Big Data Stackable Credentials

Big Data Career Pathways Project

Joyce Malyn-Smith, Ph.D.

Joe Ippolito, M.A.

Malyn-Smith, J., Ippolito, J. (2018) Waltham, MA: EDC.

This material is based upon work supported by the National Science Foundation under Grant No. DUE-1501927. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

EDC designs, implements, and evaluates programs to improve education, health, and economic opportunity worldwide. For more information, visit www.edc.org.

Big Data Career Pathways Project (DUE-1501927)
Report on Big Data Stackable Credentials

This report describes the efforts of four community colleges, Bunker Hill CC (MA), Johnson County CC (KS), Normandale CC (MN) and Sinclair CC (OH), who partnered with EDC on Creating Pathways to Big Data Careers, a project funded by the National Science Foundation's Advanced Technological Education Program (DUE-1501927) to design and implement programs leading to middle skills data careers. Middle skilled data workers who collect and analyze data to inform decision-making make up the data teams commonly found in companies across all industry sectors from health and hospitals to manufacturing and travel/tourism to transportation. Representing the nation's workforce development engine, community college partners aimed to propose a stackable credentials model that would illustrate ways individuals might access data education/training, with on and off ramps providing ample opportunities for employment and learning.

In May 2018, partners met in a working session to develop a stackable credentials model for Data Science and Analytics [1] that might be used by other community colleges to help design data/big data programs. The resulting model draws upon previous project work and is informed by the ongoing experience of the partner colleges. It joins previous materials created by the project- namely, the occupational profile of a Data Practitioner and performance-based rubrics aligned to that profile- as resources for colleges seeking to design their own big data career pathways.

Design Process

By the time the college partners convened at Normandale Community College to design the model, they had been immersed, for more than two years, in designing and implementing their own approaches to data analytic stackable credentials. To provide context for the ensuing design discussion, the college leaders shared what they were doing at their own schools to scaffold data analytic/ data science programs. Although there were many commonalities, the conversation demonstrated variety in structure and emphasis. For example:

- At Bunker Hill, programming follows the model established by the Department of Labor. That model includes two levels with prescribed ranges of credit (15-16 and 30-32). The average age of students in Bunker Hill's program is 28, many of whom have a Bachelor's degree and are returning to school for retooling. Bunker Hill's programs therefore take into account that students already holding a Bachelor's degree will not need to meet the GenEd requirements that the school's Associate degrees include;

- Normandale organizes GenEd in "buckets" that are relevant to a particular industry sector. In this way, the school meets the expectation of employers that their new hires have domain knowledge. This is all part of Normandy's strategy for training the "T-shaped employee", i.e. individuals having both industry relevant knowledge and the soft skills that ensure success in the workplace;

- At Sinclair, the student population includes more traditional students than Bunker Hill, i.e. students matriculating directly from high school. Here the strategy has focused on creating "embedded certificates." During the Fall, 2018 semester, Sinclair will launch a State approved Data Analytics degree that represents an

[1] The college representatives who contributed to the design of the stackable credentials model were Jaime Mahoney and Mike Harris (Bunker Hill), Suzanne Smith (Johnson County), Jim Polzin (Normandale) and Paul Hansford (Sinclair).

A.A.S. degree combined with a certificate in Data Analytics. The degree emphasizes statistics rather than calculus;

- Johnson County does not have "levels" of credentials. Currently, the school provides a one-year certificate involving 27 credits. Eight of the courses included in the program were specifically created for this credential. Instructors include Data Scientists from companies like Sprint.

The college leaders agreed that each school's stackable credential strategies provided elements that could be included in the model they were designing. They next established the criteria that would define common elements of their stackable credentials model:

- The model should be *generic* enough to provide for broad applicability.
- The model should be *flexible* enough to address the needs of various student populations, i.e. both traditional and non-traditional students.
- The model should be *aligned to standards* delineated by the project's occupational profile of a Data Practitioner.[2]
- The model should *incorporate the best practices* of the models already established at the partner schools and share their insights regarding *challenges and lessons learned*.

Discussion focused on the following guiding questions: What populations do the programs serve? What and how many certificates does it include? Do the certificates illustrate different levels that can be tied to jobs or knowledge/skill sets? How many credits are usually required for the various certificate levels? How do the levels relate to each other?

The Big Data Stackable Credentials Model

A. Target Populations

The model aims to address as wide an audience of prospective students as possible. It envisions serving individuals who are economically driven, who may be seeking career advancement, who may or may not be in a position to transfer degrees. It takes into account that students bring a variety of goals with them- some wanting to move immediately into a job, some intending to matriculate to four year schools, others wanting to upgrade their skills and some simply pursuing personal interests. In general, the target populations can be divided into two groups- traditional and non-traditional students. Those groups can be sub-divided further:

- Traditional students
    a. students directly from high school or with a small gap from their high school experience
    b. GED students
    c. individuals with limited work experience who want to start a career in data science
    d. International students

- Non-traditional students

[2] The profile of a Data Practitioner provides a succinct, detailed description of the work activities, skills, knowledge and behaviors embodied by effective, middle skill data workers. The profile was developed by a panel of expert data workers representing ten different industry sectors. The profile's depiction of a Data Practitioner was reviewed and validated by more than 100 data workers via a national, on-line survey. The profile is attached to this report.

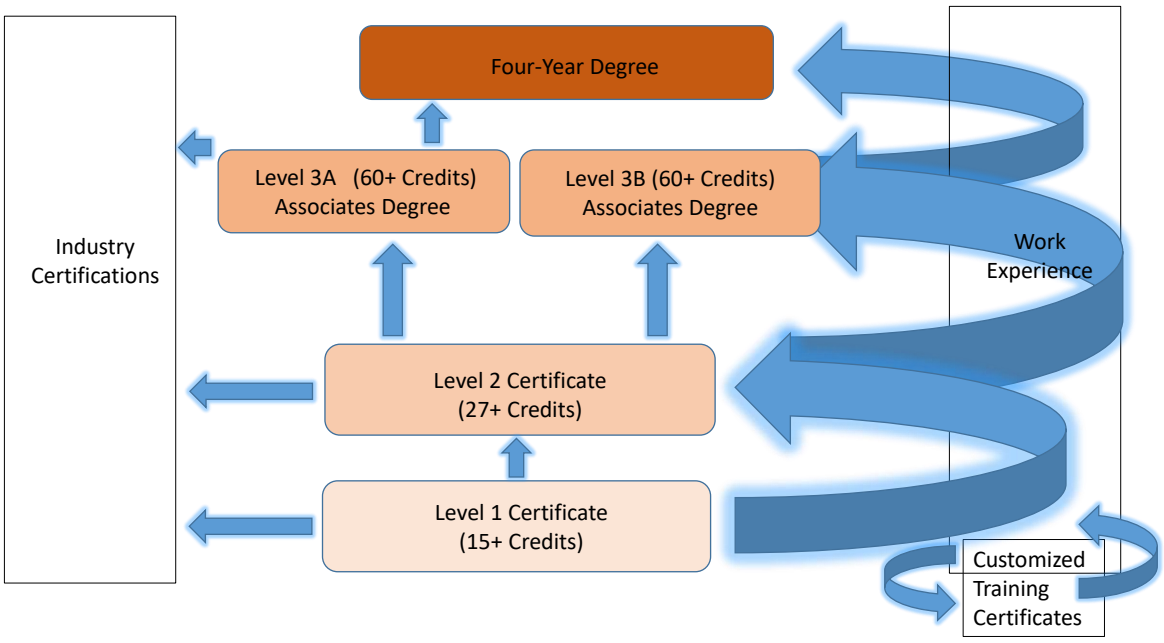
a. established professionals or technical workers seeking to upgrade skills/develop new data skills/or validate experience in working with data
b. individuals seeking a career change - move into the data field
c. veterans seeking to apply their military skills set to industry
d. people returning to work after a hiatus
e. people who want to pursue personal interest in data science

B. Diagram of Big Data Stackable Credentials Model

The diagram that follows summarizes the structure of the stackable credential model that the partner colleges agree would represent a common structure that could be shared by all. It also illustrates the paths students involved in data studies may follow. The college partners envisioned four primary certificate/ program levels to the model, with a fifth category designated to recognize the prevalence of industry certifications. The levels of the model include:

- Level 3A (60+ credits): Associates in Science Degree designed to prepare people to transition to a four-year degree program in data science. Prepares individuals to visualize data and tell the data story, and to work with big data (large amounts of data). Includes calculus course.
- Level 3B (60+ credits): Associates in Science Degree designed to prepare people for direct entry into work in a data field. Prepares individuals to visualize data and tell the data story, and to work with big data (large amounts of data).
- Level 2 Certificate (28+ credits): Prepares individuals to analyze data and build models.
- Level 1 Certificate (15+credits): Prepares individuals to collect, clean, transform and stores data; and is able to describe data (mean, median, mode…)
- Industry Certifications (1+ credits): Trains individuals in a narrowly focused skill set required for a specific work purpose. Two or more industry certifications can be horizontally stacked to prepare individuals for specific jobs in data fields.

Stackable Credentials Model for Data Science and Analytics (Data Practitioner)



The arrows in the diagram are meant to portray real life fluidity. Individuals may cycle in and out of work and training. They may pursue a path directed towards attaining a four-year degree. They may opt to build their skills by strictly enrolling in courses offered by academic institutions or by engaging in industry sponsored certifications.

In addition to the credit bearing opportunities, there are other opportunities for students to upskill in the data field. These include Boot Camps, MOOCs, skill building microcredentials/digital badges, nano degrees, and non-credit continuing education/workforce development opportunities. Non-credit opportunities are available through some Continuing Education and Workforce Development offices in community colleges, and through organizations such as: EdX, Udacity, Coursera, Khan Academy, Data Camp, Udemy, Lynda and others.

C. Alignment of Model to Data Practitioner Profile

The profile of a Data Practitioner articulates the work expected of a middle skill data worker. The profile is grounded in an occupational definition developed by the expert panel. It reads:

The Data Practitioner, in service of an organization and/or stakeholders, supports the data life cycle by collecting, transforming, and analyzing data, and communicating results in order to inform and guide decision-making.

Based upon that definition, the profile organizes the work expected of an effective Data Practitioner into major and minor responsibilities, referred to as duties and tasks respectively. Employers looking for new employees, or considering offering student internships, will want to know the extent to which students are prepared to perform these work activities. For that reason, the stackable credentials model includes an alignment between the Data Practitioner profile and the four primary certificate/ program levels.

The table below lists every duty and task found in the profile in the far left column. The remaining columns identify each of the four primary certificate/ program levels. The extent to which each of the primary certificate/ program levels prepares students for each of the profile tasks is designated using the following key:

X= Individuals have been academically prepared to perform this task

S= Individuals have been academically prepared to technically perform this task but may lack the work/ life experience to make decisions inherent in performing this task, and therefore may require supervision.

S/E=Individuals have been academically prepared to perform this task. Depending on work/life experience, they may require supervision as they make decisions inherent in performing this task.

E= Individuals have been academically prepared to perform this task are expected to have sufficient work/life experience to perform this task successfully without supervision.

Alignment of Data Practitioner Tasks to Model Levels

Work Task From Data Practitioner Profile	Level 1 (15+ Credits)	Level 2 (28+ Credits)	Level 3A A.S. Degree Leading to 4 Year School	Level 3B A.S. Degree to Work
1. INITIATES THE PROJECT				
1A Translates business problems into analytic needs			X	X
1B Interviews stakeholders		S	S/E	S/E
1C Refines stakeholder needs		S	S	S
1D Identifies appropriate data	X	X	X	X
1E Identifies whether data exists or not	X	X	X	X
1F Performs gap analysis of the data		X	X	X
1G Determines resource needs (e.g. SMEs, tools, timelines).		X	X	X
1H Determines feasibility of analysis to be done			X	X
1I Creates statement of work			X	X
2. SOURCES THE DATA				
2A Determines data sources(s)	X	X	X	X
2B Determines target structure		X	X	X
2C Collects data	X	X	X	X
2D Exercises quality control (e.g. randomizes selection)			S/E	S/E
2E Extracts data (e.g. write SWL, API code…)		X	X	X
2F Cleans data (e.g. identifies outliers/ errors)	X	X	X	X
2G Tests data	X	X	X	X
2H Creates data dictionary	X	X	X	X
2I complies with business, ethical and legal standards	X	X	X	X
3. TRANSFORMS DATA				
3A Merges data	X	X	X	X
3B Splits data	X	X	X	X
3C Derives new variables	X	X	X	X
3D Creates new data	X	X	X	X
3F Augments data		X	X	X
3F applies metadata			X	X
3G Purges data	S	S	S/E	S/E
3H Changes data structure		S	S/E	S/E
3I Changes data types	X	X	X	X
3J Normalizes data	X	X	X	X
3K Interpolates data		X	X	X
3L Finalizes data dictionary		X	X	X
3M Store data for analytics	X	X	X	X

Alignment of Data Practitioner Tasks to Model Levels

Work Task From Data Practitioner Profile	Level 1 (15+ Credits)	Level 2 (28+ Credits)	Level 3A A.S. Degree Leading to 4 Year School	Level 3B A.S. Degree to Work
4. ANALYZES THE DATA				
4A Determines what analysis to run		S	S	S
4B Applies the research method and tools		X	X	X
4C Identifies dependent and independent variables	X	X	X	X
4D Defines appropriate algorithms		S	S	S
4E Performs data mining		X	X	X
4F Separates any anomalies	X	X	X	X
4G Interprets the results		S	S/E	S/E
4H Runs additional tests as needed		S	X	X
4I Performs reasonableness tests of results			S	S
4J compares results to previous findings			S	S
4K Confirms results		X	X	X
4L Conducts causality testing		S	S	S
4M Creates data visualizations (e.g. dashboards, reports, charts, graphs, videos, animation)	X	X	X	X
5. CLOSES OUT THE PROJECT				
5A Selects documentation media	X	X	X	X
5B Describes problem method and analysis		S	S/E	S/E
5C Articulates conclusions		S	S/E	S/E
5D Compiles reports		X	X	X
5E Presents information to stakeholders		X	X	X
5F Integrates feedback from stakeholders		X	X	X
5G Defends analysis as needed		X	X	X
5H Reworks analysis as needed		X	X	X
5G Prepares final report		X	X	X
5J Archives work products	X	X	X	X
6. ENGAGES IN PROFESSIONAL DEVELOPMENT				
6A Maintains professional qualifications			X	X
6B Stays current on emerging technologies, methods and tools	X	X	X	X
6C Seeks out mentors	X	X	X	X
6D Shares best practices			E	E
6F Attends relevant conferences and seminars	X	X	X	X
6G Mentors others			E	E
6H Participates in professional organizations	X	X	X	X
6I Suggests future projects			E	E

D. Tools

As colleges consider building big data certificates or programs, they need to be mindful of the resources that faculty members will need to instruct students. The following tools were identified by the college partners as being essential to performing the tasks identified in the Data Practitioner profile:

- DataCamp
- Vocareum (Automated Software)
- Statistical Package (R Studio/ R, SPSS, SAS)
- Python IDE or Notebook (Eclipse/ Jupiter)
- Database Tools (Oracle, SQL/SQL Server, Excel)
- Open Data Portals
- Data Visualization tools (Tableau, Qlik, Spotfire, PowerBI etc.)
- Mockaroo - This website provides an easy way to create a messy data set. You can specify field names and types and can set a certain percentage of missing data for each field. It is then easy to go in and add some outliers and inconsistent category names so that students have some messy data to clean.
- Knime Data Analytics Platform - This is a tool that allows students to read, clean, transform, and visualize data, and to create models using a drag-and-drop interface. It allows students to visualize the data flow so that they can see the big picture and not get lost in the details of programming. Once they understand the big picture, they are ready to learn how to use programming to accomplish the same tasks.
- Data World -A site for using meaningful, collaborative, and abundant data as a resource.
- Lynda - a part of LinkedIn and libraries , a leading online learning platform that helps anyone learn business, software, technology and creative skills to achieve personal and professional goals.
- MOOCS (Massive Open Online Courses) - free (or low fee) Web-based distance learning program that is designed for the participation of large numbers of geographically dispersed students.
    - Examples are EdX, MIT OpenCourseWare,Udemy, Coursera, Udacity, Khan Academy

E. Challenges

Once a college decides to create a system of stackable credentials in big data, it can be certain that it will confront a series of challenges and/ or obstacles. Here is a sampling of what the college leaders encountered:

- **Identifying qualified faculty-** Because data science involves statistics, mathematics, and programming, existing full-time faculty members are rarely prepared to teach all courses in the field. Those in the math department will likely need to learn programming and those in a computer science department will need to learn some statistics. This creates a challenge in being able to find qualified full-time faculty members able to teach the range of skills required in various data courses;

- **Identifying qualified adjunct instructors**- Adjunct faculty are often hired to staff new courses and programs. The ability to recruit appropriately skilled adjunct faculty is entirely dependent on the location of a school. In large metro areas, it is possible to find more data scientists who are willing to teach. But in smaller communities, it may be more difficult to identify and recruit data scientists willing to do so;

- **Four year colleges dictate 2-year program courses.** Most 4 year degree programs require calculus. Community college programs preparing students to transfer to a 4 year data science program are required to offer calculus as part of their program offerings;

- **Mathematics skill levels can be a barrier to entry-** When 80% of students entering community colleges test into developmental level math, it limits the pool of potential students prepared to succeed in the data science program, and it is a barrier to entry;

- **Matching curriculum to transfer institution demands and industry demands-** Associate degree requirements and transfer institution requirements do not necessarily mesh easily. Careful selection of existing courses and collaborative design of new courses is needed to make a "smooth" transfer path and also satisfy local degree requirements;

- **Keeping up with industry changes-** Relationships with industry partners and faculty development opportunities are needed to keep up with current trends and technologies. Industry connections can be difficult to create and maintain, and faculty development opportunities can be limited. Industry partnerships can lead to faculty development opportunities such as externships where faculty spend time working in industry;

- **Marketing/ awareness**- Programs need to know how and where to market themselves to industry to make potential business partners aware that their college programs exist. Students/ professionals/ high schools and other institutions/ industries need informational materials to capitalize on and move into and out of these programs;

- **Open Source material grading and project work**- Utilizing open source material is great for students, but it presents it's own challenges. If you are utilizing open source material, you will need to think about how you assess the students because all of the answers to questions in the book are available to students online. This requires more time to creating and developing assessments.

F. Lessons Learned

As an additional aid to colleges working to develop stackable credentials in data analytics, the college leads described lessons learned from their own experiences. These include:

- **Professional development of faculty (Bunker Hill Community College)**
    - The development of faculty is both time and learning intensive. As a program begins, faculty is needed to champion its development, to teach the program's courses, and to learn the domain. MOOCs provide low cost avenues for instructors to develop domain expertise. The John's Hopkins data specialization on Coursera, and the Georgia Tech Micromasters in Analytics on edX are two examples of MOOCs. Georgia Tech also has an online Master of Science in Analytics (the same degree as their in house MS in analytics) available for less than $10K. Faculty can utilize these low cost options to increase their skills, and to prepare to teach in the field.

- **Networking and awareness due to rapid changing environment (Normandale Community College and Sinclair Community College)**
    - Relationships with industry are key to understanding local market demands for skills and credentials. Similarly, access to current information on skills and tools is critical for faculty and institutions to keep programs relevant. In addition, it is important to develop and maintain connections with other academic institutions. Exchange of program visions, plans and courses can provide a resource for strengthening one's own program.

- **Generic wording on faculty job postings to eliminate restrictiveness (Bunker Hill Community College)**
    - Identifying individuals who have the appropriate skill sets, education and credentials to be well prepared to develop and teach a data curriculum is difficult due to limited supply and lower compensation traditionally offered by community colleges. Bunker Hill found that many individuals who are well qualified to develop and teach a data curriculum may have fewer degrees and/ or degrees in a range of academic fields not including data science. They may also have extensive work experience in data science. Therefore, colleges should consider developing job postings for full-time faculty that invite a broader, and by extension larger, pool of candidates by including the minimum degree requirement allowed by accreditation and including degrees beyond data science. Individuals with degrees in a range of academic fields related to data (for example, mathematics, engineering, information technology, computer science, hard sciences) may also be well qualified.

- **Faculty work experience, compensation, tenure track**
    - Posting should also place a high value on work experience in the data science field or related fields. In addition, flexibility in offering higher compensation should be implemented whenever possible. Where compensation flexibility is limited, tenure track positions should be offered as a potential mechanism to somewhat overcome the compensation challenge. Consider offering release time, or additional compensation for professional development or industry partner activities required as part of offering data programs.

- **Design Certificates around FAFSA and DOL requirements (Bunker Hill Community College)**
    - The first and second level certificate requirements should be designed around federal financial aid and state guidelines to make it easier for students to access FAFSA funds to pay for their college experience. For Massachusetts, it is 15+ credits for a 1st level certificate, and 27+ credits for a second level certificate.

- **Convene a roundtable discussion with industry, partners and students to understand pathways (Bunker Hill Community College)**
    - Bunker Hill had a discussion with faculty from 2 year and 4 year colleges, along with people from industry to showcase programs the school was developing. As a result of the discussion, Bunker Hill was able to find 4 internships for its students, and was also able to build a partnership with a local industry.

- **Develop workforce development options for non-credit students (Sinclair Community College)**
    - Colleges should build foundational courses for augmenting the start of a degree program for non-traditional students advancing their career.

- **Demand no prerequisites for course entry into the certificate or degree programs (Sinclair Community College)**
    - Many community college students test into developmental courses. Lowering the bar for students to be able to upgrade their skills, to pursue their personal learning interests and/ or start a certificate or degree program increases the reach of the program by expanding the potential audience.

Conclusions

The partner colleges developed a stackable credentials model that reflects a set of common criteria and also establishes a structure allowing for program differences in individual college experiences. The model provides the language and structure for a conversation among colleges to describe their middle skilled Data Practitioner program development efforts, and between colleges and local businesses for the purposes of engaging in their program development activities. The model can be used by other colleges considering development of stackable credentials in the area of data science and analytics. Colleges interested in starting a new program can use the model to develop short and long term strategic plans, to recruit and engage business partners, to recruit new faculty, to develop student recruitment materials and to begin the institution's internal program/ curriculum development process. Colleges that already have courses/ programs developed can use the model to approach other institutions of higher education in the development of articulation and transfer agreements. Colleges and business partners can use the model to help develop student internship, faculty externship, and other work-based learning programs. Colleges can use the model as a starting place to compare and contrast their data science and analytics programs, progress, successes, issues and concerns with other colleges.