CS 205 Project Whitepaper

Eri, Fang, Torrisi, Zhao

*“To a man with a hammer, every problem looks like a nail.”*

**OVERVIEW**:

Due to a recent explosion of theoretical and experimental interest in two-dimensional systems, understanding the interaction between van der Waal heterostructures has never been more timely and exciting. **The parameter space of configurations** between systems comprised of even two layers—such as chemical composition, structural phase, strain, lattice mismatch, or angle of rotation – **is enormous. Generating datasets describing the coupling requires computation on a supercomputer cloud/cluster** lasting O(1) hours, and **these datasets can be many gigabytes in size for even one material**. Thus, the toolkits of big compute and big data can be brought to bear on this problem.

**MODEL/DATA:**

The coupling between two layers can be obtained in a variety of ways using Density Functional Theory (DFT). These methods include finding the generalized stacking fault energy (GSFE) which comes from sampling the energy of various translational configurations, and computing the Wannier function (WF) representation, which describes the shapes of orbitals and the coupling between individual electron orbitals of neighboring atoms.

We generate both data sets using the *Vienna Ab-initio Simulation Package* (VASP) and Wannier90.

**TOOLS/INFRASTRUCTURE:**

**BIG COMPUTE**: We aim to extend previously existing code for computing the generalized stacking fault energy between two layers, and for generating Wannier functions for vdW bilayers so as to enable easy parallelization and automated generation of large data sets. The outputs would be, for the GSFE, a fitted function which respects the periodicity of the lattice which describes the energetic landscape of the dislocations. For the WF’s, we receive large coupling matrices with a dimensionality of O(500 x500). **The principal tasks on the big compute side will be in generating the dataset in an efficient, standardizable, parallelizable way e.g. ‘on-demand’ generation.**

**BIG DATA**: Processing these data sets into a format easily computatble for functional fitting or machine learning will be a challenge. We anticipate that because the spark infrastructure is designed to accommodate continual referencing of a dataset via MLib, it will be well suited to manage the data sets on the scale we are working on. **The principal tasks on the big data side will be coming up with efficient schemes to reduce the dataset, and efficient ways to access the dataset e.g. efficient indexes and lookup functions.**

**INFRASTRUCTURE**: We anticipate that our project will primarily be centered around developing wrapper codes in python or on spark; the compute side will be ran on Odyssey. The data side will be run either on Odyssey, or on AWS (after transferring over datasets from one to another).