaceutical and biomedical materials

Contact Information:
Dr. Burhan Beycan
Email: burhanbeycan@hotmail.com
Office Hours: Mondays and Wednesdays, 1:00-3:00 PM
Location: Materials Science Building, Room 312