BD2K Training Coordinating Center's ERuDIte: the Educational Resource Discovery Index for Data Science

José Luis Ambite, Lily Fierro, Jonathan Gordon, Gully A. Burns, Florian Geigl, Kristina Lerman, John D. Van Horn

Abstract—Data science is a field that has developed to enable efficient integration and analysis of increasingly large data sets in many domains. In particular, big data in genetics, neuroimaging, mobile health, and other subfields of biomedical science, promises new insights, but also poses challenges. To address these challenges, the National Institutes of Health launched the Big Data to Knowledge (BD2K) initiative, including a Training Coordinating Center (TCC) tasked with developing a resource for personalized data science training for biomedical researchers. The BD2K TCC web portal is powered by ERuDIte, the Educational Resource Discovery Index, which collects training resources for data science, including online courses, videos of tutorials and research talks, textbooks, and other web-based materials. While the availability of so many potential learning resources is exciting, they are highly heterogeneous in quality, difficulty, format, and topic, making the field intimidating to enter and difficult to navigate. Moreover, data science is rapidly evolving, so there is a constant influx of new materials and concepts. We leverage data science techniques to build ERuDIte itself, using data extraction, data integration, machine learning, information retrieval, and natural language processing to automatically collect, integrate, describe, and organize existing online resources for learning data science.

Index Terms—I.2.6.g Machine learning, I.2.1.d Education, H.2.0.b Database design, modeling and management, H.2.8.c Data and knowledge visualization, I.2.12.c Ontology design.

1 Introduction

The National Institutes of Health (NIH) launched the Big Data to Knowledge (BD2K) initiative (datascience.nih.gov) to fulfill the promise of biomedical "big data" [2]. The NIH recognized that *"The ability to harvest the wealth of information contained in biomedical Big Data will advance our understanding of human health and disease; however, lack of appropriate tools, poor data accessibility, and insufficient training are major impediments to rapid translational impact.*[1]*"* The NIH BD2K program has funded 15 major centers[2] to investigate how data science can benefit diverse fields of biomedical research including genetics, neuroimaging, precision medicine, and mobile health. Ensuring that the advances produced by these centers, and other research efforts, permeate the biomedical research community and yield the expected benefits for human health requires a significant increase in the number of biomedical researchers trained in data science. To address this need, the NIH has funded the BD2K Training Coordinating Center (TCC).

Data science demands knowledge from many branches of mathematics and computer science, notably statistics and machine learning, and can be applied to multiple fields of study. Given the field's interdisciplinary nature and its growing popularity, many open learning resources have been published on the Web for anyone interested in learning about data science. However, these resources vary greatly in quality, topic coverage, difficulty, and presentation formats, making entry into the world of data science confusing and daunting for learners.

To address these challenges, the BD2K Training Coordinating Center is developing a web portal (BigDataU.org) to provide a dynamic, personalized educational experience for biomedical researchers interested in learning about data science. The portal is powered by ERuDIte, the Educational Resource Discovery Index for Data Science, a curated, richly described collection of existing web-based training materials on data science. In order to build ERuDIte, we are developing novel, automated methods to identify, collect, integrate, describe, and organize web-based learning resources.

In the collection stage, we have built a web-scraping framework that allows us to rapidly incorporate new sources and extract relevant data from them. In the integration stage, we have designed a unified schema for learning resources to integrate heterogeneous data into a single, consistent model. Under this model, the system also exposes the metadata of learning resources as linked data [3], [4], so these resources can be easily cross-referenced by others. In the description stage, ERuDIte uses methods

- *J. L. Ambite [0000-0003-0087-080X], L. Fierro [0000-0003-4440-4982], G. Burns [0000-0003-1493-865X], and K. Lerman [0000-0002-5071-0575] are with the University of Southern California's Information Sciences Institute (ISI), Marina del Rey, CA 90292. F. Geigl's [0000-0001-9759-2396] work was performed as a visiting Ph.D. student at ISI.*
- *J. Gordon [0000-0002-8896-2057] is with the Department of Computer Science, Vassar College, Poughkeepsie, NY 12604. His work was performed primarily as a postdoctoral researcher at ISI.*
- *J. D. Van Horn [0000-0003-1537-0816] is with the University of Southern California's Stevens Neuroimaging and Informatics Institute, Los Angeles, CA 90033.*
- *Primary Contact E-mail: ambite@isi.edu*

- *This paper is an expanded version of [1].*

[1] https://commonfund.nih.gov/bd2k
[2] https://commonfund.nih.gov/bd2k/centers

from machine learning, information retrieval, and natural language processing to tag resources with concepts from a hierarchical, multi-dimensional ontology designed to provide an extensible, lightweight description of core aspects of the field of data science.

In summary, both in its design and in its creation, ERuDIte uses the concepts and methods of the data science field that it aims to teach. ERuDIte will enable students and researchers to make the best use of the diverse data science learning resources available online.

2 Building ERuDIte

Since ERuDIte is itself a data science project, its construction reflects some of the key stages in the data science workflow, namely data collection, integration, modeling, and visualization. In the following sections, we detail our efforts along these stages in our development of ERuDIte.

2.1 Learning Resource Acquisition

In ERuDIte, learning resources are online resources that have a pedagogical component for data science concepts and skills. The quality and relevance of learning resources are essential to the development and success of ERuDIte, and, consequently, our initial collection efforts focused on well-known, high-quality sources, such as leading Massive Online Open Courses (MOOCs) and talks and tutorials from scientific conferences. While some sources provide learning resource data through public APIs (e.g., coursera.org and udacity.com), most sources require scraping of websites intended for human navigation. For that purpose, we built a modular framework using the popular Python packages BeautifulSoup and Dryscrape to handle both static websites and dynamic, JavaScript-based pages, which have historically been problematic.

In this framework, each source website is handled by a module designed for the site's structure and idiosyncrasies. These require some manual authoring, but, once created, the site-specific module automatically collects resource data. The scraping framework is packaged as a Docker image, so it can be used without locally managing its dependencies. As a result, we were able to increase our resource collection efforts quickly because team members could simultaneously build new site-specific modules without disturbing the core infrastructure of the scraping framework.

To date, we have collected a total of 11,320 learning resources, which vary in granularity from individual videos to online courses that include multiple video lectures and associated training material. Table 1 describes the current sources, the number of learning resources from each source, and the types of information extracted, such as resource descriptions, video transcripts, and supporting slides or other written materials.

2.1.1 YouTube Classification

To expand our learning resource collection beyond our manually curated sources, we are developing techniques to identify high-quality learning resources from large open collections, such as YouTube. We are applying information extraction and machine learning techniques to automatically assess the quality of data science videos on YouTube for inclusion in ERuDIte.

Searching for "data science" on YouTube yields over 200,000 videos – and over 9,000,000 when not constrained to the exact phrase. However, the number of these videos that are both relevant and pedagogically valuable is much lower. To filter the results, we trained a classifier to assess quality based on video metadata and content.

To find potentially relevant YouTube videos, we search for terms related to data science, drawn from the *Field* dimension of the ontology described in Section 2.3.1 (Figure 2). These queries include the names of disciplines and concepts, sometimes with additional restrictions, for example: "bioinformatics", ("python" AND "data science"), or ("regression" AND ("data science" OR "machine learning")).

We executed 98 such queries and collected metadata from the videos and playlists appearing in the first 20 pages of results for each query, yielding a dataset of 122,557 unique videos. We manually annotated 2,298 videos, sampled from across different pages of results for the queries. Initially, these were judged on a scale of 0–4, where 0 is a video that is completely unhelpful as a resource for learning about data science and 4 is most helpful. For simpler classification, these labels are binarized, with resources labeled 0–1 considered too low quality to index and 2+ considered sufficiently good. This annotation effort yielded 1,217 high-quality videos and 1,081 low-quality ones.

We trained a random forest classifier using a variety of features from the YouTube videos, including the uploader ID, upload date, number of views, likes and dislikes, average rating, duration, tags and categories, and encodings of the title, description, and transcripts in a 50-dimension Word2vec vector space. With five-fold cross-validation, the classifier achieves precision 0.82, recall 0.81, and F_1 score of 0.82. This performance is sufficient to select highly promising videos from YouTube for final human curation. As the size of our training data improves, we expect the automatic classification quality to approach human levels of agreement and minimize human effort.[3]

2.1.2 Google Books

For pedagogical reasons, we initially focused collection on video materials, but we have begun to extend our data collection to scientific written materials. We queried the open Google Books API[4] with a set of 54 queries specific to data science, similar to our YouTube searches. This yielded 19,666 records, consisting not only of simple metadata for each book (title, authors, description, publisher, URL), but also snippets of text from within the book that surrounded hits of the search terms. To clean the corpus from off-topic books, we generated a 200-topic latent Dirichlet allocation (LDA) topic model using the MALLET toolkit [5] and manually examined the word distributions of each topic to determine whether the topic is relevant to data science. We then removed any document that had an irrelevant topic in any of the top three topics provided by LDA. We

[3] Currently, the discovered 122,557 unique videos have been automatically classified. Predicted high quality videos are under review using our curation interface (cf. Section 2.4.1), so these videos are not fully included in the YouTube totals in Table 1.

[4] https://www.googleapis.com/books/v1/volumes?q="terms"

TABLE 1
Currently Indexed Learning Resources

Provider/Source	Types	Total	With Descriptions	With Transcripts	With Slides or Documents
BD2K	Video, Written	681	602	277	72
edX	Course, Video	89	88	69	53
Coursera	Course, Video	256	256	81	83
Udacity	Course, Video	17	17	17	0
Videolectures.net	Video	8,577	6,166	7,994	4,699
YouTube	Video	988	873	749	0
ELIXIR	Course, Written	237	48	0	0
Bioconductor	Course, Written	5	2	0	0
Cornell Virtual Workshop	Course, Written	38	19	0	0
OHBM	Video	78	6	0	51
NIH	Video	1	1	0	0
Bioinformatics.ca	Course, Video	86	63	0	0
Google Books	Written	267	213	0	0
Total		**11,320**	**8,354**	**9,187**	**4,958**

manually assessed the documents generated by this filtering and found the relevancy to be acceptable. This procedure provided a total of 12,379 book records for subsequent analysis and curatorial review.[5]

2.2 Resource Integration

To integrate the heterogeneous resource data we collected under a uniform schema, we designed a metadata standard to represent learning resources in ERuDIte.

2.2.1 Global Schemas for Learning Resources

To facilitate cross-institution data sharing, we first reviewed existing standards, including classes and properties from the Dublin Core,[6] Learning Resource Metadata Initiative (LRMI),[7] IEEE's Learning Object Metadata (LOM),[8] eXchanging Course Related Information (XCRI),[9] Metadata for Learning Opportunities (MLO),[10] and Schema.org vocabularies. Our initial model had three classes: *LearningResource* (with 27 properties), *Person* (with 8 properties), and *Provider* (with 10 properties). However, with the aim of creating a standard that is more universal, which allows for greater detection, discovery, and interchangeability, we updated our model based on our participation in the World Wide Web Consortium (W3C) Schema Course Extension Group[11] and our collaboration with the ELIXIR consortium,[12] which uses the Schema.org-based standard defined by Bioschemas.org.[13]

There are a variety of large-scale efforts across the world developing training resources, including MOOC providers as well as large research consortia like the BD2K program. One effort of particular importance in the biomedical space is the ELIXIR consortium, which seeks to provide a distributed infrastructure for life-science across Europe, in a spirit akin to the NIH BD2K Initiative. The ELIXIR Programme includes a training component, the Training e-Support System (TeSS),[14] which plays a role analogous to the BD2K TCC.

We have established a collaboration with ELIXIR TeSS to develop joint metadata standards for learning resources and to share data synergistically. As part of this collaboration, we have redefined our metadata standard to adopt Schema.org vocabularies, only defining additional properties when critically needed. Pages with embedded Schema.org markup, such as the resource pages at BigDataU.org, are preferentially indexed by major search engines, such as Google and Bing, so by using this vocabulary, we facilitate the discovery and dissemination of the resources indexed in ERuDIte.

A graphical overview of the ERuDIte metadata standard appears in Figure 1. The key classes of our standard are *CreativeWork* (used for learning resources), *Person* (for instructors or material creators), and *Organization* (for affiliations and learning resource providers). The schema definition is also available for download at https://github.com/bioint/erudite-training-resource-standard under a Creative Commons Attribution-ShareAlike License (version 3.0) license.

2.2.2 Integrated Resource Database

All learning resources collected by ERuDIte are stored in an integrated relational database, which we refer to as the Resource Database. This database uses views to map source tables to our metadata standard, which we have translated into a relational schema, in order to remain flexible for any future changes and extensions. The scraping framework outputs source-specific tables, and the views in the database integrate the source data into a single schema model. We then use an additional reporting materialized view that joins relations defined by the schema to form a composite table that generates the data for resource detail pages for display and use on the BD2K TCC web portal (http://BigDataU.org). We also generate an Elasticsearch

[5] These books are partially included in Table 1. The rest are under curatorial review, and we expect to add them incrementally as they are reviewed (Section 2.4.1).

[6] http://dublincore.org

[7] http://lrmi.dublincore.net

[8] https://standards.ieee.org/standard/1484_12_1-2002.html

[9] https://shop.bsigroup.com/ProductDetail/?pid=000000000030259242

[10] https://joinup.ec.europa.eu/solution/metadata-learning-opportunities-mlo-advertising

[11] https://www.w3.org/community/schema-course-extend

[12] https://www.elixir-europe.org

[13] http://bioschemas.org

[14] https://tess.elixir-europe.org

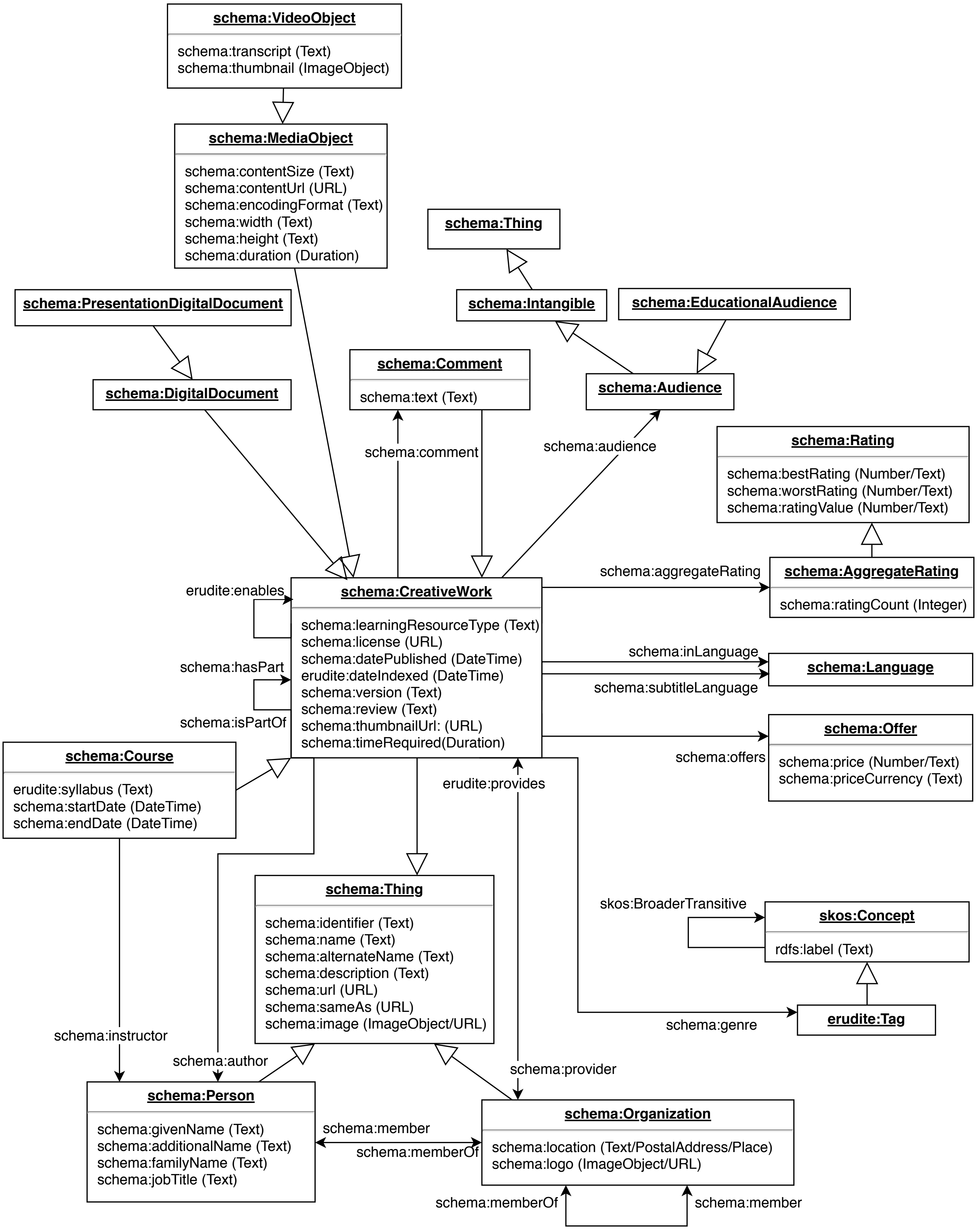


Fig. 1. ERuDIte metadata standard based on Schema.org vocabularies. Learning resources in ERuDIte are instances of *CreativeWork*; the people who create or teach the learning resources are instances of *Person*; and, the institutions that provide and/or are affiliated with the people who create or teach the learning resources are instances of *Organization*.

(https://elastic.co) index from a query to this table, and that index powers the search interface on the web portal.[15]

2.2.3 Learning Resource Metadata as Linked Data

The linked data movement [3] seeks to make data available on the Web not only readable to humans, but also to machines. The JSON-LD format is a popular way to insert structured data into regular web pages and contribute to the web of linked data. These structured data snippets can then be easily extracted by external tools and indexed by search engines. In particular, Google encourages the use of JSON-LD over the Schema.org vocabulary for this purpose.[16] In the spirit of open data sharing, we expose all metadata for each learning resource in the ERuDIte collection as linked data in the JSON-LD format, both embedded in each of the learning resource pages on BigDataU.org that are reachable through our faceted search interface, and as a complete data file published as a versioned Digital Object Identifier (DOI) at Zenodo.[17] ERuDIte metadata is made available under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license.[18] Listing 1 shows a sample JSON-LD markup for a single educational resource. The JSON-LD data has recently been enhanced with links to DBpedia for organizations, and to DBpedia [6], DBLP [7], and ORCID (orcid.org) for instructors [4].

As part of our integration pipeline, we developed an automated mapping functionality from the Resource Database's relational schema into our Schema.org-based standard (cf. Figure 1) as JSON-LD, using our previous work on data exchange [8]. Augmenting our published learning resources with JSON-LD structured data allows current and future collaborators to easily cross-reference any resource we collect, increasing data interchangeability across global efforts for educational resource indexing.

```
{"@context": {"@vocab": "https://schema.org/",
              "bdu-resource":
              "http://bigdatau.org/resource/",
              "bdu-organization":
              "http://bigdatau.org/organization/",
              "bdu-person": "http://bigdatau.org/person/",
              "dseo": "http://bigdatau.org/dseo#"},
  "@id": "bdu-resource:14783120385630643790",
  "@type": "CreativeWork",
  "author": {"@type": ["Person"],
             "@id": "bdu-person:Andrew_Y._Ng"},
  "description": "Machine learning is ...",
  "genre": ["dseo:image_data", "dseo:introductory",
            "dseo:MATLAB_Octave", "dseo:data_analysis",
            "dseo:written_documents", "dseo:video",
            "dseo:machine_learning" ],
  "provider": [{"@type": ["Organization"],
                "@id": "bdu-organization:Coursera"},
               {"@type": ["Organization"],
                "@id": "bdu-organization:
                        Stanford_University"}],
  "name": "Machine Learning",
  "url": "https://www.coursera.org/learn/machine-learning"}
```

Listing 1: JSON-LD markup for http://BigDataU.org/resource/14783120385630643790

2.3 Resource Description

As an additional layer of description beyond the learning resources' collected metadata, we designed a hierarchical, multi-dimensional ontology known as the Data Science Education Ontology (DSEO), to provide further categorization of the learning resources in ERuDIte. This ontology provides learners with concepts that can assist them with resource exploration and discovery.

2.3.1 Defining the Data Science Education Ontology

To design the Data Science Education Ontology (DSEO), we combined top-down and bottom-up approaches. First, we identified relevant concepts based on our knowledge of the data science domain and organized them hierarchically along six dimensions. Each dimension represents a facet that learners would want to use in searching the resource collection. For example, if the learner is a neuroscientist who just received a recent set of fMRI data and who needs to visualize the data using Python, the learner should be able to filter and target his or her searches with terms from the ontology to cover these needs. With this top-down structure in mind, we then collected and reviewed categories used to describe learning resources in each of the existing sources (e.g., videolectures.net provides a categorization of its video collection), and those concepts were used to discover and fill gaps in our defined ontology.

We then incorporated two semi-automated methods to refine and extend the ontology further in a bottom-up manner. As a first semi-automated method, we developed a system that analyzes the textual information associated with the learning resources (including titles, descriptions, syllabi, transcripts, slides, etc.) to automatically generate concepts from bigrams, trigrams, nouns, and shallow noun phrases[19] extracted from sentence trees constructed by the Stanford Parser [9]. In evaluating these automatically identified concepts, we found that shallow noun phrases from the parser provided the richest terms. We reviewed the 8,160 automatic concepts from the parser and eliminated ambiguous and irrelevant ones. We also added tags related to the depth, field (i.e. knowledge domain), and format of the course. This process identified a total of 861 candidate tags.

As a second semi-automated method, we used non-negative matrix factorization (NMF) [10] to discover topics in our resources. We analyzed the most significant words associated with each topic and defined a concept for each of the topics. Much of this analysis confirmed the concepts identified earlier, but it also yielded ten additional concepts. More recently, we have created another ten new concepts from the topics generated by the LDA model used in Section 2.1.2 in order to increase DSEO's coverage for the written text we have collected through Google Books.

We apply the following criteria for a concept to be included in the DSEO:

1) Is there enough support for the concept within our resource collection? (Currently, we require more than five resources to be relevant to the concept.)
2) Does the proposed concept capture an abstract phrase that cannot be automatically extracted from

[15] For example, http://BigDataU.org/search?query=machine+learning

[16] https://developers.google.com/search/docs/guides/intro-structured-data

[17] https://doi.org/10.5281/zenodo.1214375

[18] https://creativecommons.org/licenses/by-nc-sa/4.0

[19] We define "shallow noun phrases" as ones constructed with words at a single node level in the parse tree of resource descriptions.

text (i.e., it would not be easily found by an information retrieval search over the resource text)?

3) How does the proposed concept impact a user's ability to discover a resource?
4) Does a clear definition for the concept exist?
5) Can the proposed concept be automatically assigned using machine learning? (cf. Section 2.4).

Given that the DSEO describes the learning resources in ERuDIte to enable searching and filtering in the interface of the TCC Web Portal, well-defined concepts are essential. Consequently, this set of criteria was used to reduce the ontology to a total of 126 concepts, which we organized hierarchically along six dimensions. For clarity, we define a specific question that every dimension aims to answer for a learner. These questions are listed below, along with how many of the 126 concepts fall under each dimension.[20]

Data Science Process (7)
What stages of the data science process will this resource help me with?

Field (83)
What field of study does this resource focus on?

Datatype (18)
What types of data does this resource address?

Programming Tool (14)
What programming tool is used in or taught by this resource?

Resource Format (2)
How is this resource presented?

Resource Depth (2)
How advanced is this resource?

Figure 2 shows all concepts in the DSEO, as seen at http://BigDataU.org/explore_erudite. We expect the DSEO to be a living, breathing ontology that can adapt to innovations in data science. As we discover and assess more resources, we expect new concepts to emerge.

2.3.2 Publishing the Data Science Education Ontology

DSEO is formally a Simple Knowledge Organization System (SKOS)[21] vocabulary, with the hierarchical relationships encoded by the *skos:broaderTransitive* property. DSEO is publicly available at GitHub[22] and at the BioPortal ontology repository[23] (for convenience of visualization in BioPortal, we also defined a version of DSEO using *rdfs:subClassOf*). Beyond its applications to the web portal at BigDataU.org, DSEO's greatest value lies in its concern for the intricacies and developments of data science. Consequently, by making DSEO public, we welcome community suggestions and edits to extend DSEO with any emerging, relevant concepts.

2.4 Automatic Concept Assignment (Tagging)

In order to scale up ERuDIte, we need to develop automated methods to assign concepts from our ontology to the collected learning resources, i.e., tagging. Here, the tagging problem is a multi-label one; each learning resource can have multiple tags from each dimension of the DSEO.

TABLE 2
Training + Cross-Validation and Testing Set Sizes (Total Resources)

Dimension	Training/CV Set Size	Testing Set Size
Field	7,904	1,885
Resource Depth	1,241	299
Resource Format	7,870	1,989
Data Science Process	1,725	447
Programming Tool	429	109
Datatype	1,866	466

We initially explored both machine learning and information retrieval methods using resource text as inputs, and we found that one-versus-all logistic regression was the method that yielded the best-performing classifier in five-fold cross-validation [1]. In order to improve our classifiers and to assess them further, we expanded our gold standard dataset. For each dimension, we developed a gold standard of hand-curated resources (data science courses from Coursera, Udacity, edX, and Cornell's Virtual Workshop, and videos from Videolectures.net and YouTube) labeled with the appropriate tags from each DSEO dimension. For each dimension, we left aside approximately 20% of each gold standard set for testing, and we used the rest for cross-validation and training. For each dimension's testing set, we made sure that every tag in the dimension is represented by randomly selecting resources on a per tag basis. Table 2 shows the total number of learning resources in each dimension's training/cross-validation and testing set.

We take the hierarchy of DSEO into account by also assigning parent and ancestor tags to a resource. For example, if a resource is tagged with "clustering," we also tag it with that concept's parents and ancestors: "unsupervised learning," "machine learning," "artificial intelligence," "probability statistics," "computer science," and "mathematics."

For classifier training, we defined an experimental procedure that used the same dataset, cross-validation folds, and performance measurements for every method tested in order to select the best model. We then created a training framework using the popular Python machine learning package scikit-learn [11] to perform grid-searches over configurable parameters, which are defined by configuration files that handle parameters for data access (e.g., table to query for source data), document vectorization (e.g., n-gram range, minimum document frequency threshold, maximum document frequency threshold), and classifier methods (e.g., C value for regularization strength, probability threshold). The framework takes the source data, which includes resources' titles, subtitles, descriptions, syllabi, transcripts, and text from slides and additional written documents, and combines them to form a single text document for each resource. It then vectorizes each resource document as a bag-of-words TF–IDF vector, forming the input feature matrix to our classifiers. Next, the input features are sent to classifiers specified by the configuration, and the results of training and cross-validation are output into a standardized, reviewable format. The dashed arrow path through steps A, B, C, D, and E in Figure 3 graphically presents this workflow of the framework.

[20] The top-level concept for the *Programming Tool* dimension can be assigned to resources, while the top-level names of the other dimensions are only used for organization of concepts.

[21] https://www.w3.org/2004/02/skos

[22] https://bioint.github.io/DSEO

[23] https://bioportal.bioontology.org/ontologies/DSEO

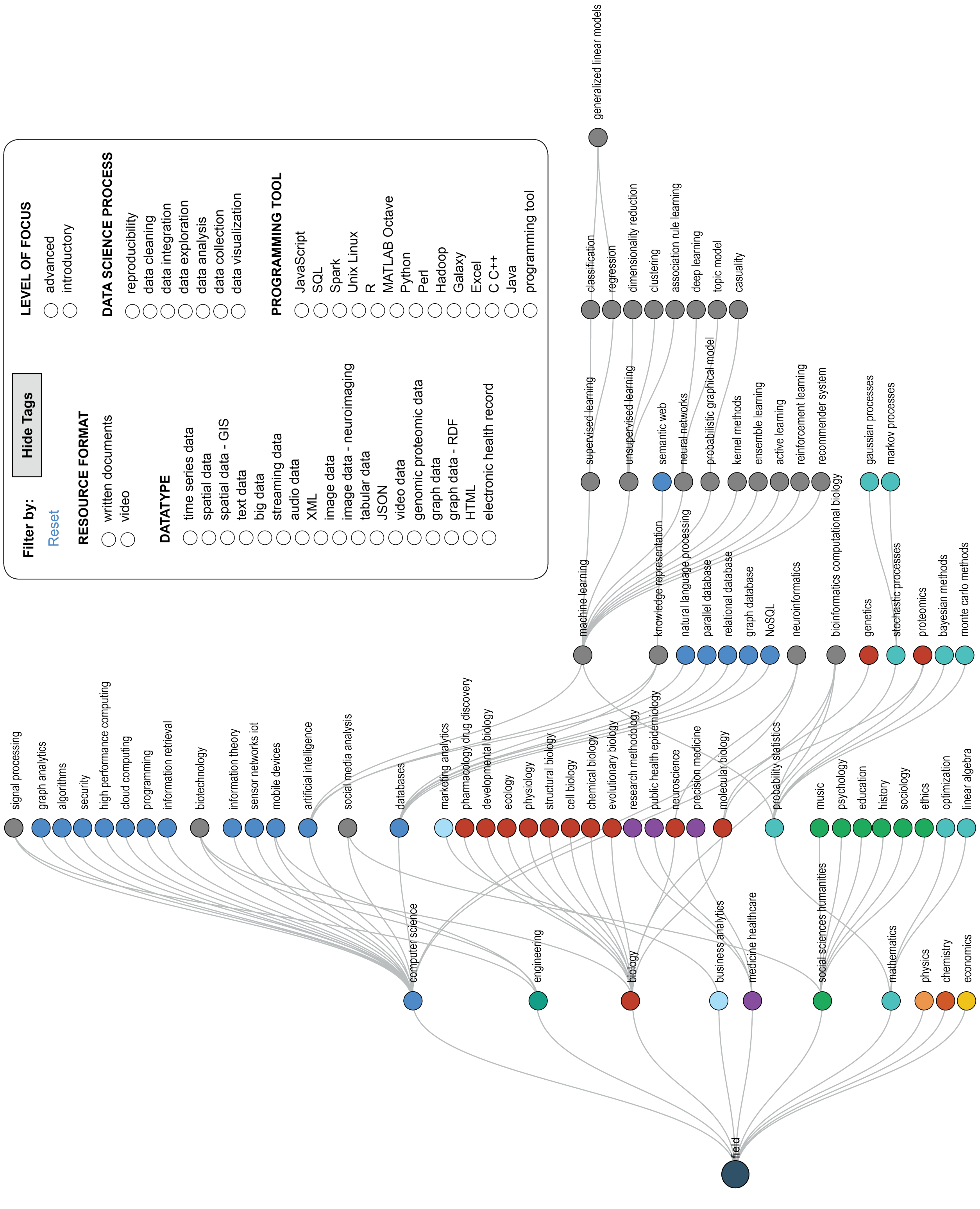


Fig. 2. Data Science Education Ontology (DSEO), as shown at http://BigDataU.org/explore_erudite.

In recent exploration, we noticed performance improvements with more support per tag. Thus, to ensure that tags have an adequate amount of support for each fold, we now only include tags that have a minimum of five examples of support [24] and use the multi-label stratified k-fold approach of [12] to assign fold numbers. We then perform five-fold cross-validation grid searches over hyperparameters defined for two classifier types: one-vs-all logistic regression and one-vs-all random forest. In [1], we found that one-vs-all logistic regressions using L2 regularization were the most successful. However, given the increases in our feature matrices due to the increases in the sizes of our training/cross-validation sets, we favor one-vs-all logistic regressions using L1 regularization in order to give weight to the most discriminative terms. Based on the success of our YouTube classifier (cf. Section 2.1.1), which similarly uses text features from learning resource metadata, we also perform grid searches using one-vs-all random forest classifiers. Our performance metric for classifier comparison was the F_1 score, which is the harmonic mean of precision (positive predictive value) and recall (sensitivity). We calculated the weighted average F_1 score, with the weights equal to the number of true positives of each tag in the validation fold, in each fold, to select the best hyperparameter combination for each classifier. Afterwards, we predicted tags for each test set we left aside and calculated the weighted F_1.

Table 3 shows the performance of the best classifiers for each dimension on its respective test set. Overall, the larger training datasets for each dimension help improve classifier performance. With our continuous curation efforts (cf. Section 2.4.1), we expect the classifiers to continue to improve while also easing the burden of curation.

2.4.1 Continuously Improving Tag Assignments

To improve our tagging classifiers beyond their current performance, we still need more gold standard data, particularly for under-represented concepts in DSEO. However, asking curators to manually label every resource from our collection would require too much time. Consequently, in order to assess our existing classifiers and to reduce curation time, we have created a pipeline where tag predictions on novel learning resources are made by the classifiers and sent to a curation interface (Figure 4) as recommended tags for curators to confirm or reject. In the interface, curators can add tags that were not predicted by the classifiers, and they can also suggest tags that are relevant, but are not currently in the DSEO. In addition, here, curators are given the opportunity to assess the quality of a resource (good, bad, skip, or remove), and these quality labels can be used to inform and expand the resource quality classifiers discussed in Section 2.1.1. As such, this curation process allows us to continuously update our classifiers with more gold standard data, as shown by the solid arrow path through steps A, B, C, E, F, G, and H in Figure 3.

Furthermore, with the curation interface, multiple curators can provide tags for a resource, which allows us to assess inter-rater reliability in order to solidify tag assignments. This process will allow us to find any further gaps in the DSEO and to address any ambiguity between concepts in the ontology. While curation is currently only internal to the project, we envision later opening it to users of the web portal or crowdsourced curators, allowing us to re-train and validate our automated tagging algorithms at scale.

2.5 Resource Visualization

We ran MALLET's LDA-based topic modeling on an aggregated corpus made up of all available video resources in the ERuDIte catalog with enough text to analyze, combined with the records obtained from Google Books (for a total of 18,458 documents). Having generated topic signatures across all documents, we then applied t-SNE [13] to these signatures to project them into a two-dimensional space. We then used the Bokeh visualization library[25] to generate a two-dimensional scatterplot of available documents that also permits end users to interact directly with the mapped data (Figure 5). Within this visualization, each dot represents a single document and is colored according to the main topic of that document. Mousing over a dot reveals metadata concerning that document and clicking a node navigates to the appropriate page within ERuDIte. This visualization is available at http://BigDataU.org/erudite_cluster. The source code supporting this visualization is also publicly available on GitHub.[26]

3 Ongoing Work

The ERuDIte system is under active development. To reach our vision for ERuDIte as a dynamically updated, personalized system suited for self-directed learning, we are pursuing the following research directions.

3.1 Dependencies and Prerequisites

To enable personalized learning plans, we are studying how to automatically infer what data science concepts are presented in each resource and what other concepts are prerequisites for these; e.g., if the learner is interested in a course on machine learning, but his or her user profile does not indicate experience in mathematics, ERuDIte should suggest starting with a resource on probability.

There are many ways to infer the concepts involved in a set of resources, with topic modeling such as latent Dirichlet allocation (LDA) being a traditional approach [14]. LDA is unsupervised and requires no external resources, but the topics it produces can be unclear. After exploring several approaches, we chose to use Wikipedia articles as concepts, since they are well-defined and have broad coverage of data science concepts. Given a resource, we can infer a distribution over Wikipedia concepts using explicit semantic analysis (ESA) [15].

In previous work [16], we created an unsupervised method for inferring prerequisites based on an information theoretic analysis of large corpora of technical text. We are now pursuing an approach that exploits "naturally occurring" ordering relations between concepts, such as textbook tables of contents, course syllabi, and – our current

[24] This is consistent with the minimum of five relevant resources required to add a new tag into the DSEO (cf. Section 2.3.1)

[25] https://bokeh.pydata.org

[26] https://github.com/SciKnowEngine/sciknowmap

TABLE 3
Precision, Recall, and F_1 Scores on Each Dimension's Independent Test Set for the Dimension's Best Classifier, Overall and Over Tags with at Least a Given Level of Support

Dimension	Classifier Type	Support ≥5: P	R	F_1	Support ≥10: P	R	F_1	Support ≥15: P	R	F_1	Support ≥20: P	R	F_1
Field	Logistic Regression	0.74	0.88	0.80	0.74	0.88	0.80	0.74	0.88	0.80	0.74	0.88	0.80
Resource Depth	Random Forest	0.66	0.91	0.76	0.66	0.91	0.76	0.66	0.91	0.76	0.66	0.91	0.76
Resource Format	Logistic Regression	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Data Science Process	Logistic Regression	0.69	0.77	0.73	0.69	0.77	0.73	0.69	0.77	0.73	0.69	0.77	0.73
Programming Tool	Logistic Regression	0.80	0.71	0.74	0.80	0.71	0.73	0.79	0.70	0.73	0.81	0.76	0.77
Datatype	Logistic Regression	0.75	0.86	0.79	0.75	0.86	0.79	0.75	0.87	0.80	0.76	0.87	0.80

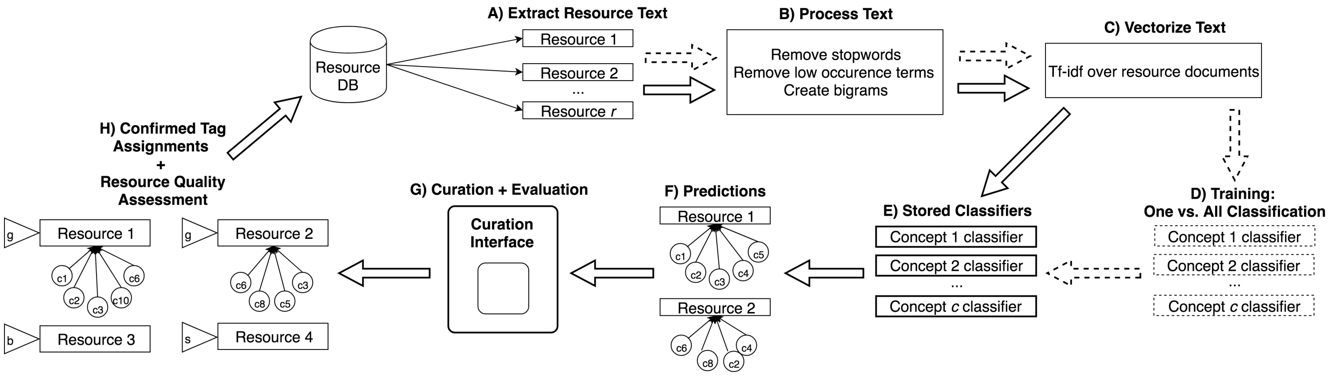
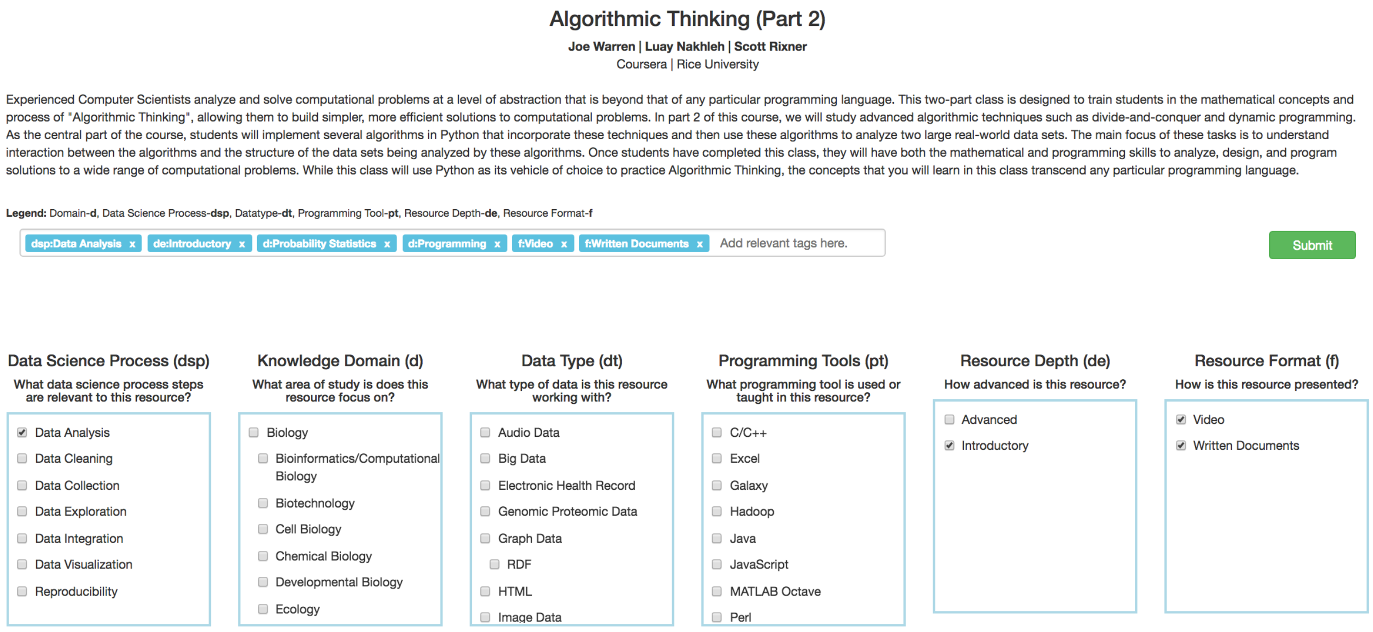


Fig. 3. Processing workflow for training (path drawn by dashed arrows) and evaluating and updating (path drawn by solid arrows) automated tagging classifiers. Circles in steps F and H represent concepts tagged to a resource. Triangles in step H represent quality assessments, with g=good, b=bad, and s=skip.



Algorithmic Thinking (Part 2)

Joe Warren | Luay Nakhleh | Scott Rixner

Coursera | Rice University

Experienced Computer Scientists analyze and solve computational problems at a level of abstraction that is beyond that of any particular programming language. This two-part class is designed to train students in the mathematical concepts and process of "Algorithmic Thinking", allowing them to build simpler, more efficient solutions to computational problems. In part 2 of this course, we will study advanced algorithmic techniques such as divide-and-conquer and dynamic programming. As the central part of the course, students will implement several algorithms in Python that incorporate these techniques and then use these algorithms to analyze two large real-world data sets. The main focus of these tasks is to understand interaction between the algorithms and the structure of the data sets being analyzed by these algorithms. Once students have completed this class, they will have both the mathematical and programming skills to analyze, design, and program solutions to a wide range of computational problems. While this class will use Python as its vehicle of choice to practice Algorithmic Thinking, the concepts that you will learn in this class transcend any particular programming language.

Legend: Domain-**d**, Data Science Process-**dsp**, Datatype-**dt**, Programming Tool-**pt**, Resource Depth-**de**, Resource Format-**f**

dsp:Data Analysis x | de:Introductory x | d:Probability Statistics x | d:Programming x | f:Video x | f:Written Documents x | Add relevant tags here.

Data Science Process (dsp)
What data science process steps are relevant to this resource?
☑ Data Analysis
☐ Data Cleaning
☐ Data Collection
☐ Data Exploration
☐ Data Integration
☐ Data Visualization
☐ Reproducibility

Knowledge Domain (d)
What area of study is does this resource focus on?
☐ Biology
☐ Bioinformatics/Computational Biology
☐ Biotechnology
☐ Cell Biology
☐ Chemical Biology
☐ Developmental Biology
☐ Ecology

Data Type (dt)
What type of data is this resource working with?
☐ Audio Data
☐ Big Data
☐ Electronic Health Record
☐ Genomic Proteomic Data
☐ Graph Data
☐ RDF
☐ HTML
☐ Image Data

Programming Tools (pt)
What programming tool is used or taught in this resource?
☐ C/C++
☐ Excel
☐ Galaxy
☐ Hadoop
☐ Java
☐ JavaScript
☐ MATLAB Octave
☐ Perl

Resource Depth (de)
How advanced is this resource?
☐ Advanced
☑ Introductory

Resource Format (f)
How is this resource presented?
☑ Video
☑ Written Documents

Fig. 4. The curation interface for reviewing predicted/recommended tags for a resource. This is the tagging screen, where curators review recommended tags and add additional tags if needed.

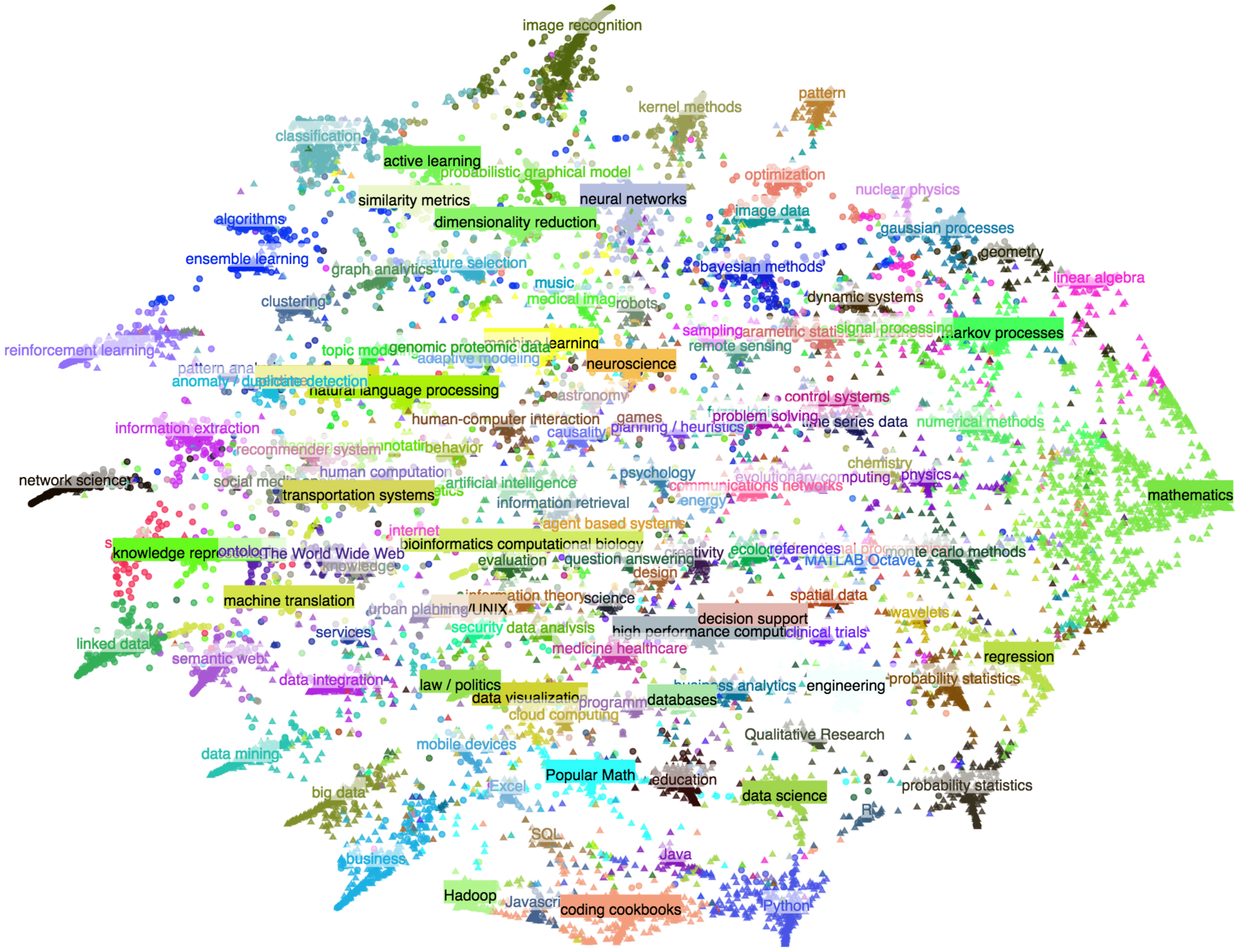


Fig. 5. Visualizing the structure of Learning Resources using LDA and t-SNE

focus – user navigation on Wikipedia. Using the released "clickstream" data [17], we find that learners most often look up the concept they are interested in learning about and then navigate to more basic concepts, e.g., from "Deep learning" to "Neural networks." Based on this insight, we have trained a classifier to identify prerequisite pairs and are in the process of evaluating its accuracy against existing sets of manually judged prerequisites, such as those defined in the Metacademy[27] guide of machine learning concepts.

3.2 Personalization

We plan to explore personalization methods in ERuDIte through recommendations tailored for an individual learner via collaborative filtering. To do this, we have instrumented the web portal to collect user activity data. This will allow us to benefit from a large, consistently engaged user base to build our recommendation engine.

4 Related Work

We briefly review work related to ERuDIte. There are a number of commercial "MOOC aggregators" (such as Class Central, CourseBuffet, CourseTalk, TubeCourse, etc.), developed as social web applications, but the techniques for automatic identification, description, and organization of learning resources we propose in ERuDIte go beyond what these sites provide. The TechKnAcq project serves as an example of the possibility of such methods, attempting to structure the underlying organization of a pedagogical resource based on analyses of the content of that resource [16]. The concept hierarchies we use to describe resources can also be learned from existing resources [18]. For our visualization approach, we build on our previous work on the NIHMaps project [19], which provided a navigable map of all grants issued by the NIH, allowing users to explore the high-level structure of funded grants across several years. Other efforts have also used NMF to drive the creation of visual clusters. The multi-view NMF of [20] shows the potential to use more resource metadata in the generation of future resource visualizations. In the BD2K program, there is a parallel effort, bioCADDIE, to catalog scientific datasets [21], but it is not focused on learning resources. ELIXIR-UK Training e-Support System (TeSS), has similar goals as the BD2K TCC. As discussed in Section 2.2.1, we are coordinating with ELIXIR TeSS to share resources and exploit synergies. As part of this collaboration, we are also working with EDAM [22], a comprehensive ontology for data, topics, and operations in bioinformatics to connect our concepts and to collaborate in areas where our respective ontologies can address each other's gaps in coverage.

[27] https://metacademy.org

5 Conclusions

When looking at ERuDIte as its own data science project, we have made significant progress on the data collection, data integration, data exploration, and data analysis steps. In the development of ERuDIte so far, we have designed and implemented a flexible scraping framework, a unified schema, a tagging ontology, a visualization approach for resource exploration, and a collection of automated tagging algorithms. We are more than halfway towards completing the vision of making ERuDIte a platform that aggregates and organizes relevant resources and provides a personalized and engaging experience for the self-directed data science learner.

Our immediate future plans are to curate several thousand data science videos from YouTube and books from Google Books and add further high-quality resources to our collection. A major ongoing effort is to identify prerequisite relations between learning resources/concepts, e.g., that linear regression should be learned before logistic regression. We plan to provide personalized training paths using our resource descriptions, prerequisite relations, and from mining user interactions (searches, creation of educational plans, ratings, etc.) in the BigDataU.org web portal. In future work, we plan to explore active learning techniques to optimize curation and classifier advancement by prioritizing resources that would address key areas where our classifiers need to improve.

Although ERuDIte currently focuses on knowledge about data science, the techniques used in its construction are general, therefore we expect that the ERuDIte platform can be applied to other fields. Most careers demand continuous, self-directed learning well outside of degree programs, and few tools exist to help learners navigate through the heterogeneous resources on the Web. Consequently, ERuDIte has the potential to expand interaction with an important subset of scholarly data: web-based educational resources. Historically, when thinking about the web of scholars, we look at journal publications and citations, but now, in the age of digital learning, scholars also produce open-access educational resources, creating a source of data that connects, informs, and educates not only scholars, but also anyone interested in learning more about a field, concept, or technique. With this type of educational resource, the web of scholars can strengthen across disciplines, for understanding others' work is easier through an open educational resource as compared to a journal article, and can grow because many more people have access to the materials they need to learn in order to become scholars themselves.

Acknowledgments

This work is supported by the National Institutes of Health[28] under Grant U24ES026465. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Institutes of Health.

We thank the rest of the BD2K TCC team: Jeana Kamdar, Sumiko Abe, Xiaoyu Lei, Avnish Bhattrai, Crystal Stewart, Caroline O'Driscoll, Carmen Tan, Shobhit Agarwal, and Krithika Ramaswamy for their support and their work on the TCC web portal. We thank Rochelle Tractenberg for her guidance in the process of defining skills and knowledge areas essential for data science. We thank Priyambada Jain for her contributions to website modules in the scraping framework, Joel Mathew for his contributions to the curation interface, Supriya Anand and Mohsen Sayyadi for their contributions to prerequisite discovery work, Aakanchha Sinha for her contributions to the YouTube classification work, and Pedro Szekely, Dipsy Kapoor, and Amandeep Singh for support with the Karma system.

[28] https://dx.doi.org/10.13039/100000002

References

[1] J. L. Ambite, L. Fierro, F. Geigl, J. Gordon, G. A. P. C. Burns, K. Lerman, and J. D. Van Horn, "BD2K ERuDIte: The educational resource discovery index for data science," in *Proceedings of the 26th International Conference on World Wide Web Companion*, ser. WWW '17 Companion. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2017, pp. 1203–11. [Online]. Available: https://doi.org/10.1145/3041021.3053060

[2] L. Ohno-Machado, "NIH's big data to knowledge initiative and the advancement of biomedical informatics," *Journal of the American Medical Informatics Association* (*JAMIA*), vol. 193, 2014, doi: 10.1136/amiajnl-2014-002666.

[3] T. Heath and C. Bizer, *Linked Data: Evolving the Web into a Global Data Space*, ser. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers, 2011. [Online]. Available: http://dx.doi.org/10.2200/S00334ED1V01Y201102WBE001

[4] J. L. Ambite, J. Gordon, L. Fierro, G. Burns, and J. Matthew, "Linking educational resources on data science," in *Proceedings of the 31st Innovative Applications of Artificial Intelligence Conference (IAAI)*, Honolulu, Hawaii, 2019.

[5] A. K. McCallum, "MALLET: A machine learning for language toolkit," 2002. [Online]. Available: http://mallet.cs.umass.edu

[6] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A nucleus for a web of open data," in *The Semantic Web*. Springer-Verlag, 2007, pp. 722–735.

[7] M. Ley, "The DBLP computer science bibliography: Evolution, research issues, perspectives," in *String Processing and Information Retrieval SPIRE*, Lisbon, Portugal, 2002, pp. 1–10.

[8] M. Taheriyan, C. A. Knoblock, P. Szekely, and J. L. Ambite, "Semi-automatically modeling web APIs to create linked APIs," in *Proceedings of the ESWC 2012 Workshop on Linked APIs*, 2012. [Online]. Available: http://usc-isi-i2.github.io/papers/knoblock12-eswc.pdf

[9] D. Chen and C. D. Manning, "A fast and accurate dependency parser using neural networks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), 2014.

[10] F. Shahnaz, M.W. Berry, V.P. Pauca, and R.J. Plemmons, "Document clustering using nonnegative matrix factorization," *Information Processing & Management*, vol. 42, no. 2, pp. 373–86, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306457304001542

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[12] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," in *Machine Learning and Knowledge Discovery in Databases*, D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 145–158.

[13] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–605, Nov. 2008.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Mar. 2003.

[15] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using Wikipedia-based explicit semantic analysis," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, ser. IJCAI'07. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007, pp. 1606–11.

[16] J. Gordon, L. Zhu, A. Galstyan, P. Natarajan, and G. A. P. C. Burns, "Modeling concept dependencies in a scientific corpus," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (*ACL*). Association for Computational Linguistics, Aug. 2016, pp. 866–75.

[17] E. Wulczyn and D. Taraborelli, "Wikipedia clickstream," 2 2017. [Online]. Available: https://figshare.com/articles/Wikipedia_Clickstream/1305770

[18] A. Plangprasopchok, K. Lerman, and L. Getoor, "A probabilistic approach for learning folksonomies from structured data," in *Proceedings of the 4th ACM Web Search and Data Mining Conference* (*WSDM*), Feb. 2011. [Online]. Available: http://arxiv.org/abs/1011.3557

[19] E. M. Talley, D. Newman, D. Mimno, B. W. Herr, H. M. Wallach, G. A. P. C. Burns, A. G. M. Leenders, and A. McCallum, "Database of NIH grants using machine-learned categories and graphical clustering," *Nat. Meth.*, vol. 8, no. 6, pp. 443–4, Jun. 2011. [Online]. Available: http://dx.doi.org/10.1038/nmeth.1619

[20] Y. Liu, Z. Huang, Y. Yan, and Y. Chen, "Science navigation map: An interactive data mining tool for literature analysis," in *Proceedings of the 24th International Conference on World Wide Web*, ser. WWW '15 Companion. New York, NY, USA: ACM, 2015, pp. 591–6. [Online]. Available: http://doi.acm.org/10.1145/2740908.2741733

[21] P. McQuilton, A. Gonzalez-Beltran, P. Rocca-Serra, M. Thurston, A. Lister, E. Maguire, and S. A. Sansone, "BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences," *Database: the journal of biological databases and curation*, 2016.

[22] J. Ison, M. Kalaš, I. Jonassen, D. Bolser, M. Uludag, H. McWilliam, J. Malone, R. Lopez, S. Pettifer, and P. Rice, "EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats," *Bioinformatics*, vol. 29, no. 10, pp. 1325–32, 2013. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btt113

José Luis Ambite received the M.S. and Ph.D. degrees in computer science from the University of Southern California in 1998. He is currently an Associate Research Professor in Computer Science, and a Research Team Leader at the Information Sciences Institute, both at the University of Southern California. His expertise is in information integration, including query rewriting under constraints, learning schema mappings, and entity linkage. His research interests include databases, knowledge representation, semantic web, semantic search, automated planning and learning, scientific workflows, and biomedical informatics. His current focus is on biomedical data science.

Lily Fierro received the S.B. degree in brain and cognitive sciences from the Massachusetts Institute of Technology in 2013. She is currently a business data analyst at the University of Southern California Information Sciences Institute. Her research interests include databases, ontology design, machine learning, and designing effective data-oriented applications, all as applied to problems in health, medicine, and education.

Jonathan Gordon received the B.A. degree in computer science from Vassar College in 2007 and the M.S. and Ph.D. degrees in computer science from the University of Rochester in 2009 and 2014 respectively. He was a postdoctoral researcher at the University of Southern California Information Sciences Institute and is now a Visiting Assistant Professor of Computer Science at Vassar College. He works on the problems of learning and organizing knowledge from text.

Gully A.P.C. Burns received the B.Sc. degree in physics from Imperial College in England in 1991 and the D.Phil. degree in physiology from Oxford University in 1997. He is currently a Project Leader at the University of Southern California Information Sciences Institute. His research interests are effectively encapsulated within the formulation of the domain of "Discovery Informatics" that is an attempt to bring together many facets of AI research in service of scientific research. Within this work, he is focused on constructing tools and approaches that support the cycle of scientific investigation by accelerating it, tracking its provenance, populating it with data, and automating processes of reasoning within it.

Florian Geigl received the B.Sc. and M.Sc. degrees in computer science in 2011 and 2013 respectively from Graz University of Technology. He received the Ph.D. degree in computer science in 2017 from the Institute of Interactive Systems and Data Science at the Graz University of Technology. He was a Visiting Ph.D. Student at the University of Southern California Information Sciences Institute in 2016. From 2017 to 2018, he was a data scientist for Detego. Currently, he is the Chief Digital Officer (CDO) of IBS Austria.

Kristina Lerman received the A.B. degree in physics from Princeton University in 1989 and the Ph.D. degree in physics from University of California at Santa Barbara in 1995. She is currently an Associate Research Professor in the Computer Science Department at the University of Southern California and a Project Leader at the University of Southern California Information Sciences Institute. Her research includes statistical text analysis, semantic modeling of data, and mathematical modeling of multi-agents systems, as well as social networks and social computing. Her current work revolves around deciphering the structure and dynamics of social media and crowdsourcing platforms, such as Twitter, Digg, Facebook, and Stack Exchange among others. She is an active member of the social computing research community, as a chair (SocInfo'14, SocialCom'14, Hypertext'13) or senior PC (ICWSM, IJCAI) of leading conferences, and has organized several computational social science workshops.

John D. Van Horn is an Associate Professor of Neurology with additional appointments in Neuroscience and in Electrical Engineering at the University of Southern California (USC). Prior to joining the faculty at USC in 2013, he was a faculty member in the Neurology Department at the University of California Los Angeles (UCLA), and, prior to that, was faculty in the Department of Psychological and Brain Sciences at Dartmouth College. He received the Ph.D. degree in psychology from the University of London in England in 1992 for his studies performed on the subjects of phenotypic expressions of brain abnormalities using neuroimaging and biobehavioral metrics. He conducted his post-doctoral training in the area of human brain imaging using PET and functional and structural MRI at the National Institute of Mental Health. He also received the M.S. degree in engineering from the University of Maryland, College Park in 2000. From 2002 to 2006, he directed the fMRI Data Center, as well as a high performance neuroimaging, computational, analysis, and visualization facility based at Dartmouth College. At UCLA from 2007-2013, he was a professor in the Department of Neurology and served on the executive committee for the Staglin Center for Cognitive Neuroscience imaging center. Based now at the USC, he currently leads a team of researchers focused on the informatics associated with large-scale biomedical data. He has authored over 125 publications (h-index >45), and he is known internationally as an expert in neuroinformatics and data sharing. His current research includes multimodal neuroimaging in clinical populations, databasing, and data mining.