COLERIDGE INITIATIVE

TRAINING MODULE

Contents

1. Motivation

It is important to establish a set of standard practices, tools, and develop the inhouse technical proficiency to access, link, and analyze datasets. The ADRF training module is designed to address such needs. This document provides the following:

(i) a summary of the existing ADRF approach to training and
(ii) a discussion of how the approach might be customized and built upon

2. Context

It is important to build human capacity as well as technical capacity when developing and implementing new systems for working with data. The ADRF is a research and analysis sandbox that demonstrates the value of linking data by bringing together data and analysts from different administrative agencies to collaborate on shared problems through hands-on projects defined by senior management. The development of training programs delivered within the ADRF: (i) promotes system and tool adoption by cultivating teams of skilled practitioners who can demonstrate the value of new ways of linking and analyzing data for solving real world problems; (ii) promotes adoption of these practices to the benefit of the practitioners' own organizations; (iii) creates a pipeline of new product prototypes central to agency missions; and (iv) generates a growing set of linked data available as an ongoing asset for use in other projects.

3. Current Approach

The Applied Data Analytics (ADA) training program develops the key computer science and data science skills necessary to take best advantage of the wealth of data made available when agencies share data and collaborate. The main learning objectives are to apply modern techniques to analyze social problems using and combining large quantities of heterogeneous data from a variety of different sources. The program has been taught five times, and each time has successfully brought together analysts with different skills from multiple agencies to collaborate on substantial projects. Below we break down the current training program as it pertains to curriculum (or what is taught), tools (or how it is taught), and people (or who is taught and through which ways).

3.1 Curriculum

At the core of the ADRF training module is a seasoned and fine-tuned curriculum of course content which exposes participants to the constellation of skills and practices involved in performing data science on heterogeneous data. This exposure is, by necessity, tempered with opportunities for hands-on engagement through class exercises and project work around real-world data analysis challenges brought by participants from their own agencies.

The ADRF training program instructs participants how to: (i) Evaluate which data are appropriate to a given research question and statistical need; (ii) Identify the different data quality frameworks and apply them to big data problems; and (iii) Learn a broad array of basic computational skills required for data analytics, typically not taught in social science, economics, statistics, or survey courses.

With these skills in mind, the curriculum is structured around four key components:

- **Foundations:** The social science of measurement; Formulating research questions; Basics of program evaluation; Differentiating data sources; "Big Data" - definitions, technical issues, quality frameworks and varying needs; Introduction to the data that will be used in this class and case studies; Introduction to Python; Working with Jupyter Notebooks[1]; Exploring data visually.
- **Data Curation:** Database concepts - database taxonomies, characteristics of large databases, and building a data schema; ETL in different databases; Building datasets to be linked; Introduction to APIs and web scraping; Linkage in the context of big data; Creating a big data workflow; and Data hygiene: curation and documentation.
- **Data Analysis:** Machine learning fundamentals - examples, process and methods; Fundamentals of network analysis - directed and undirected graphs and relational analysis of graphs; Value of text data - different text analytics paradigms, discovering topics and themes in large quantities of text data; The importance of geographic information - Basics in spatial data analysis, Geographic information systems, and Mapping your data.
- **Presentation, Inference, and Ethics:** Using graphics packages for data visualization; Error sources specific to found (big) data; Examples of big data analysis and erroneous inferences; Inference in the Big Data context; Methods to correct for data errors; Big data and privacy - Legal framework, Statistical framework, Disclosure control techniques, Ethical issues, Practical approaches.

The above elements were first outlined in the series of similar training programs delivered at the United States Census Bureau in Spring 2014 – Fall 2015 which culminated in the textbook: *Big Data and Social Science: A practical guide to models and tools*[2]. This textbook

[1] http://jupyter.org/

[2] Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, and Julia Lane. 2016. Big Data and Social Science: A Practical Guide to Methods and Tools. Chapman & Hall/CRC.

is provided as accompanying reference material to the participants in the current Applied Data Analytics training program run within the ADRF.

Figure 1. Cover to the Applied Data Analytics Course textbook: Big Data and Social Science

To deliver training in the four key components, course topics expose participants to hands on training in the following practical skills:

- Working within a secure computing environment with introductions to restricted data, the command line interface in a Linux environment, and using Git
- SQL, databases, and hands-on data exploration
- Python for data analysis and exploration
- Data Visualization and using APIs for fetching data
- Record linkage methods and tools
- Network analysis and geospatial analysis
- Machine learning (introduction and practical uses)
- Text analysis
- Web scraping and working with Big Data
- How to handle errors and inference in data analysis
- Understanding and enacting privacy and confidentiality requirements in working with administrative records

Employing this broad reaching curriculum within a single data analysis environment – where people, projects, and data come together – fosters an approach to data science around administrative data which is both comprehensive and coherent. It allows analysts, executives, and administrators to see the big picture of managing and analyzing linked administrative data

while learning at a level of detail and interactivity that helps them see how these tools and skills might be implemented at their own organizations.

3.2 Tools

As mentioned above, we focus training using a set of standard tools for data analysis which both make hands on instruction easier, allows for easier sharing and collaboration, and promotes adoption of industry standard, open source tools for analyzing data once participants return to their own organizations.

Central to instruction in the Training module is Project Jupyter (also mentioned above as Jupyter Notebooks), which is an open source, Integrated Development Environment[3] that allows participants to work with data and code in a browser and step through operations and commands, examining and documenting what is happening at each point (see **Figure 2**).

[3] https://en.wikipedia.org/wiki/Integrated_development_environment

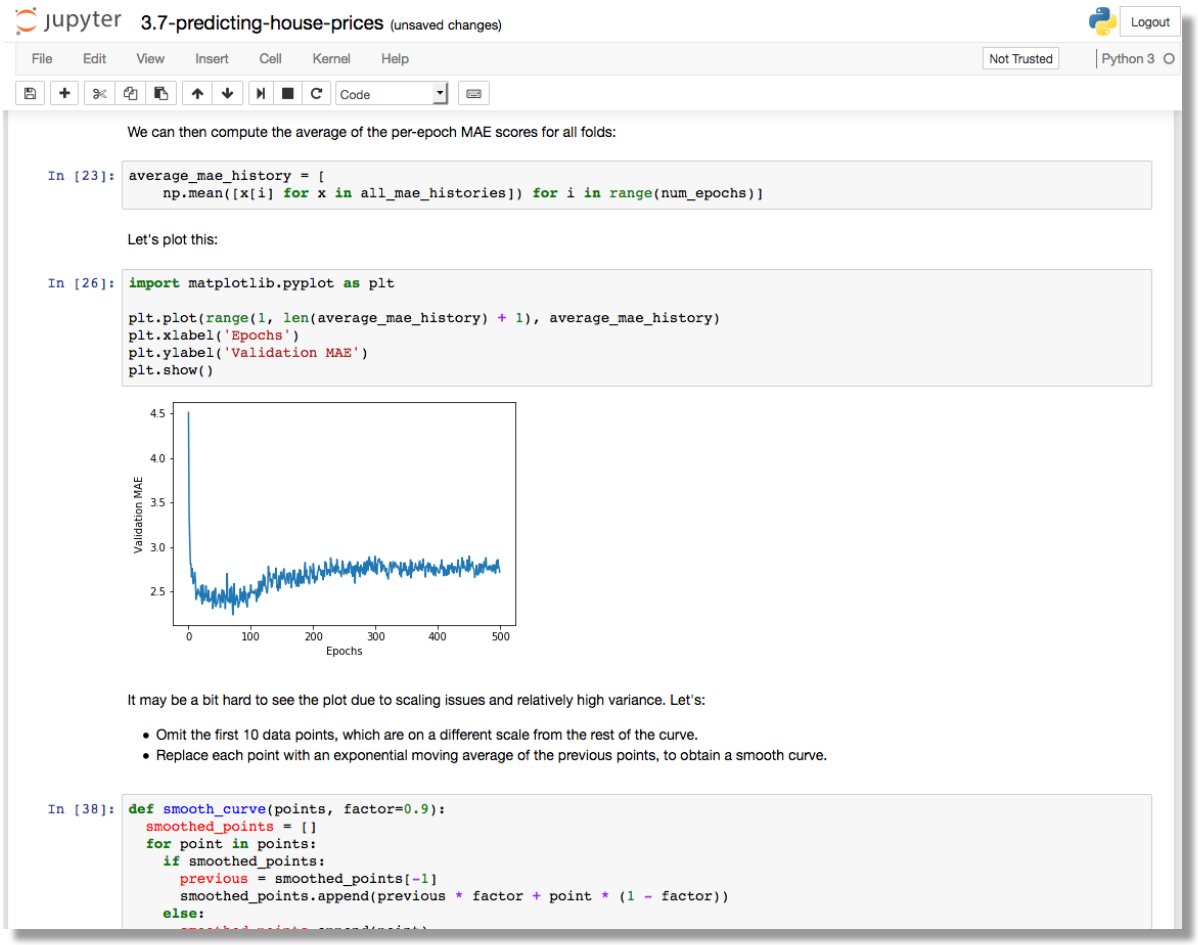

Figure 2. Screenshot of Jupyter notebook showing combination of code, documentation, and data visualization all in one place (source: https://github.com/fchollet/deep-learning-with-python-notebooks/blob/master/3.7-predicting-house-prices.ipynb)

That it is browser-based lowers the barrier to entry for participants who have not had extensive prior experience with programming tools like IDE software and a command line terminal. Jupyter Notebooks also allow instructors to add extensive explanation and comment to each section of code, allowing participants to better understand how the code works.

In addition to using Jupyter Notebooks, the course also introduces participants to popular Python libraries that facilitate powerful data processing and analysis. Libraries including in instruction are data analysis frameworks like Pandas[4] and NumPy[5], machine learning libraries

[4] http://pandas.pydata.org/
[5] http://www.numpy.org/

like scikit-learn[6], database interfaces like Psycopg[7] and SQLAlchemy[8], and libraries for conducting geospatial (e.g., GeoPandas[9]) and network[10] analyses of course and project datasets.

3.3 People

Finally, the training module prioritizes on-site course and project work mixed with remote training to ensure that the people involved in these courses – including participants, instructors, and training support staff – get the most out of the time and energy they invest in the Applied Data Analytics course.

Setup of class

Participants apply for the course by submitting a project team and data analysis proposal which identifies a real-world data challenge they work with in their own work. While the course is built to deliver instruction on skills and tools that will help them address the challenge in their project proposal, the main focus of the class is to provide opportunities, support, and feedback to participants to work on their stated project. This work includes defining an appropriate scope and addressing the technical and analytical demands of that project in realizing a viable solution or set of outcomes. In addition to providing participants an opportunity to work on a real-world data problem, it also allows them to do so in collaboration with participants across agencies.

On-site training

Following an introductory in-person meeting where all participants and staff come together in one location to get started with the course content and infrastructure, classes are then conducted live in three biweekly installments at five individual sites (New York City, the Washington DC area, Chicago, Hartford, Connecticut and Seattle, Washington). These meetings, allow instructors, support staff, and participants to work in the same room, get immediate feedback, and get hands-on training with the core elements of the curriculum while working with their project teams.

Realizing that participants come from many different locations both inside and outside of the United States, on-site training and lectures are intentionally mixed with remote learning. During each live meeting, other sites are connected together via video conferencing software to maximize delivery of content while providing everyone an opportunity to attend class at a location most convenient to them.

---

[6] http://scikit-learn.org/stable/
[7] http://initd.org/psycopg/
[8] https://www.sqlalchemy.org/
[9] http://geopandas.org/
[10] https://networkx.github.io/

4. The Future

4.1 Overview of Customizations

The training module for the Applied Data Analytics class within the ADRF is already flexible in how it may be implemented. That is, how classes are structured how they are delivered, and how participants might be organized to maximize engagement can be designed based on the specific needs of participants. Determining what to customize would rely mostly on what content to deliver, to who, and around what types of data would be included as it would be preferable to design training around data that participants actually use. Therefore, the following elements would have to be visited:

1. Understand what content needs are required for member organizations and tailor content to those needs.
2. Understand what data analysis tools and workflows member organizations might require. Should they require training in tools or practices not already covered above, we would need to determine whether that would be feasible and to what extent.
3. Understand whether project work would be required for training for the ADRF or not and tailor to those needs.
4. Develop Jupyter notebooks around data which members actually use. This would require some input from participants to identify which data and to help instructors better understand the specific data requirements of participants in developing instructional materials.

Below is a detailed breakdown for implementation of the ADRF Training module and the above customizations specifically within the two separate scenarios of 1) individual institutions using their own stand-alone versions of the ADRF that then connects with installations at partner institutions and 2) one instance of the ADRF implemented in one place for the usage of all partner institutions centrally.

4.2 Scenario 1

In the scenario where each individual organization installed and ran their own ADRF, training customizations would greatly depend on which features each ADRF installation implemented and what types of data they intended to place in the installation. Should member organizations have different needs, this might require more staff and instructor resources to meet the different needs. However, should member organizations agree that one training module would

suffice for multiple different installations, this could reduce costs and resources associated with training.

Benefits

1. Greater ability for customization of training materials.

Considerations

1. Depending on depth of customization, could increase cost of training as more resources would need to be allocated to different training cohorts.
2. Less opportunity for different institutions to work together through training, which would hinder standardization and sharing of practices.

4.3 Scenario 2

In the scenario where all institutions agreed to use one central ADRF installation, the training module could be more easily generalized to all involved institutions. This would come with the benefits of greater standardization of practices and tools used for analyzing data. As only one instance of the training module would need to be deployed, this would also consolidate resources and lower costs across participating institutions.

Benefits

1. Greater standardization of training.
2. Lower costs as training module would be shared across institutions.
3. Greater potential for group project work, which would engender more cross-institution collaboration.

Considerations

1. Potentially less customization towards each member institution.
2. Might be more difficult to schedule more people for one training opportunity.