Challenge of Preparing for Careers in Big Data in Geosciences

Larry Zheng[1], *Gabriele Morra[2,1], David A. Yuen[3,1], Davin Loegering[1], Henry Tufo[4,1], Chuck Li[5,1]

1. Mc Data , Wuhan, 2. Department of Physics, University of Louisiana at Lafayette, LA, 3. Department of Earth Sciences, University of Minnesota, Minneapolis, MN , 4. Department of Computer Science, University of Colorado, Boulder, CO, 5. Wuhan Huawei Technology Co., Ltd

In the aftermath of the 2008 financial crisis we have seen the steady encroachment of Big Data into every facets of society, from finances to medical services. Students graduate lacking technological skills despite needing them in the lab and on the field. We believe that putting a stronger emphasis on programming and technology will prepare them for the demands of today’ s modern job market in the geosciences and to use better measurement and analysis technology.

Our curriculum in educating students needs some changes, but universities move too slow. Therefore training centers are sorely needed. For this reason, we have established Mc Data Consult Ltd., based now in Wuhan, but poised to move anywhere.

Our aims are four fold:
(1) To establish training courses at both fundamental and advanced levels, which will be taught with customized software embedded within a affordable data-analytic tool box built with (a) cheap processors such as Raspberries Pi and (b) higher-end Nvidia TX1. Students can learn and perform exercises according to their available time slots.
(2) To provide professional consulting for various Big Data challenges encountered in industries.
(3) To hold workshops and international conferences where we can mix people from various disciplines and engage them in Big Data immersion.
(4) We also see the need to prepare suitable textbooks , focusing on high-performance computing, visualization and data analytics. We maintain that Python holds the key for preparing the students in Big Data analytics.

To be sure, the big data problem is not a new paradigm for geoscience. For instance, Peter Shearer (1991) used a relatively simple 1-dimensional velocity model to stack thousands of long-period body waves, revealing two upper mantle discontinuities, which was the first successful "big data" application: the primary computing happens for data processing, not for artificial modeling. Thus, we believe that geoscientists can be prepared to adapt to the big data era once they master the modern tools: they should master an open programming language suitable for large data, such as Python, and know how to harness parallel and distributed systems. They should learn sound software engineering skills, just as a wet chemist needs to learn to wash glassware. They should learn to produce a reproducible work: all analyses should be scripted and point-and-click tools should be avoided. They should have skills in data visualization and should master the rudiments of nonparametric, computationally based statistical inference, such as permutation tests.

Keywords: Big Data, Machine Learning, High Performance Computing, Python, Education

