Carnegie Mellon University

TALKS & ABSTRACTS

MONDAY, MAY 13

Welcome & Opening Remarks:

Keith Webster, Dean, Carnegie Mellon University Libraries
Michael McQuade, Vice President for Research, Carnegie Mellon University
Beth A. Plale, Science Advisor, CISE/OAC, National Science Foundation

Keynote 1:

Tom Mitchell, Interim Dean, E. Fredkin University Professor, School of Computer Science,
Carnegie Mellon University

Title: Discovery from Brain Image Data

Abstract: How does the human brain use neural activity to create and represent meanings of words, phrases, sentences and stories? One way to study this question is to collect brain image data to observe neural activity while people read text. We have been doing such experiments with fMRI (1 mm spatial resolution) and MEG (1 msec time resolution) brain imaging, and developing novel machine learning approaches to analyze this data. As a result, we have learned answers to questions such as "Are the neural encodings of word meaning the same in your brain and mine?", and "What sequence of neurally encoded information flows through the brain during the half-second in which the brain comprehends a word?" This talk will summarize our machine learning approach to data discovery, and some of what we have learned about the human brain as a result. We will also consider the question of how reuse and aggregation of such scientific data might change the future of research in cognitive neuroscience.

Session 1: Automation in data curation and metadata generation

Session Chair: Paola Buitrago, Director for AI and Big Data, Pittsburgh Supercomputing Center

1.1 Keyphrase extraction from scholarly documents for data discovery and reuse

Cornelia Caragea and C. Lee Giles

Abstract: The current Scholarly Web contains many millions of scientific documents. For example, Google Scholar was estimated to have more than 160 million documents in 2014 and Microsoft Academic expanded from 83 million records in 2015 to 168 million in 2017. Open-access digital libraries such as CiteSeerX , which acquire freely-available research articles from the Web, witness an increase in their document collections as well. These rapidly-growing scholarly document collections offer benefits for knowledge discovery, learning, and staying up-to-date with recent research advances. However, navigating in these digital libraries and finding useful information in the huge amounts of text have become very challenging. Keyphrase extraction is the task of automatically extracting a small set of keyphrases from a text document to concisely summarize its content and can allow for efficient information processing and data discovery. In addition, reuse and applications of keyphrases are very diverse ranging from text summarization to contextual advertisement, topic tracking over time, discovery and evolution of science, indexing, searching, query formulation, identifying reviewers for submissions, recommending articles to readers in digital libraries, or simply as aid for navigation through large text corpora.

Despite their strong value, manually annotated keyphrases are not always provided with the documents, but they need to be gleaned from the content of these documents. In this project, we discuss artificial intelligence (AI) approaches to keyphrase extraction from scholarly documents and study what types of information can aid keyphrase extraction and what aspects of this task seem to be the most difficult. Among the AI approaches, we highlight the benefits of neural learning models that capture not only label dependencies, but also hidden semantics in text. Moreover, we will discuss AI approaches to keyphrase extraction that use information from citation contexts in novel ways. These citation contexts, or the short text segments surrounding a citation mention, which are currently available from the CiteSeerX digital library, are not arbitrary, but they serve as brief summaries of a cited paper and can be effectively exploited to improve the performance of supervised and unsupervised keyphrase extraction. Through our research, we identify aspects that bring additional challenges such as incomplete and/or noisy data and the subjectivity of the annotation process. Last, we present several applications of keyphrases for data discovery and reuse, including researcher homepage discovery, topic classification, and keyphrase boundary detection and show how keyphrases can aid these tasks.

1.2 Dynamic System Explanation, DySE, a framework that evolves to reason about complex systems

Cheryl Telmer, Khaled Sayed, Adam Butchy, Kara Nicole-Simms Bocan, Emilee Holtzapple, Casey E. Hansen, Gaoxiang Zhou, Yasmine Ahmed and Natasa Miskov-Zivanov

Abstract: The scientific literature contains data collected from decades of research. In these papers the authors consider background knowledge, methods and results, and then discuss and interpret their findings in a peer-reviewed published manuscript. This voluminous knowledge can be mined using natural language processing and utilized to automatically construct models of complex networks in order to obtain a greater understanding of the system. The Dynamic System Explanation, or DySE, framework uses methods developed for circuit design automation to configure models and execute simulations over time. The use of discrete update rules allows for the modeling of systems where precise rates of change are unknown. DySE can use the output of natural language processing (NLP) engines that extract directed causal interactions from text or can use inputs from databases or directly from users. The knowledge is represented in a tabular format that is simple for humans and machines to read, interact with, and for the inclusion of common sense connections that may not be stated explicitly in the literature. The tabular schema utilizes a notation to represent various logical or algebraic

relationships between elements and regulators, and numerical data is discretized to parameterize the model. Additionally, in this tabular schema, users can adjust relationships between elements that machines may have difficulty recognizing automatically. The Discrete Stochastic Heterogeneous, DiSH, simulator outputs stochastic trajectories of state changes for all model elements, thus providing a means for thousands of in silico experiments in a matter of seconds. Numerous scenarios, created for testing hypotheses or interventions, are automatically evaluated, utilizing the statistical model checking approach, and known or desired system properties. DySE can automatically extend models when additional knowledge is available, allowing for the exploration of hypotheses or interventions. Model extension is integrated with model checking to test the validity of additional interactions so that the model is iteratively expanded. This has proven to be a very powerful method for creating models to incorporate data or explain correlations in data with causative mechanisms. The directed nature of the interactions, ability to incorporate feedforward and feedback loops, and iterative selection to grow the model, form a framework that learns and reasons about data to provide explanations rooted in knowledge. Automated reading and incorporation of data, a standardized notation in an interactive tabular schema, coupled with iterative growth and testing of models and scenarios, results in an intelligent system for learning and reasoning from text and data.

1.3 Searching for similarities and anomalies in a pool of galaxy images using Deep Learning

Matias Carrasco Kind

Abstract: Large Astronomical surveys like the Dark Energy Survey and the Large Synoptic Survey Telescope are generating an overwhelming amount of catalog data and images. With the current advances in Big Data and Machine Learning there are modern techniques to store, classify and extract some useful information from such data. There are a large number of Deep Learning algorithms, supervised and unsupervised, to classify galaxy images but there has been very little done in the area of data discovery based on images. For example, the Dark Energy Survey has imaged over 300M galaxies, and there are no current methods to find galaxies based on their image similarity or find anomalies within the data from the image point of view. In this talk, I'll show what approaches we are using to leverage current Deep Learning techniques including CVAE and GAN for setting up a discovery service for galaxy images. This will allow scientist to rank galaxy images based on a similarity search (for a given query image), to find anomalies and to constrains a query service in order to find galaxies with properties which otherwise are lost in the catalogs created from those images. The powerful combination of catalog and images in one query interface will enable the scientific discovery. I will show how these techniques can be easily expanded to other fields where images carry important information.

1.4 Ask the Doctor if YouTube is Right for You: An Augmented-Intelligence Video Recommender System for Patient Education

Xiao Liu, Anjana Susarla and Rema Padman

Abstract: The availability of health information on blogs, social networks, YouTube, Twitter, and hospital review sites presents an unprecedented opportunity to examine how social media can facilitate patient-centric health and to investigate how social media can be a channel to inform and communicate healthcare information to patients. YouTube hosts over 100 million healthcare related videos on a variety of medical conditions. The plethora of user-generated content could be leveraged by patients to improve adherence to evidence-based guidelines and self-care required for a variety of chronic diseases.

Alongside, several healthcare providers and government agencies have expressed concern about the quality and reliability of healthcare information available through YouTube. The existing studies on healthcare information on YouTube rely primarily on small samples and qualitative evaluation of informational quality and reliability. It is critical for visual social media platforms to provide information retrieval capabilities to enable health consumers to efficiently access helpful medical knowledge and avoid misleading health information. In practice, though, limited interventions have been developed to improve the process by which users obtain useful medical information through video search.

With patients regularly turning to YouTube for health advice, we propose an augmented intelligence-based approach that effectively combines human input from domain experts and computational methods to recommend relevant video materials to patients. The problem of identifying the most relevant videos from a patient perspective provides an immense innovation space for this approach. YouTube provides metadata about videos in a standard and semi-structured format, which can be processed by computers. Assessing the value of healthcare videos still requires domain expertise to gauge the medical information in the videos. At the same time, readability, content organization, and presentation are critical to healthcare consumers. We leverage a co-training machine learning framework and incorporate inputs from patient education experts and clinicians to assess videos on two dimensions: the amount of medical information in the videos and video understandability. We develop a user-centric patient education video recommender system by integrating these two dimensions with the YouTube video ranking results. Recommending relevant materials leveraging user-generated content is one way to deliver personalized healthcare information. As technology continues to advance and evolve, our methods can be used to improve patient education to benefit and empower patients, physicians, and the broader healthcare community.

1.5 Image Recognition for Archaeological Research

Claudia Engel, Peter Mangiafico, Justine Issavi and Dominik Lukas

Abstract: This project focuses on the image repository of the Çatalhöyük Archaeological Project (catalhoyuk.com). Inscribed on the UNESCO World Heritage List in 2012 the 9000-year-old neolithic settlement of Çatalhöyük in central Turkey is widely recognized as one of the most important archaeological sites in the world. Since 1993 an international team of archaeologists has been carrying out excavations and research. Today, after 25 years of research, the project has accumulated close to 5TB of data, including an image repository with a total of about 150,000 images, which is currently being ingested for long term archiving into the Stanford Digital Repository (sdr.stanford.edu).

The images are used to identify artifacts, to document the excavated objects in their excavation contexts, and to record the excavation process, which is by nature destructive. However, extracting the wealth of relevant information from these images for research remains a challenge for several reasons. In many cases, metadata recorded with the images are incomplete and inconsistent. Secondly, researchers require access to information captured in the images that is not contained in the metadata. And lastly, the number of images cannot be reasonably processed by hand.

While machine learning has been applied in archaeology it has typically focused on single objects and patterns to support archaeologists in their assessment and classification of individual archaeological finds. Information about the context these objects are found in, i.e. the relation among artifacts and between artifacts and their

excavators, is largely inaccessible. Context and process information are fundamental for archaeological research and the composition of archaeological photographs can be linked to specific practices of archaeology.

We apply existing image recognitions models (Google Vision API and Clarifai Predict API) on our images and train custom machine learning models from a subset of labeled images with Google AutoML. We report on our initial explorations discuss their usefulness for extracting and enriching metadata for discovery and research.

Session 2: Automation in data discovery

Session Chair: Huajin Wang, Biomedical Data Science Liaison, Carnegie Mellon University Libraries

2.1 Google Dataset Search: An open ecosystem for data discovery

Invited Talk: Natasha Noy, Staff Scientist, Google AI

Abstract: Google Dataset Search enables users to find datasets across thousands of repositories on the web. They range from various scientific disciplines, such as environmental or social science, to government or economic data from around the world. I will discuss the ideas behind Dataset Search, the open ecosystem for describing and citing datasets that we hope to encourage, and the technical details on how we went about building Dataset Search. I will conclude with the discussion of technical challenges both in using datasets for AI applications and for using AI and ML to help in dataset discovery.

2.2 A Dataset Search Engine for Data Augmentation

Fernando Chirigati, Remi Rampin, Aecio Santos and Juliana Freire

Abstract: Data augmentation is the task of augmenting, enlarging a dataset with the goal of improving the performance of a machine learning task, potentially by employing external datasets to add more columns (new features) or more rows (new data points) to the original dataset (training and testing data). For structured data, this involves discovering datasets that can be joined or unioned with the original one. Augmenting structured data, however, comes with myriad challenges. First, the sheer number of datasets available on the Web makes it hard to find relevant data: manually searching for them is cumbersome and data scientists can easily miss a large fraction of relevant data sources. Second, looking for potential join and unions is far from trivial and depends on several database techniques, from schema matching to data fusion, and given the large amounts of data, this task is vastly computationally expensive. Also, as Web data often come with incomplete metadata, methods are needed to efficiently extract metadata that can be used for discovery, in particular, to identify data that can be joined and unioned. Last but not least, search queries must ideally have nearly-interactive response times and results must be properly ranked to avoid overwhelming users with numerous result datasets. In this talk, we will introduce our ongoing work on a dataset search engine tailored to data augmentation. We will describe our first steps in solving some of the aforementioned challenges, outline the different interfaces and APIs for accessing the search engine, and discuss the many avenues for future research. This work is part of the DARPA D3M program for developing automated model discovery systems, and it has already been used by different performers in the program. The tool is currently available at https://datamart.d3m.vida-nyu.org/ (GitHub: https://gitlab.com/ViDA-NYU/datamart/datamart).

2.3 A Semantalytic Approach to Accelerated Data Reuse for Reproducible Scientific Discovery

Alexander New, Shruthi Chari, Miao Qi, Sabbir M. Rashid, John S. Erickson, Deborah L. McGuinness and Kristin P. Bennett

Abstract: We develop a semantics-driven, automated approach for dynamically performing rigorous scientific studies. This framework may be applied to a wide variety of data and study types. Here, we demonstrate its suitability for conducting retrospective cohort studies using openly available population health data, including identifying lifestyle factors associated with chronic diseases, or genetic mutations associated with diseases in precision health. Our semantically-targeted analytics (semantalytics) approach addresses the end-to-end data science workflow, ranging from intelligent data selection to dissemination of derived data and results in a rigorous, reproducible way. Our approach seeks to eliminate what we see as the critical bottleneck in this process: the availability of skilled data analysts. Ideally, the data analyst must translate an ill-defined problem from a client and one or more datasets to an analysis task; transform the data; design an appropriate modeling and validation approach; and optimize and execute a workflow to create and validate the analytical results. Ultimately, the data analyst translates the results into viable information products, including reports, texts, and presentations documenting the study, and disseminations of the results via mechanisms appropriate to the target domain. Our semantalytics approach treats the end-to-end data science workflow as data itself, semantically represented using domain-specific and authoritative ontologies in a comprehensive knowledge graph.

Our evaluation is motivated by the challenge of identifying risk factors that, for a given cohort, have significant associations with some health condition. Our solution is a semantics-augmented machine learning system that applies a novel health analysis ontology and knowledge graph to dynamically discover risk factors for selected cohorts. Semantalytics introduces the notion of cartridges: application-specific fragments of the underlying knowledge graph that extend its analytic capabilities. Defined using ontologies, the cartridges capture the necessary metadata for source data, workflows, and results/derived data. Cartridges enable an automated architecture allowing analysts to dynamically conduct studies exploring different health outcomes, risk factors, cohorts, and analysis methods. Semantalytics extends existing automated machine learning approaches by enabling the analysis pipeline to be configured from interchangeable modules in a reusable, domain-specific way. We apply the semantalytics framework to generalized risk analysis using the National Health and Nutrition Examination Survey, demonstrating that it can both reproduce existing studies and discover new findings. Semantalytics supports interoperability, reusability, reproducibility, and explainability within the automated system. The ontology and semantic framework developed here can be readily extended to other learning tasks and datasets in the future.

2.4 An innovative approach to scalable semantic search

Shenghui Wang, Rob Koopman, Titia van der Werf and Jean Godby

Abstract: Semantic search, in addition to keyword based search, is a desirable feature for many digital library systems. Even in the largely structured library data world, there is still a lot of tacit information locked in the free-text fields. Embedding words and texts in compact, semantically meaningful vector spaces allows for computable semantic similarity/relatedness which would make search more intelligent. The recent success of local context predictive models such as Word2Vec has initiated the development of more complex and powerful deep learning models for embedding words, sentences and documents. The resulting embeddings combine compactness and discriminating ability, but the associated computational requirements are substantial (often requiring powerful machines with GPUs) and the optimal hyperparameter settings are not

easy to find. It is, therefore, more common that embeddings are pre-trained on large corpora and plugged into a variety of downstream tasks, as in sentiment analysis, classification, translation, and so on. However, such transfer learning might fail to capture domain-specific semantics, which can be crucial for certain applications, such as medical information retrieval, special collection exploration, etc. Standard benchmarks and evaluation methods often do not answer practical needs either.

Unfortunately, most libraries and even large-scale aggregators do not have the processing capacity nor the skills to embrace powerful deep learning and therefore stick with the traditional keyword-based approach. In our quest for practical solutions to support libraries in this field, we revisit the global co-occurrence based embedding and propose a conceptually simple and computationally lightweight approach. Our method extends random projection by weighting and projecting raw term embeddings orthogonally to an average language vector. As a result, the discriminating power of the term embeddings is increased, and even more meaningful document embeddings can be built by assigning appropriate weights to individual terms. We describe how updating the term embeddings online, as we process the training data, results in an extremely efficient method, in terms of both computational and memory requirements. Our experiments show highly competitive results with various state-of-the-art embedding methods on different tasks, including the standard Semantic Textual Similarity (STS) benchmark and an extreme multi-label classification (automatic subject prediction) task, at a fraction of the computational cost. Our method also has important practical benefits, including increased speed, more modest hardware requirements and effective handling of very rare words.

2.5 Building Specialized Collections from Web Archiving

Cornelia Caragea and Mark Phillips

Abstract: A growing number of research libraries and archives around the world are embracing web archiving as a mechanism to collect born-digital material made available via the Web. Between the membership of the International Internet Preservation Consortium, which has 55 member institutions, and the Internet Archive's Archive-It web archiving platform with its 529 collecting organizations, there are well over 584 institutions currently engaged in building collections with web archiving tools. The amount of data that these web archiving initiatives generate is typically at levels that dwarf traditional digital library collections. As an example, in a recent impromptu analysis, Jefferson Bailey of the Internet Archive noted that there were 1.6 Billion PDF files in the Global Wayback Machine. If just 1% of these PDFs are of interest for collection organizations and can be reused, that would result in a collection larger than the 15 million volumes in HathiTrust.

While the number of web archiving institutions increases, the technologies needed to provide access to these large collections have not improved significantly over the years. At this time, the standard way of accessing web archives is with known URL lookup using tools such as the OpenWayback. The use of full-text search has increased in many web archives around the world, but often provides an experience that is challenging for users because of the vast amount of content, and large heterogeneous collections. Another avenue of access to web archived data that is of interest to web archiving institutions is the ability to extract high-quality, content-rich publications from the web archives in order to add them to their existing collections and repository systems.

Our research is conceived to understand how well Artificial Intelligence can be employed to provide assistance to collection maintainers who are seeking to classify extracted PDF documents from their web archive into being within scope for a given collection or collection policy. By identifying and extracting these documents,

institutions will improve their ability to provide meaningful access to collections of materials harvested from the Web that are complementary, but oftentimes more desirable than traditional web archives. Our research focus is on three generic use cases that have been identified for the reuse of web archives and include populating an institutional repository from a web archive of an university domain, the identification of state publications from a web archive of a state government, and the extraction of technical reports from a large federal agency.

2.6 Reuse and Discovery for Scholarly Big Data

C. Lee Giles and Jian Wu

Abstract: Over the past 20 years, CiteSeerX has created searchable scholarly big datasets and made them freely available for research. In order for digital library search engines like CiteSeerX to be scalable and effective, AI technologies are employed in nearly all essential components of the system, including document classification, metadata extraction, author name disambiguation, data linking, etc. To effectively train these, a sufficient amount of labeled data is generated, providing the ground truth for specific training and evaluation tasks. For example, the document type corpus, randomly selected from the focused crawl repository, includes 3000 PDF documents, manually labeled as papers, theses, slides, books, resumes, and other. The corpus also includes URLs where these documents were downloaded. These ground truth datasets can be reused for training improved supervised and semisupervised models. Figures, tables, algorithms, and math equations extracted from scholarly papers can be reused for research. They can also be reused as a derived scholarly big dataset to design and evaluate images, compression, classification, ranking algorithms, and others. Rich semantic information, such as keyphrases and scientific entities, can also be reused for tasks such as summarization and automatic author profile generation.

We discuss how to discover datasets in scholarly articles and make them more available via a search interface. Although there are open access repositories such as figshare, we believe that a large number of datasets are mentioned in scholarly papers. We explore natural language parsing algorithms to find the datasets that are actually used in papers. The datasets, once discovered, will be associated with research topics and sorted chronologically. Now when users search for a research topic, the search engine returns all datasets related to this topic and its popularity over time. The goal is to free researchers from a large amount of readings and provide them with an overview of data usage over time, helping them to explore and make decisions on datasets they choose for specific purposes. To this end, CiteSeerX becomes a suitable platform since it has the full-text data of over 10 million documents with a goal is to ingest all open access scholarly papers, estimated to be 30-40 million. We have designed machine learning algorithms to link these documents to the Web of Science, Medline, and other digital libraries. Users will have more opportunities to explore a comprehensive history of datasets that can be reused and discover other datasets for new research projects.

Panel 1: Challenges and Opportunities in Data Reuse Using the Power of AI

Keith Webster, Dean, Carnegie Mellon University Libraries **(Moderator)**
Cliff Lynch, Executive Director, Coalition for Networked Information
Natasha Noy, Staff Scientist, Google AI
Casey Greene, Assistant Professor of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania
Alex London, Clara L. West Professor of Ethics and Philosophy, Department of Philosophy, Carnegie Mellon University

TUESDAY, MAY 14

Keynote 1:
Glen de Vries, President and Co-founder, Medidata Solutions

Title: New Evidence Models: Clinical trials in the age of AI and Precision Medicine

Abstract: The "gold standard" by which physicians, scientists, and regulators have evaluated safety and efficacy in clinical research, the randomized controlled trial, has been consistent for over two hundred years. Today, digital transformation and AI are making new evidence models possible. These evidence models address both practical and ethical issues in human research. We will examine those issues, how precision medicine make them more acute, how digital and AI are allowing for new comparators and unprecedented data reuse, and how these new experimental and statistical models are -- at long last -- allowing that gold standard to evolve.

Session 3: Integrating datasets and enabling interoperability

Session Chair: Nick Nystrom, Chief Scientist, PSC and Carnegie Mellon University

3.1 Data reuse through domain adaptation AI algorithms for the self driving industry

Evgeny Toropov, Alex B. Weiss and José M.F. Moura

Abstract: Currently, many companies and research groups rely on annotated image data to solve a multitude of tasks in the computer vision domain. For example, the self-driving industry requires images annotated with rectangles drawn around each car or pedestrian to feed the computer vision algorithms. The academic community released a number of publicly available well-annotated datasets that could be reused. However, despite being similar in general, images from public datasets differ from images of any particular group in details. For example, traffic signs and lane markings differ across countries, cameras differ in the field of view or position on the car. The combination of such factors reduces the performance of computer vision algorithms trained on public datasets. One big step towards overcoming such problems and reusing previously collected data is a set of AI algorithms collectively called domain adaptation. The goal of domain adaptation is to generalize to the target domain while being trained on annotations available in the source domain. Here, we describe our experience of applying computer vision to detect road in images taken from a car. We succeeded in reusing annotated public datasets by the means of adopting domain adaptation AI tools.

3.2 A self-organized Scenario-based Heterogeneous Traffic Database for Autonomous Vehicles

Jiacheng Zhu, Wenshuo Wang and Ding Zhao

Abstract: A multitude of open driving datasets and data platforms have been raised for autonomous vehicles (AV). However, the heterogeneities of datasets in structure and driving context make existing datasets practically ineffective due to a lack of uniform frameworks and searchable indexes. In order to overcome these limitations on existing public datasets, we proposed a data unification framework based on traffic primitives with the ability to automatically unify and label heterogeneous traffic data. This is achieved by two steps: 1)

Carefully arrange raw multidimensional time series driving data into a relational database and then 2) automatically extract labeled and indexed traffic primitives from traffic data through a Bayesian nonparametric learning method. Thus, We make it possible for establishing an accessible framework capable of automatically arranging, unifying and storing heterogeneous driving data.

The heterogeneity of driving data can be categorized into (1) Between existing datasets; (2) Within datasets; and (3) Between completed datasets and on-going collecting datasets. The heterogeneity between existing datasets is caused by diversity in driving settings and data collection platforms since most existing datasets differ from each other in terms of format, traffic context, application, sensor statue, and support tools. The heterogeneity also exists within one single dataset. A variety of traffic scenarios should be contained to investigate and test algorithms under different working conditions, which will cause the diversity in vehicle behaviors and sensor records and hence two parts from this dataset may greatly different from each other. Last but not least, the existing datasets could also be greatly different from the on-going collection data due to uncertainty and diversity of driving scenarios.

In order to build a dataset unification framework to accelerate autonomous driving technologies, we propose a traffic primitive-based framework. First, information in the dataset will be rearranged into time series according to data attributes and timestamps. Then a nonparametric Bayesian learning approach will be applied to extract the primitives, which is believed to be fundamental buildings block of traffic scenarios, by segmenting the multidimensional time series. After classification and learning statistical and physical properties of primitives, the traffic scenarios finally will be defined as the composition of primitives, and thus all the data could be labeled and indexed according to traffic scenarios. Since this database framework is able to recognize traffic scenarios automatically without human prior knowledge, the autonomous vehicle driving data can be self-organized.

3.3 Model Tracking using the Keras API - Simple Metadata Management

Catherine Ordun

Abstract: Data scientists rely on open source frameworks for building complex “blackbox” neural network algorithms. There are two reasons for why model explainability is important - the first is to increase human trust and insight on algorithm decision-making, and the second is to be able to reproduce results following a scientific method. This abstract focuses on the latter, demonstrating how existing functionality of a popular open source framework can be used to improve the organization and tracking of model development, similarly to what is available in expensive, proprietary tools.

Keras is the second most popular deep learning framework according to a 2018 analysis [1], followed by TensorFlow, largely for its “API designed for human beings, not machines”. While thousands of data scientists across the world use Keras, this framework is typically used for prototyping due to the ease of syntax and ability to quickly learn how to use the sequential and functional APIs. However, as data scientists build out prototype models, they quickly lose track of the algorithms they build, much less develop a systematic and organized way to track hyper-parameters, components of each layer, accuracy and classification error, gradients and more. Recently proprietary frameworks such as MXNET used for Amazon Web Services’ Sage Maker platform offer powerful systems to manage metadata and provenance information [2], similar to other software that leverages automated machine learning such as DataRobot [3] and H2O [4]. This abstract introduces simple functionality intrinsic to the Keras API to extract these metadata including hyper parameters, model meta-data, evaluation, and prediction metadata, so that data scientists have a practical way to develop their own "leaderboard" of models to compare and contrast, without added cost.

3.4 Visual and Statistical Analysis and Comparison of Handwritten and Font Datasets

Daniel Clothiaux and Ravi Starzl

Abstract: Previous work of optical character recognition of typed forms will often get strong results by training over synthetic data, created by rendering words from fonts. Similar methods have been used to create handwriting data from fonts, especially in Chinese and Japanese, but such methods in English are often done ad hoc over small hand selected font sets that look like handwriting, while OCR of typed forms will use much larger datasets. Here, we looked at comparing large scale font datasets with handwritten datasets, as a first step towards integrating them into large scale synthetic data. We did this by first collating large numbers of both fonts and 'handwritten fonts', that is, fonts created to look like handwriting. We then use these fonts to create a large database consisting of all the numbers 0-9 as well as both uppercase and lowercase letters a-z. We then ran a large series of qualitative and quantitative comparisons between them and human handwritten data from the full NIST SD 19 dataset, which we use to categorize characters (such as seven written with and without a line through it). Our qualitative comparisons are visually oriented. For example, we calculated heatmaps for handwritten characters, fonts, and handwritten fonts, which allows for visual comparisons where some differences are immediately visible, such as slant or most handwritten fours being closed, and most font fours, even handwritten fonts, being open. We also cluster characters using PCA and t-SNE, and create visualizations of these reduced data, both directly applied on the font, and applied on the final hidden layer of a recognition convolutional neural network. Our quantitative methods encompass comparing statistics such as character orientations, number of loops, eccentricity, bounding ellipse major and minor axis, density, and aspect ratio, as well as results recognition of fonts on handwritten data. We find that although handwritten fonts are in general closer to actual handwriting than the full font set, it still has distinct characteristics of fonts-such as the aforementioned four, or letters with stems on the left (lowercase b and h) having a much stronger trend of leaning left in real handwriting. The statistics and especially visualizations are also interesting in and of themselves.

3.5 Lowering the barriers to experiment, data, and method reproducibility in AI research with a cloud-based computational reproducibility platform

Xu Fei and April Clyburne-Sherin

Abstract: The progress of AI research relies on robust methods. The explosion of interest in the field of AI resulted in many more publications in the past decade. However, this tremendous growth of AI research activity makes evaluating the robustness of a publication increasingly difficult and time consuming. The robustness of computational publications is reflected by 3 degrees of reproducibility: experiment, data, and method. Experiment reproducibility means that after applying the same data and implementation of the method from the original authors, we are able to get the same results. Data reproducibility means after applying an alternative implementation of the method on the same data from the original authors, we are able to get the same results. Method reproducibility means after applying an alternative implementation of the method on different data from the original authors, we are able to get consistent results.

Although experiment reproducibility confirms the results are indeed reproducible on the specific data provided by the same research team, it only suggests that the generality of the results is limited to the specific implementation and dataset provided by the original authors. However, method reproducibility confirms that the results are valid when re-implemented on a different dataset, hence the method is more generalizable and more robust.

Traditionally, the reproducibility of a publication depends on the documentation provided by the original research team. Unpublished code and unpreserved public data both contribute to the frustrations that researchers experience when trying to evaluate a publication. Moreover, the rapidly changing technological infrastructure makes implementing the method even with original authors' code and data an unnecessarily time-consuming task.

Our talk will present Code Ocean -- a computational reproducibility platform designed to help researchers evaluate any publication more efficiently by addressing all 3 degrees of reproducibility. Code Ocean uses container technology along with designs rooted in computational research. Any work published on Code Ocean contains code, data, and the computational environment exactly the same as the original authors and executable in the browser. Hence publications on Code Ocean are experimentally reproducible by default. This provides a great starting point for researchers to investigate whether they want to investigate more resources to validate higher degrees of reproducibility of a publication. Code Ocean also supports the reuse and remix of existing publications, allowing a novel way of showing the robustness of a study from the number of successful attempts at experiment, data, and method reproducibility.

3.6 LabelBee: a web platform for large-scale semi-automated analysis of honeybee behavior from video

Rémi Mégret, Ivan Rodriguez, Isada Claudio Ford, Edgar Acuna, Jose L Agosto and Tugrul Giray

Abstract: The LabelBee system is designed to facilitate the collection, annotation and analysis of large amounts of honeybee behavior data from video monitoring. It is developed as part of NSF BIGDATA project "Large-scale multi-parameter analysis of honeybee behavior in their natural habitat", where we analyze continuous video of the entrance of bee colonies. One key aspect is the presence of bees with barcode tags, which can be identified to individually to monitor the variability of their individual behavior on the long-term (from days to seasons). Due to the large volume of data and its complexity, LabelBee provides advanced AI and visualization capabilities to enable the construction of good quality datasets necessary for the discovery of complex behavior patterns. It integrates several levels of information: raw video, honeybee positions, decoded tags, individual trajectories and behavior events (entrance/exit, presence of pollen, fanning, etc.). This integration enable to combination of manual and automatic processing, where automatic results can be corrected and extended by multiple users connecting to the platform.

This flexibility facilitates the close collaboration between Biologists and Computer Scientists. Biologists, as users of the platform, can annotate the behaviors of interest by looking directly at the videos, which is an improvement over the traditional approach of direct observation in the field. Typical dataset creation involves three steps: first step is the main annotation, which involves several users working on intervals of one hour of video, the second step is the cleaning by a different user, the last step is curation by an expert user who ensures uniform standards. Visualization aids such as highlighting potentially redundant annotation helps with this process. Computer Scientists, as contributors to the platform capabilities, have developed several additional AI modules to automatize part of the analysis. First, automatic tag detection helped focus the

attention of the users on the precise times where tagged bees are passing through the entrance. Then, modules for the automatic detection of untagged bees and pollen were trained based on already annotated data. We are currently in the process of integrating these newly trained modules in the platform, to accelerate the main annotation step. The data constructed by this semi-automatized approach can then be exported for the analytic part taking place on the same server: long-term behavior datasets can be visualized and manipulated through Jupyter notebooks for the extraction and exploration of behavior patterns.

Session 4: Biomedical applications

Session Co-chairs:
Sean Davis, Senior Associate Scientist, National Cancer Institute, NIH
Andreas Pfenning, Assistant Professor, School of Computer Science, Carnegie Mellon University

4.1 Invited Talk: Data reuse enables ML-based analysis of rare diseases

Casey Greene, University of Pennsylvania

Abstract: We sometimes speak of "big data" in biology. In most cases, these data have many more features than examples. This is particularly pronounced in the case of rare diseases, where we may have tens of samples but tens of thousands of measurements. I'll discuss how we can reuse compendia of data with many training examples as a training dataset and then transfer the resulting model to rare disease datasets where the number of samples is particularly limited.

4.2 Prototype ML Software for Several Distinct Classes of Biomedical Data Science Problems Developed in NIH-Hackathons!

Ben Busby

Abstract: Machine Learning techniques and concepts can be explored more freely in hackathons than other settings. Here we present some examples of Machine Learning tools we have prototyped in hackathons.

The possibility of identifying the evidence of genetic engineering and other subtle indicators of DNA in the wrong context is an open scientific question. We built a prototype that distinguishes natural DNA sequences from model plant organisms from artificially modified plant DNA sequences. We think this is expandable to many other types of searches for sequences that are not in public databases.

We have developed a machine learning prediction algorithm for neurological injury in pediatric cardiac patients. We have built a framework to statistically assess actionable events leading to stroke, the biggest risk factor for Pediatric EcMo patients, to improve neurological outcomes.

We have developed a workflow to apply ensembles of machine learning methods for feature selection and selection consensus, to determine sets of best discriminating gene biomarkers using RNA-seq data from heterogeneous cancer populations (e.g. pediatric AML, etc.).

Moreover, we have used machine learning to calculate the immunogenicity of peptides and produce training images with Generative Adversarial Networks to use in image analysis workflows.

This body of work demonstrates that hackathons are a useful tool to generate machine learning prototypes.

4.3 Invited Talk: Data Privacy: Control, Use, and Governance

Lisa S. Parker, Professor and Director, Center for Bioethics & Health Law, University of Pittsburgh

Abstract: For some people, protecting their privacy is a matter of controlling the flow of information about themselves. Given the development and adoption of myriad technologies and social practices, controlling data about oneself has become an unrealistic goal. Development of systems—social, legal, and technological—to control and/or govern the use of data is more realistic and ethically necessary. Drawing on lessons from bioethics and healthcare, this talk will address basic reasons—traditional and evolving—to care about privacy and to develop systems to govern the use of data.

4.4 Invited Talk: Data engineering tools and approaches to facilitate data reuse and data science

Sean Davis, Senior Associate Scientist, National Cancer Institute, NIH

Abstract: Data engineering, the close cousin of data science, is the necessary first step in any data science project. As data projects become larger and more complex, awareness and understanding of the existing ecosystem for collecting, transforming, exploring, and labeling data at scale can accelerate the data science process. In this talk, I will present an opinionated view of data engineering and a pragmatic view of the data engineering landscape, with particular focus on scalable infrastructure and tools.

4.5 Invited Short Talk: Predicting Tissue-Specific cis-Regulatory Elements Across Mammals to Identify Potential Evolutionary Mechanisms

Irene Kaplow, Postdoctoral Fellow, School of Computer Science, Carnegie Mellon University

Abstract: The Vertebrate Genomes Project (VGP) is generating high-quality genomes from every order of mammals, enabling us to compare the DNA sequences of species whose most recent common ancestors lived tens of millions of years ago. These genomes enable us to investigate mammalian evolution, human uniqueness, and endangered species preservation. However, genome sequence alone is insufficient for identifying the connections between a species's DNA sequence and the traits the sequence controls. While every cell in the body has the same DNA, many traits that differ between species have evolved due to differences in the extent to which genes – encoded within the DNA – make RNA, which in turn make proteins, which varies across tissues. These differences arise from cis-regulatory elements (CREs), parts of the DNA usually found outside genes that help initiate the creation of RNA, and different CREs are active in different tissues. To understand how species differ, we need to identify which CREs are active in each tissue of every species. Obtaining primary tissue samples from many species is impossible due to cost, feasibility of obtaining animals, and the fact that some mammals are endangered.

We are developing novel machine learning models for using CRE activity data from a small number of species to impute CRE tissue-specificity in all of the mammals from the VGP. We are employing a widely used proxy for

CREs: Genomic regions identified in ATAC-seq experiments. ATAC-seq is a genome-wide assay used to identify regions of the genome that are accessible for interaction with proteins that regulate RNA levels. We are obtaining brain and liver ATAC-seq data from mouse, macaque, and bat. We will use this data to train our models to predict TS-CREs – regions with stronger DNA accessibility in one tissue relative to another – directly from DNA sequence. As a proof-of-concept, we have trained a convolutional neural network on TS-CREs from mouse and shown that we can successfully predict TS-CREs in parts of the human genome whose corresponding regions in mouse are not CREs in the same tissue (AUROC = 0.93, AUPRC = 0.90). We have already identified brain- and liver-specific CREs near important genes that our model predicts will not be brain- or liver-specific in a subset of mammals. We look forward to predicting TS-CREs in all of the VGP mammals and using our predictions to identify potential CRE tissue-specificity changes underlying differences between species.

4.6 Invited Talk: Standards, incentives, tools – Which are the necessities for data discovery in academia vs industry?

Fiona Nielsen, Founder and CEO, Repositive

Abstract: Based on her experiences in genomic data sharing in both the charity DNAdigest and the social enterprise Repositive, currently enabling a marketplace for data access within pre-clinical cancer drug discovery, Fiona Nielsen presents her learnings and perspective on how to move the needle for data discovery and reuse. Evaluating the scientific approach to solve data sharing challenges and calling out where it fails, as well as giving examples of communities where data sharing are successful, Fiona suggests we learn from approaches outside the academic setting to attack our challenges of data sharing to make the most impact with our AI and ML tools sets.

4.7 Invited Talk: Understanding the Role of Explainaiblity and Verification in Medical AI

Alex London, Clara L. West Professor of Ethics and Philosophy, Department of Philosophy, Carnegie Mellon University

Abstract: Breakthroughs in machine learning are enabling the creation of automated systems to perform a wide range of diagnostic and predictive tasks in medicine. Essential to securing and maintaining trust in health care providers and health systems are clear practices and procedures to ensure accountability and respect for the freedom of stakeholders from arbitrary interreference. A common proposal for achieving these goals imposes requirements like explainability or interpretability that seek, in different ways, to lay out the operation of such systems to human inspection. Because the most powerful AI systems are often “black-boxes,” these requirements may be purchased at the price of reduced predictive accuracy. In this talk I argue that such requirements are misguided in domains—such as medicine—where our theories of disease pathophysiology and drug mechanism are often precarious and under-developed. Instead, I argue that we should promote trust and accountability by clearly defining the tasks such systems can perform, the conditions necessary to ensure acceptable system performance, and rigorously validating their accuracy under those well-defined conditions in real-world contexts.

4.8 Invited Talk: Enabling Data Discoverability in the Human BioMolecular Atlas Program (HuBMAP)

Nick Nystrom, Chief Scientist, PSC and Carnegie Mellon University

Abstract: TBD

4.9 Invited Talk: AI for Biological Discovery: Data Integration and Self-Driving Instruments

Bob Murphy, Ray and Stephanie Lane Professor, Head of Computational Biology, School of Computer Science, Carnegie Mellon University

Abstract: Self-driving cars are changing transportation and setting a precedent for automated decision-making in the physical world. Following that precedent, self-driving instruments will revolutionize the way in which experimental science is done. A major task in biomedical research is creating *empirical* models that can predict, for example, how a person with a particular genetic makeup will respond to a particular drug. Predictive models are created from the results of some experiments to predict the outcome of others. The accuracy of a predictive model will in general be higher the more data is used to build it. This is where integration of as many existing data sources as possible is critical. However, we have data for only a tiny fraction of possible experiments, and there are experiments that we can't or don't want to do (such as treating a person with all possible drugs to see which works the best against their disease). We need a means of deciding which new experiments are best to do to improve our models. An important idea is to use our current predictive model to decide which experiments to do next (and rebuild the model) – an approach called *active* machine learning. The idea is to continuously try to improve the model (e.g., by choosing experiments whose outcome the model cannot confidently predict). The experiments themselves can nowadays often be executed using automated instruments. The combination of active machine learning and automated instruments leads to *automated science*: experiments are not only *executed* using robotic equipment but the *choice* of experiments is also made by computer – a combination we term "self-driving instruments." We have shown that automated science can produce an accurate model of drug effects while performing only a fraction of possible experiments. Of course, just as self-driving cars don't decide where you want to go, self-driving instruments need to be given a goal. The promise is to build useful predictive models much more efficiently (in terms of time, money and reducing duplication of effort) than is possible with current human directed efforts.

WEDNESDAY, MAY 15

Outcome and future planning meeting - all invited to participate

Moderator: Huajin Wang, Biomedical Data Science Liaison, Carnegie Mellon University Libraries

Panel 2: Enabling Smart and Safe Communities Through AI

Karen Lightman, Executive Director, Metro21: Smart Cities Institute **(Moderator)**
Robet Tamburo, Senior Project Scientist, Robotics Institute, Carnegie Mellon University
Bob Gradeck, Project Manager, Western Pennsylvania Regional Data Center
Santi Garces, Director, Department of Innovation and Performance, City of Pittsburgh

Session 5: Data security, privacy and algorithmic bias

Session Chair: Sayeed Choudhury, Associate Dean for Research Data Management, Johns Hopkins University Libraries

5.1 Invited Talk, Finding Bias, Discrimination, and Private Data Leakage in Machine Learning Systems

Matt Fredrikson, Assistant Professor, School of Computer Science, Carnegie Mellon University

Abstract: The use of machine learning to make decisions that impact people has become common, with applications to areas like criminal justice, child welfare, and medicine. This trend has generated excitement and interest, but has also been accompanied by legitimate concerns about the threat that these systems pose to values such as privacy and fairness. Among the primary factors leading to such concerns is the fact that these systems are opaque, meaning that it is difficult to explain their behavior, and in particular why a certain decision was made. Opacity poses a challenge for developers, users, and auditors who wish account for these systems' behavior.

In this talk, we show that privacy and fairness violations can often be viewed in terms of an inappropriate use of individuals' personal information. Central to this approach is a notion of "proxy use", which characterizes programs that employ strong predictors of a protected information type, rather than making direct, explicit use of the data in question. Viewing machine learning models as programs that operate on random data, we describe techniques that identify proxy use in large-scale models, and illustrate how these techniques provide a form of transparency by isolating and explaining behaviors that amount to privacy and fairness violations. We conclude by showing that these explanations can be put to use in removing certain types of undesirable behavior from models, often without compromising their performance.

5.2 Sharable Cyber Threat Intelligence Using Weak Anonymization

Lena Pons and Jeffrey Chrabaszcz

Abstract: There are several barriers to sharing cyber threat intelligence (CTI): lack of common data models, risk associated with centralized storage of unencrypted indicators, and lack of motivation for data collection beyond a single incident handling model. The cost of maintaining updated collections of CTI is also prohibitive for most industries.

Without sharing CTI it is impossible to perform longitudinal analyses, which are necessary for detecting tools, techniques and procedures which are harder than hashes and IP addresses for attackers to obfuscate and change.

We propose reducing the risk of sharing CTI by using hashing.

Common cyber incident data models have limitations, however models like MITRE's ATT&CK framework can provide insight into some areas such as kill chain phase. Providing pairs of obscured indicators and associated kill chain phase can provide enough information for intrusion detection system rule generation.

Using one-way cryptography only CTI that exists in a recipient's system will be exposed, which reduces the risk of constructing and maintaining centralized repositories of CTI.

5.3 Privacy Preserving Synthetic Health Data

Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao and Kristin Bennett

Abstract: We examine the feasibility of using synthetic medical data generated by GANs in the classroom, to teach data science in health informatics. Teaching data analysis with actual medical data such as electronic healthcare records is greatly restrained by laws protecting patient's privacy, such as HIPAA in the United States.

While beneficial, these laws severely limit access to medical data thus stagnating innovation and limiting educational opportunities. The process of obfuscation of medical data is costly and time consuming with high penalties for accidental release. Health histories recovered from deidentified data may result in harm to the subject. Research and education is biased towards the few publicly-available datasets such as the ICU dataset "Medical Information Mart for Intensive Care" (MIMIC). Our focus lies on the problem of making available to medical students and researchers a wider variety of medical datasets, by creating synthetic data which retains utility for teaching purposes, and ideally even for research, while definitively preserving privacy. Our proposed workflow consists of training a generative model of synthetic data, using real data in a secure sand-boxed environment, exporting the model to the outside, and then synthesizing data. This procedure complies with our healthcare partners' regulatory requirements. We develop a novel Wasserstein GAN and conduct a benchmark study on MIMIC data comparing it to 5 other approaches using novel metrics, based on nearest neighbor adversarial accuracy, for defining the resemblance and privacy of synthetic data generated from real data. We performed a comparison of 6 data generative methods on the MIMIC-III mortality problem: Gaussian Multivariate, Wasserstein GAN (WGAN), Parzen Windows, Additive Noise Model (ANM), Differential Privacy preserving data obfuscation, and (6) Copy the original data. Using our resemblance metric only the WGAN and Parzen windows show enough resemblance when compared to both the training and testing data. With our privacy metric the WGAN, Gaussian Multivariate, and Parzen Windows all excel. The Gaussian Multivarite method is ultimately ruled out as it does not retain a high enough utility level for education and research purposes. The Parzen Windows method is eliminated because it stores the original data and therefore sacrifices privacy in the model itself. Therefore the WGAN was the only effective method that maintained privacy and that allowed model export. This workflow can be used to address the vital need to create datasets for health education and research without undergoing deidentification which can be both costly and risky, and lose information.

5.4 Protecting fMRI data from unforeseen privacy attacks in a distributed machine learning environment

Michael Ellis, Naseeb Thapaliya, Lavanya Goluguri, Harini Booravalli Suresh and Shan Suthaharan

Abstract: The functional magnetic resonance imaging (fMRI) data [1] are highly sensitive to privacy attack, when they are distributed and scattered over multiple platforms to support data reuse and research reproducibility for the enhancement of machine learning based interdisciplinary research. The fMRI data, when sophisticated predictive models are developed using them, can easily disclose regions of interests (ROIs) information, from which one could deduce privacy information of an individual's thoughts or opinions. Therefore, it is important to develop meaningful machine learning models that lead to high predictive or classification accuracy, while protecting privacy.

In our recent research, we studied compressed sensing [2] and compressed learning [3] techniques with two-state Markov chain to characterize the transition behavior of fMRI signals. We then used these transition states to construct compressed sensing matrix and transformed the signals for compressed learning to build a privacy-preserving predictive model. This approach served as a feature selection mechanism. We have presented this work with our preliminary results and findings at the Stanford Compression workshop [4]. This model showed its strengths with Logistic Regression (LR) and Sparse Multinomial Logistic Regression (SMLR), while showing contradictory results with Naïve Bayes (NB) and Artificial Neural Network (ANN). However, we observed that our approach, as a feature selection mechanism, eliminates the distinguishable characteristics between the classes, which led to drawbacks for the techniques (NB and ANN) that rely heavily on statistical

nature of the data. We have now transformed this compressed sensing with two-state MC approach into a feature extraction approach and able to obtain very high accuracy for NB and ANN. We will present this enhanced approach along with our results and findings in our short-talk presentation.

CLOSING REMARKS:

Keith Webster
Dean, carnegie mellon university libraries

POSTER ABSTRACTS

1. Promoting open data by empowering stakeholders: what data should authors share?

Timothy Vines and Kristen Ratan

A growing number of funders, institutions, and publishers are developing policies that encourage or require the sharing of research data. However, authors rarely comply with these policies, often because there is confusion over exactly which datasets should be shared and where they should be placed. Similarly, publishers, funders, and other research stakeholders struggle to check policy compliance because they too are unsure what should have been shared. The overall result is that most research data remain with authors and will ultimately be lost to science.

We are developing DataSeer, a free, open source, web application. It uses AI to check articles for data-related keywords, lists the datasets associated with the article, then tells the authors how each dataset should be shared. DataSeer thus bridges the information gap between stakeholders' general data sharing policies and the actions required for a specific paper. Authors are clearer about their responsibilities, while funders, institutions, and journals are empowered to check whether authors have complied with their data sharing mandates. DataSeer concomitantly fosters the development of best sharing practice for each type of data, and facilitates broad uptake of that best practice across the research community.

2. Approaches to systematic transaction data reuse: machine learning support of information discovery

Jim Hahn

This poster will present a case study on an account-based recommender system at an academic research library to detail approaches in systematic transactional data (e.g. topic clusters found within book circulations) reuse with machine learning. Machine learning workflows were undertaken in WEKA and made use of an FP-growth algorithm. The pilot recommender system (RS) was derived from data mining topic clusters belonging to library items that were checked out together. With these topic metadata clusters a rule set for the recommender system was developed. The prototype recommender study began in October 2016 with seed data of 33,060 consequent subject association rules from initial machine learning processes. At the time of writing (2019) there are over 600,000 topic clusters collected from library transactions. These clusters form the basis for the prototype library account-based recommender incorporated into the library mobile app (https://minrvaproject.org/modules_recommendations.php). This poster will analyze antecedent and consequent subject clusters that were utilized for seeding the library account-based recommender. Finally, the poster will consider additional implications for other data reuse with machine learning approaches in research library settings; the foundations of which will serve to support information discovery for users, especially for content that may have been previously overlooked. Recent findings of student interviews indicate a desire for personalized account-based recommender systems to help support interdisciplinary scholarship. Graduate students in particular indicated that RS could help them see the way other research areas approach similar problems—this is a feature of academic libraries' RS that is under-appreciated—they are not simply "more like this" search engines—they also may be designed for novelty and support increased capability to browse the library collection, and ultimately support discovery.

3. Custom named entity recognition with spaCy - A fast and accessible library that integrates modern machine learning technology

Andrew Janco

This short paper will introduce spaCy, a free and open-source library for text analysis. Developed by Matthew Hannibal and Ines Montari in Berlin, spaCy offers a suite of tools for applied natural language processing (NLP) that are fast, practical and allow for quick experimentation and evaluation of language models. These tools make it possible for an individual scholar to quickly train models that can infer customized categories in named entity recognition tasks, match phrases, and visualize model performance. While comparable to the Natural Language Toolkit (NLTK), spaCy offers neural network models, integrated word vectors and dependency parsing. The paper will detail how the author trained a model to identify Spanish-language name components to automatically identify distinct persons in a collection of human rights documents from Guatemala. The use of custom entities, such as father's and mother's last name (apellido), helps to prevent misidentification and confusion where two persons have the same family names, but in different positions. Finally, the paper will introduce Prodigy, which is an annotation and active learning tool from the makers of spaCy. Prodigy allows a single researcher to quickly fine-tune a model for greater accuracy on a specific task or to train new categories and entities for recognition. This simple web-application can also be used to crowdsource annotations.

4. The RIALTO Project: A Stanford Research Intelligence System

Peter Mangiafico, Tom Cramer, Vivian Wong, Mike Giarlo, Justin Coyne, Justin Littman and Aaron Collier

Stanford University is a large research institution, with over $1.6 billion in sponsored research. Along with other sources of funding, this produces an enormous amount of research. The results of this work are published in journals or books, presented at conferences, taught in classes and workshops, and captured in data sets. In order to help Stanford work towards new opportunities and fulfill its mission, we seek to better understand and catalog this research output, capture it in preservable form, make it discoverable and understand how it is interconnected.

Stanford currently maintains separate systems for tracking researchers, grants, publications and projects, but it has no canonical system for combining this information and further tracking and managing its research output. The RIALTO project attempts to close that loop, helping provide a holistic picture of the University's activity and impact, while also eliminating waste through inefficient, duplicate data entry and opportunity costs stemming from lack of information.

RIALTO currently uses linked data principles and a linked data store to interconnect the various data elements being aggregated. It currently holds publications, researchers, affiliations, and sponsored projects, and provides a browse and search feature as well as specific reports. In the future, we plan to provide an API so that other applications can build upon the aggregated data.

Artificial intelligence, algorithmic approaches and machine learning principles provide an opportunity to further enhance the data, by allowing for additional interconnections and discoverability not currently possible with the metadata available from the source systems being aggregated. For example:

* by analyzing the full text of publications and its associated metadata, we can better understand and facilitate patterns of intra and inter-institutional collaborations among researchers across disciplines
* by looking at patterns of awarded grants by topic area, we can provide recommendations on funding opportunities to researchers

* by analyzing the full text of publications, we can better determine which publications resulted from awarded grants for better compliance reporting and ROI analysis

The RIALTO project is currently in beta testing as we continue to seek feedback and ideas for future work. The proposed short talk will be an overview of the project and its intended future directions, with the hope of spurring discussion and collaboration.

5. Modeling and analyzing cohort descriptions in research studies

Shruthi Chari, Miao Qi, Oshani Seneviratne, James P. McCusker, Kristin P. Bennett, Deborah L. McGuinnesss and Amar Das

We are developing an ontological workflow to operationalize cohort descriptions in clinical trials and observational case studies (referred here as research studies), with the ultimate goal of allowing physicians to better understand and target clinical recommendations. Treatment recommendations within Clinical Practice Guidelines (CPGs) are justified by findings from research studies, that are often based on highly selective populations. As a result, when physicians are faced with the treatment of complicated patients who don't align with guideline recommendations wholly, they face challenges in locating study evidence with results applicable to their clinical population. To address these challenges, we are developing a Study Cohort Ontology (SCO) – an ontology that captures the overall structure and patterns of cohort variables and control/intervention groups defined within the structured population descriptions (commonly referred to as Table1's or Cohort Summary Tables). We are also building a suite of ontologies (Diseases, Drugs, Medications, LabResults etc.) to accommodate Diabetes related terminology for Table1’s of research studies, cited in the Standards of Medical Care 2018 guidelines. We believe that SCO can be extended to publications targeting other diseases, and we plan to develop a workflow/set of tools to ensure longevity and extensibility of our ontology. Table1's vary in their descriptions of cohort variables and interventions, usage of descriptive statistical measures to cover aggregations, and in unit representations. Our modeling of the descriptive statistical measures (mean + standard deviation, median + interquartile range etc.) on cohort variables in RDF knowledge graphs, makes it possible to run queries and inferences across measure and unit representations, to ascertain the cohort similarity of a patient to a study population. For example, to determine the similarity of a young patient of color to a study whose population is largely of older white adults. Additionally, deep-dive visualizations (such as a star plot diagram) driven off query results, help visualize the fit of patient with a patient group, at a quick glance. In the future, we plan to quantify cohort similarity with scores learned through Semantically Targeted Analysis techniques, empowered both by analysis models to rank variables and semantics provided by our ontology. SCO also enables capabilities to compare reported populations with inclusion criteria, making it possible to assess if the actual study population is reflective of the intended population. Our cohort ontology supports provenance to maintain lineage of content, incorporates terms to support subject alignment evaluation, and enhances the transparency and understandability of study-supported guideline recommendations.

6. Bias amplification in AI systems

Kirsten Lloyd

As Artificial Intelligence (AI) technologies proliferate, concern has centered around the long-term dangers of job loss or threats of machines causing harm to humans. All of this concern, however, detracts from the more pertinent and already existing threats posed by AI today: its ability to amplify bias found in training datasets, and swiftly impact marginalized populations at scale. Government and public sector institutions have a responsibility to citizens to establish a dialogue with technology developers and release thoughtful policy around data standards to ensure diverse representation in datasets to prevent bias amplification and ensure that AI systems are built with inclusion in mind.

7. A common data model for repeatable and scalable cyber analytics - Empowering ai and novel defensive posture

Will Badart and Michael Dahlberg

In cyber security, analysts aggregate hundreds of intelligence streams, and hand-craft integration and processing programs on an ad hoc basis, reinventing the proverbial wheel for every experiment. What's more is that these manual processes may "only work on my machine," and may not preserve their data or results when completed (let alone archive and catalog them in accordance with any standard). The result: researchers can be bogged down by data acquisition and engineering problems and thusly struggle to advance the state of the art.

Booz Allen's approach to streamlining the integration of all those sources, with their proprietary formats and ontologies, is our Common Data Model (CDM), a universal schema for cyber security information. The CDM is the core of an ecosystem of parsers and analytic enrichments which automate the curation, flow, and archive of live, real-time cyber datasets.

Once a common data model is established, researchers can use intelligent data tagging and synthetic data generation to create large sets of labeled data for novel methodology development. By basing these methods off of a common data model, disparate groups can work to generate novel tags and simulated attacks simultaneously, better keeping pace with the scale of the cyber mission. With more consistent data types and better data sets, software engineers and data scientists can focus on reproducible data pipelines that can lead to improved automation at scale, more mature and novel algorithm development, and reproducible deployment for consolidated learning.

This presentation will explain the development process which led to the CDM, and use that discussion as a basis to explore its varied impacts. We will present a series of case studies as evidence of how the CDM promotes both cyber defense efficiency and the pace of new research. Finally, we will go on to show how we have embedded this philosophy of consistency and repeatability into our broader practice of operational architecture design and network defense deployment.

By unifying proprietary data sources, automating their curation, and building scalable, repeatable pipelines, Booz Allen is successfully coordinating the efforts of cyber domain experts and data scientists alike. The CDM lays out a unified language for these communities to share, fostering collaboration, enforcing consistency, and paving the way for the next generation of cyber research.

8. Entity name disambiguation in scholarly big data

Kunho Kim, Bruce A Weinberg and C. Lee Giles

Entity resolution (ER) is the problem of identifying or matching the same entities from a single collection or across multiple collections into unique clusters. Scholarly documents have these problems with many entities such as authors, datasets, affiliations, etc. We have investigated this matching problem in scholarly databases such as CiteSeerX, Web of Science, and PubMed in order to discover unique author name entities and clusters in publication records. These datasets have almost 300 million author mentions of which we believe about 20 million are unique. We present our scalable supervised pairwise classification method using gradient boosted trees (GBT) and a deep neural network (DNN). Pairwise classification is used to predict either each pair of records are the same authors or not. Then we discuss three different ER problems that utilize the classification method. We demonstrate our author name disambiguation (AND) on PubMed, which uses a record-to-record similarity calculated from the pairwise classifier to cluster publication record of each authors. The method successfully identifies 5.5 million unique authors along with a list of publication records for PubMed. In addition, we describe our author name search engine, PubMedSeer. This search engine indexes the disambiguated author names from the previous task and allows for advanced queries that use an external publication record. We then present our idea of using the pairwise classification method to measure record-to-cluster similarity and show that our algorithm can accelerate the response to such a query by a factor of 4 compared to a baseline naive approach. Finally, we discuss our method of measuring cluster-to-cluster similarity to match author entities across multiple databases. Our method enables us to accurately match advisors in ProQuest dissertation records to authors in PubMed publication records, matching approximately 800 thousand named entities in total. We believe these disambiguation methods can be applied to other entities besides author. Using our methods, users should be able discover and give proper attribution of research to authors, collaborators, funding agents, methodologies, and datasets.

9. Prediction of seismic wave arrivals using a convolutional neural network

Jorge Garcia and Lauren Waszek

Using energy from earthquakes that traverse Earth, seismology is able to observe the physical structure of the planet's interior. Large amounts of seismic data are needed to have a detailed image of the Earth's internal structure and its evolution; typical datasets are compromised of over 100,000 seismic records. With the exception of some basic processing methods, going through the datasets is done by hand using simple visualization software. Consistency of datasets is an issue, due to discrepancies between their compilation done by different researchers and the inconsistency in human decisions over time. This can result with repeating work in an attempt to achieve consistent results, which can even take over a year's worth of time. There are then other procedures that have to be done, such as analyzing the measurements from the data set and generating the seismic observables; the total time needed for a seismic project to acquire results can take several years. The goal of the project is to be able to develop tools with the ability to automate the processing and identification of seismic waves, thus reducing human error and time spent sorting data. We implement a Convolutional Neural Network and train a model capable of predicting the arrival time of a desired seismic wave within obtained seismograms to accelerate the task of data processing. We compare the results obtained from the model to those of an experienced seismologist.

10. Identification of causal genetic network for Alzheimer’s disease

Biwei Huang, Yandong Wen and Zhipeng Yang

Alzheimer’s Disease (AD) is a chronic neural degenerative disease characterized by progressive cognitive impairment. Commonly as it occurs, the underlying mechanism of AD is poorly understood. It is generally agreed that AD is a heterogeneous disease with a combination of genetic and environmental risks. Identifying new AD-related loci in human genome would provide valuable information for the underlying mechanism of the disease and potential new therapeutic targets for treatment. Single nucleotide polymorphisms (SNPs) refer to the variations of single nucleotide in specific positions in the genome. They are the most common genetic variations among people. Here we analyzed the SNP data based on the whole genome sequencing (WGS) from Alzheimer’s Disease Neuroimaging Initiative (ADNI). We identified 186 AD-associated SNPs in the correlation analysis (threshold set to 0.2). These SNPs fall in 46 genes, mostly protein-coding. In addition, we generated a causal network of 27 SNPs directly connected to AD. Our analysis on existing AD patients’ genetic data could potentially help researchers to better understand the underlying nature of this complicated disease and therefore help to better predict and prevent the progression of the disease.

11. On applications of bootstrap in continuous space reinforcement learning

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari and George Michailidis

In reinforcement learning problems for models with continuous state and action spaces, linear dynamics together with quadratic cost functions are central. For this setting, some randomized methods for data-driven strategies have been employed in recent literature to address the exploration-exploitation trade-off. However, little is known about resampling policies such as bootstrapping observed states and actions. In this work, we establish efficiency of bootstrap-based policies showing the square root scaling of regret with respect to time. We also obtain results on the accuracy of learning the model’s dynamics. Corroborative numerical analysis that illustrates the technical results is also provided.

12. Seismic fault mapping and facies prediction using parallel computing on big data

Patitapaban Palo and Aurobinda Routray

In several fields of research, a small dataset doesn't represent the complete and real scenario. The oil industry is one such. A small and real hydrocarbon dataset can't give the whole picture of a geographic structure which is essential pre-information for oil exploration. In such cases the use of large scale data becomes inevitable. These datasets refer to Big data that is too complex to handle. Big data comprises of four key concepts, popularly the 4V characteristics. These characteristics are Volume, Variety, Velocity and Veracity. However, the data acquired from hydrocarbon fields can be big only in terms of volume and variety. Cores, well-logs, seismic, trends, production and geology are among the varieties of big data that can be acquired. Quantitative data is not only acquired from various sources but also varying the time of acquisition and space resolution. Huge volumes of such varieties of data are to be integrated expertly in order to get precious information about thin layers, fault mapping and facies prediction.

Improving resolution and enhancing detectability of thin beds are crucial for getting the correct seismic data. Whereas fault mapping and facies prediction are essential in getting information about reservoirs. A fault is typically imaged as the discontinuity in seismic data. Whereas, facies characterizes of a rock unit that reflects its origin and permits its differentiation from other rock units. Fault mapping and facies prediction are important tools for correctly estimating the reservoir capacity. Existing methodology for these tools considers well log data only. Current practice also involves repeated validation and visual inspection by experts. However, integrating big data analysis tools into the techniques of fault mapping and facies prediction is an attempt to evolve an automated process. Retrieving such vital information from big data will require the employment of scalable data management and algorithms. Parallel algorithms are the most effective for this.

Support Vector Machine (SVM), being one of the effective two-class classifiers, is implemented for fault mapping and facies prediction. But SVM has the problem of scalability. The parallel SVM (PSVM) plays a key role in reducing computing cost. Instead of using the huge quantity of data as the whole set of training vectors, data is split into subsets. Next SVM is applied separately to the subsets and then individual results are combined and filtered to get the final result. Modelling of complex channels and multiple fault intersections are carried out from geophysical datasets.

13. Developing methodologies for data reuse from free-text radiology reports

Glen Ferguson, Michoel Snow, Boudewijn Aasman, G Anthony Reina and Parsa Mirhaji

As in any large healthcare system, the Montefiore Medical Center produces thousands of radiological exams and associated reports each day. These exams depend on expensive infrastructure and require radiologists with years of experience to interpret them, but are only reviewed a few times, and very rarely in the context of other patients. This treasure trove of untapped data is ripe for reuse for research, clinical decision support, operational efficiency, and care coordination. However, there is no established methodology for extracting medically relevant information from unstructured clinical texts, such as radiology or pathology reports. This is partly due to a wide range of potential tasks, from identifying rare phenomena among hundreds, if not thousands, of normal findings, to differentiating between multiple competing diagnoses.

More recent approaches to solve these problems center around translating free-text to dense embeddings suitable as inputs for deep learning. These methods (e.g., Word2Vec or Doc2Vec), have significant unknowns when applied to clinical data, such as the necessary size of the corpus, and the effectiveness of pre-trained (context-independent) embeddings. In addition, the type and parameters of the neural network, (e.g., recursive versus convolutional neural networks) that are most effective for a specific medical application are not well established and may be task specific. This uncertainty results in trial and error or blind reuse of models between applications. The wide range of potential use-cases and modeling practices warrants a systematic approach for mapping the task at hand to the requisite combined methodologies that would optimally solve it.

In this presentation we will discuss a generic framework for determination of a combined methodology to discriminate between patients with 'spinal cancer' and 'no cancer' using only the text of radiology reports. Our final model predicts the occurrence of spinal cancer with a sensitivity of 0.81 and specificity of 0.97, using document embeddings fit on approximately 20,000 training samples of a highly imbalanced dataset. We will outline the full technology stack, developed in collaboration with Intel, that has allowed us to develop an end-to-end DL/ML platform for healthcare data science. We will then describe the generalization of the platform to multi-class classification for automating the detection of spinal cord compression in radiological images. Understanding the process for developing a combined embedding/deep learning methodology can significantly increase data reuse from clinical free-text reports.

14. Using machine learning to automate metadata generation in academic libraries

Harish Maringanti

High quality metadata is the bedrock of Digital Library systems, as it helps in users discovering the unique content in various collections housed in digital libraries. Creating metadata is a time-intensive manual process and is done by experts who are trained in metadata schemas, and taxonomies; but the process is a bottleneck in adding content to digital libraries. J. Willard Marriott Library at University of Utah received funding to explore the feasibility of using image analysis to generate descriptive metadata for archival images. Our project aims to extract meaningful metadata from archival images, by applying advanced image analysis techniques. Fundamental to any digital library system is the quality metadata, without which, discoverability of collections is not possible. If algorithms can assist human experts in creating quality metadata, it will help libraries and archives, in accelerating their "collection processing" workflows.

15. Automatically extracting and generating traffic scenario primitives from naturalistic driving data

Wenshuo Wang and Ding Zhao

A desired automated vehicle that can handle complicated driving scenarios and appropriately interact with other road users requires the knowledge of the driving environment. This capability is usually attained by analyzing massive amounts of naturalistic driving data. An important paradigm that allows automated vehicles to learn from human drivers and gain insights is revealing and understanding the principal compositions of traffic, termed as traffic scenario primitives. However, the exploding data growth presents a great challenge in extracting traffic scenario primitives from high-dimensional time-series traffic data. Therefore, automatically extracting primitives is becoming one of the cost-efficient ways to help autonomous vehicles gain insights into the complex traffic scenarios. In addition, the extracted traffic scenario primitives from raw data should 1) be appropriate for automated driving applications and also 2) be easily used to generate new traffic scenarios. However, existing literature does not provide a method to automatically learn and reuse these traffic scenario primitives from large-scale traffic data. Our contribution has three-manifolds. The first one is that we proposed a new framework to extract and reuse new traffic scenarios from a handful of limited traffic data. The second one is that we introduce a Bayesian nonparametric learning method, i.e., a sticky hierarchical Dirichlet process with a hidden Markov model to automatically extract primitives from multidimensional traffic data without knowledge a priori. The third one is that we develop a disentangled generative adversarial network (dGAN) to generate new traffic scenarios by reusing these extracted traffic scenarios primitives. The developed dGAN allows us to generate the desired traffic scenarios by adjusting disentangled factors. The developed framework is then validated using naturalistic driving data over 3 years. Experiment results show that our proposed framework provides a semantic way to reuse limited on-hand traffic data for autonomous driving applications such as analyzing complex traffic scenarios.

16. 3d conceptual design using deep learning

Lan Zou, Zhangsihao Yang and Haoliang Jiang

This article proposes a data-driven methodology to achieve a fast design support, in order to generate or develop novel designs covering multiple object categories. This methodology implements two state-of-the-art Variational Autoencoder deal- ing with 3D model data. Our methodology constructs a self-defined loss function. The loss function, containing the outputs of certain layers in the autoencoder, ob- tains combination of different latent features from different 3D model categories.
Additionally, this article provide detail explanation to utilize the Princeton Model- Net40 database, a comprehensive clean collection of 3D CAD models for objects. After convert the original 3D mesh file to voxel and point cloud data type, we enable to feed our autoencoder with data of the same size of dimension. The novelty of this work is to leverage the power of deep learning methods as an ef- ficient latent feature extractor to explore unknown designing areas. Through this project, we expect the output can show a clear and smooth interpretation of model from different categories to develop a fast design support to generate novel shapes. This final report will explore 1) the theoretical ideas, 2) the progresses to imple- ment Variantional Autoencoder to attain implicit features from input shapes, 3) the results of output shapes during training in selected domains of both 3D voxel data and 3D point cloud data, and 4) our conclusion and future work to achieve the more outstanding goal.

17. Machine-learning-guided optimization of battery electrolytes

Adarsh Dave and Venkat Viswanathan

Electrolytes play a key role in battery performance, affecting power capabilities and cycle life. Electrolytes can often be complex mixtures of solvents, salts, and trace additives, which have been optimized through extensive trial-and-error testing over many years. Electrolyte data is too sparse and scattered to apply a rigorous optimization method to this problem.

To solve this issue, we have built an automated test apparatus with the ability to both generate a complex mixed electrolyte and to characterize two key properties of an electrolyte, conductivity and electrochemical stability. Data is generated, stored in clean forms, and shared over a web-server with API integrations for experimenters.

Given an objective function of these two properties, a machine-learning model (Dragonfly) integrates with this API to guide experimentation to deliver an optimal electrolyte. Dragonfly leverages a state-of-the-art electrolyte simulation to build background knowledge on the electrolyte design space, enabling the model to more effectively guide experiments. Machine-learning thus enables the apparatus to converge on an optimal solution to a high-dimensional problem in the fewest iterations.

This talk/poster will illustrate the hardware, software, and data design of the apparatus, the effectiveness of machine-learning methods in guiding this optimization, and key results.

18. Proposal of a wide-area supply chain restoration simulation method for the Nankai Trough massive earthquake tsunami: Applying Deep Q network

Yoshiki Ogawa, Shaofeng Yang, Yuki Akiyama, Yoshihide Sekimoto and Ryosuke Shibasaki

In this study, we developed a deep reinforcement learning simulation method of reconstruction for economic damage to intercompany transactions spread through supply chain (SC) for a predicted earthquake tsunami at the Nankai Trough near Japan in the future . We also attempted to optimize an enterprise decision making scenario in the event of a disaster by clarifying the impact of the industrial location on the supply chain using 1.6 million data sets from the SC networks of different companies. We first estimate the number of employees and the building damage of each company in the case of a Nankai Trough earthquake tsunami for the initial conditions using various geospatial data, such as building information and mass mobile phone GPS data. In the reconstruction simulation, a multi-agent system (MAS) was constructed with each company as the agent, and a deep reinforcement learning method with Deep Q Network (DQN) was employed. Through the multi-agent simulation, an optimal behavior model was constructed with the objective function that each company can maximize its market value. In addition, the disaster-affected companies will learn each action by modeling their recovery actions for their facilities, the occurrence of alternative transactions when the original transactions are lost, and optimization of the recovery actions. In the model, we adopted an evolutionary algorithm that references other agents for better action by sharing the Q functions of the other agents; thus, cooperative action could be restored more quickly. In the simulation, at the beginning of the learning, each agent took action at random by selecting the actions and taking time to recover. As the agent learned, it became possible for each company to achieve earliest recovery by optimizing the combination of actions to the shortest possible set of actions for reconstruction. In other words, it suggests that the speed of recovery is better when strategic actions are taken. Further, from the estimated reconstruction results by industry type of the affected enterprises, the construction industry and transportation and communication industry were observed to have better speed of recovery than the other industries such as wholesale and financial industry.

19. Genetic algorithm for evolving and strategizing the game of Tetris

Abhinav Nagpal and Goldie Gabrani

Since the inception of evolutionary algorithms, the capabilities of genetic algorithms are showcased by games. This paper proposes an algorithm that uses an evolutionary approach i.e. genetic algorithm (GA) followed by frequent pattern growth algorithm (FP) (hence, named GAFP) to evolve a player's gameplay in the game of Tetris. Tetris is a block-matching game that was designed and developed by a Russian game developer, Alexey Pajitnov. It consists of seven different geometric shaped blocks called tetrominoes. The objective of the game is to place the randomly falling tetrominoes onto the game board in such a way that it does not cross the top margin of the board. Each row filled completely with tetrominoes is cleared, thus decreasing the height of the stack by one. As the game proceeds, the game levels and the speed by which the pieces fall also increases, thus making it difficult to place the tetrominoes at the right position. The proposed approach uses a novel set of parameters to decide which move needs to be taken by the Tetris bot for each falling Tetromino. The parameters represent the various genes present in the chromosome. The Tetris bot developed would be able to play the game with the goal of clearing as many rows as possible. A fitness function has been used to select better performing individuals and continuously evolve the bot with time. Each individual is being allowed to play the game once. Once the entire population of individuals has played, the population undergoes crossover and mutation. In this way, the parameters are evolved to get a better bot. Due to time constraints, an upper limit has been set to the fitness function. The most evolved bot as per the fitness is allowed to simulate 150 rounds of Tetris during which all its actions are stored in a database. Further, Frequent Pattern Growth algorithm, an association rule mining technique, is used to extract knowledge from the given stored actions. The extracted knowledge is used for mining association rules and identifying the correct and useful strategies used by the evolved bot to play the game.