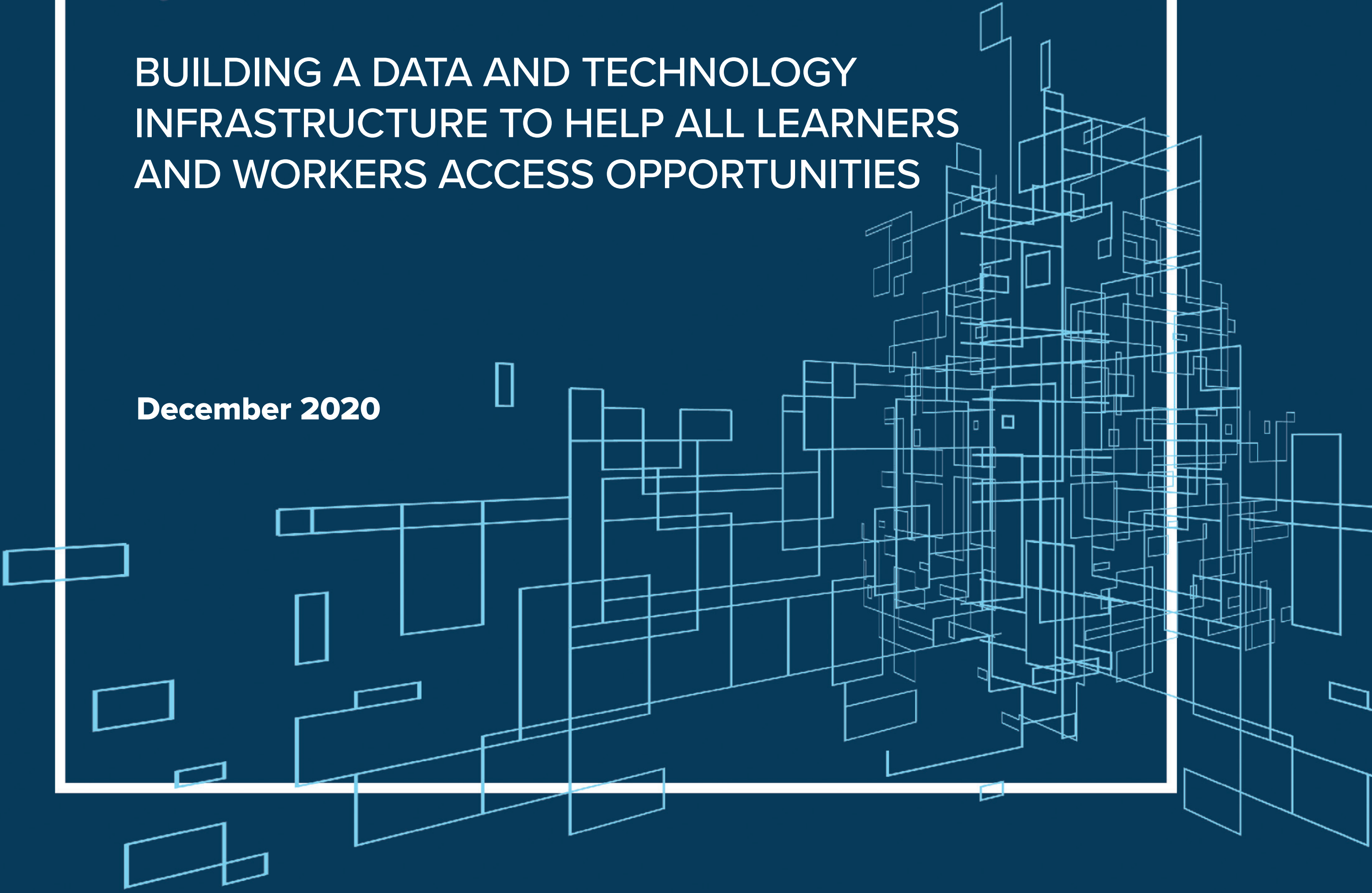
T3 INNOVATION NETWORK

SKILL & COMPETENCY DATA TRANSLATION AND ANALYSIS

BUILDING A DATA AND TECHNOLOGY INFRASTRUCTURE TO HELP ALL LEARNERS AND WORKERS ACCESS OPPORTUNITIES



December 2020

U.S. CHAMBER OF COMMERCE FOUNDATION

TABLE OF CONTENTS

INTRODUCTION

Skill and Competency Data Can Reduce the Burden on Individuals to Obtain Opportunities

Skills and competencies are the currency of the labor market. They describe all the things an individual can do which can be observed, measured, or otherwise assessed. Skills and competencies can include any range of knowledge and abilities, from "graphic design" to "bending and threading conduits," and more.[1] By 2030, employers are expected to require increasingly cognitive, technological, and interpersonal skills and competencies, as opposed to manual labor and basic skills.[2] This shift presents a challenge for our current and future workforce. In a study of 500 U.S. human resource (HR) professionals, 74% agreed there is already a skills gap, a difference between the skills that employers need and the skills that job seekers possess.[3] COVID-19 has worsened this disconnect by accelerating existing workforce trends of automation and increasing the demographic divide between who benefits from technology and who is left out.[4] Empowering learners and workers for current and future periods of career transition through clearer signaling of supply and demand in the labor market may help create a more inclusive workforce. The use of data to facilitate the exchange of skills as "currency" in the workforce will be vital for discovering and connecting these individuals to reemployment and upskilling opportunities.[5]

Analyses of data about skills and competencies can be used to match open positions to job seekers, enable more robust career planning, tailor educational curricula to job market demand, and identify opportunities for non-traditional or underserved learners. This requires structured data, data sharing and governance, algorithms, applications, research, advocacy, and communities of practice. Together these can improve communications between labor market systems to enhance and scale services for individuals across the touchpoints they use to access information about their education and career, e.g. by supporting the creation of career navigation software for use by career coaches and jobseekers or by improving the filters jobseekers can use to search on job boards. If authors of skill and competency frameworks begin to structure and provide open data licenses for use by the talent ecosystem (everyone involved in the training, education, and workforce sectors), this will enable applications and services using this data to reduce the burden on individuals to match themselves with opportunities and find support.

[1] Though the terms skill and competency are often used interchangeably, many stakeholders tend to view skills as a subset of competencies. We will refer to skills and competencies in this report to encompass all statements of what a person can do.

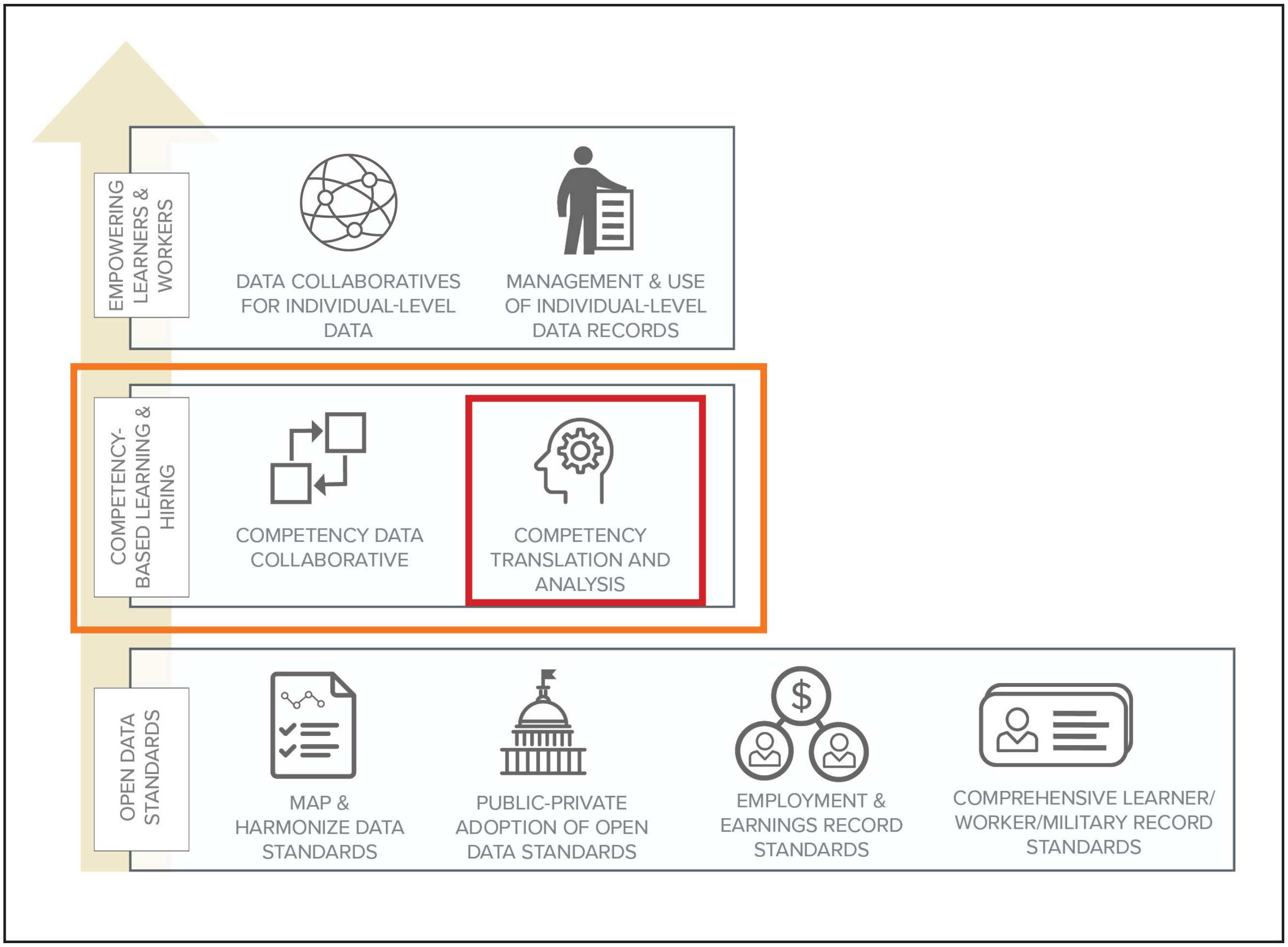
[2] Sandra Durth and Kristina Störk, "Thriving after COVID-19: What skills do employees need?," Mckinsey Accelerate, 2020, https://www.mckinsey.com/business-functions/mckinsey-accelerate/our-insights/accelerate-blog/thriving-after-covid-19-what-skills-do-employees-need.

[3] U.S. Chamber of Commerce Foundation, "Hiring in the Modern Talent Marketplace," U.S. Chamber of Commerce Foundation, 2020, https://www.uschamberfoundation.org/reports/hiring-modern-talent-marketplace.

[4] Phaedra Boinodiris Frsa and Rebecca Kemper, "Reprioritising Workforce Development," RSA, 2020, https://www.thersa.org/comment/2020/04/reprioritising-workforce-development.

[5] Molly Scott, Lauren Eyster, Christian Collins, Semhar Gebrekristos, and Yipeng Su, "Better Connecting Students to Job: A Guide for Policymakers to Encourage and Support Integrating Competencies in Postsecondary Education and Training," Urban Institute, 2020, https://www.urban.org/sites/default/files/publication/102281/better-connecting-students-to-jobs_1.pdf and Sapana Agrawal, Aaron De Smet, Sébastien Lacroix, and Angelika Reich, "To emerge stronger from the COVID-19 crisis, companies should start reskilling their workforces now," Mckinsey & Company, 2020, https://www.mckinsey.com/business-functions/organization/our-insights/to-emerge-stronger-from-the-covid-19-crisis-companies-should-start-reskilling-their-workforces-now.

The T3 Innovation Network Improves Skill and Competency Data


EMPOWERING LEARNERS & WORKERS
DATA COLLABORATIVES FOR INDIVIDUAL-LEVEL DATA
MANAGEMENT & USE OF INDIVIDUAL-LEVEL DATA RECORDS
COMPETENCY-BASED LEARNING & HIRING
COMPETENCY DATA COLLABORATIVE
COMPETENCY TRANSLATION AND ANALYSIS
OPEN DATA STANDARDS
MAP & HARMONIZE DATA STANDARDS
PUBLIC-PRIVATE ADOPTION OF OPEN DATA STANDARDS
$
EMPLOYMENT & EARNINGS RECORD STANDARDS
COMPREHENSIVE LEARNER/ WORKER/MILITARY RECORD STANDARDS

The T3 Innovation Network (T3 Network) strives to provide guidance, coordination, and an open data and technology infrastructure to reduce silos between employment, education,[6] government, and supporting systems across the talent marketplace. The T3 Network launched in 2018 and is managed by the U.S. Chamber of Commerce Foundation with support from the Annie E. Casey Foundation, Bill & Melinda Gates Foundation, Google, Lumina Foundation, Microsoft, and Walmart. The network is comprised of more than 500 organizations representing business, government, education, and technology stakeholders working together to build an open, decentralized, public-private infrastructure for education and workforce data. This infrastructure will ultimately support applications that will make it easier to search for and find information on the open web and improve data exchange across the public and private sectors.

[6] We will use "education" or "education providers" to refer to the stakeholder group including education, training, and credentialing providers.

In Phase 2 (January 2019 - December 2020), the T3 Network launched the Competency Translation and Analysis project (informally referred to as Pilot Project 6) along with eight other projects, all of which were outputs of work groups held in 2018. This project uses and builds off of the findings from two open-source tools that the T3 Network is developing and pilot testing to make competency and skills data more accessible and machine-actionable. The Competency Framework Extraction Module (CFEM) is an open-source tool that can convert skill and competency data into a variety of common, digital, machine-actionable formats used by learning, training, and credentialing software.[7] Additionally, the Open Competency Framework Collaborative (OCF Collab) is an open member trust network focused on making competency and skill frameworks readily available to people and machines through agreed-upon search, retrieval, and retention rules. The Competency Translation and Analysis project will use competency and skills data from CFEM and the OCF Collab to help the community analyze, compare, and translate competencies within and across industries using artificial intelligence and machine learning. This paper summarizes takeaways from those projects and serves as a reference for using technology in conjunction with skills and competency data to improve efficiency and outcomes. These recommendations build off the “T3 Network Phase 1 Report,”[8] the “Work Group 3 Report,”[9] the “Landscape Analysis: Improving Signaling of In-Demand Skills,”[10] and collaboration with the other T3 Network and community projects and work groups (i.e. the Competency Advisory Group (CAG)).

USING SKILL AND COMPETENCY DATA TODAY

Supply and Demand are Mismatched

Communication skills are both in demand and commonly listed on resumes and profiles. According to LinkedIn, “92% of talent professionals reported that soft skills are equally or more important to hire for than hard skills” and they are the most common reason for a hire not working out.[11] Yet “communication” and other soft skills are compensated at half the rate of hard skills,[12] creating a disincentive to invest in developing or improving them. Skills and competencies with high importance to the talent marketplace (e.g., communication skills) are often undervalued and may suffer from the difficulty of understanding what they mean in different contexts such as industry or seniority of a position. This is part of a broader signaling problem between employers, education providers, individuals, and workforce intermediaries.

[7] T3 Innovation Network, “LER Hub Pilots Community T3 Network Tools,” https://lerhub.org/s/curators/specs-0/s6wb9CdEi3qoxJQ48-0.

[8] US Chamber of Commerce Foundation, “T3 Innovation Network Phase 1 Report: Developing an Open, Public-Private Data Infrastructure for the Talent Marketplace,” US Chamber of Commerce Foundation, 2018, https://www.uschamberfoundation.org/reports/t3-network-phase-1-report.

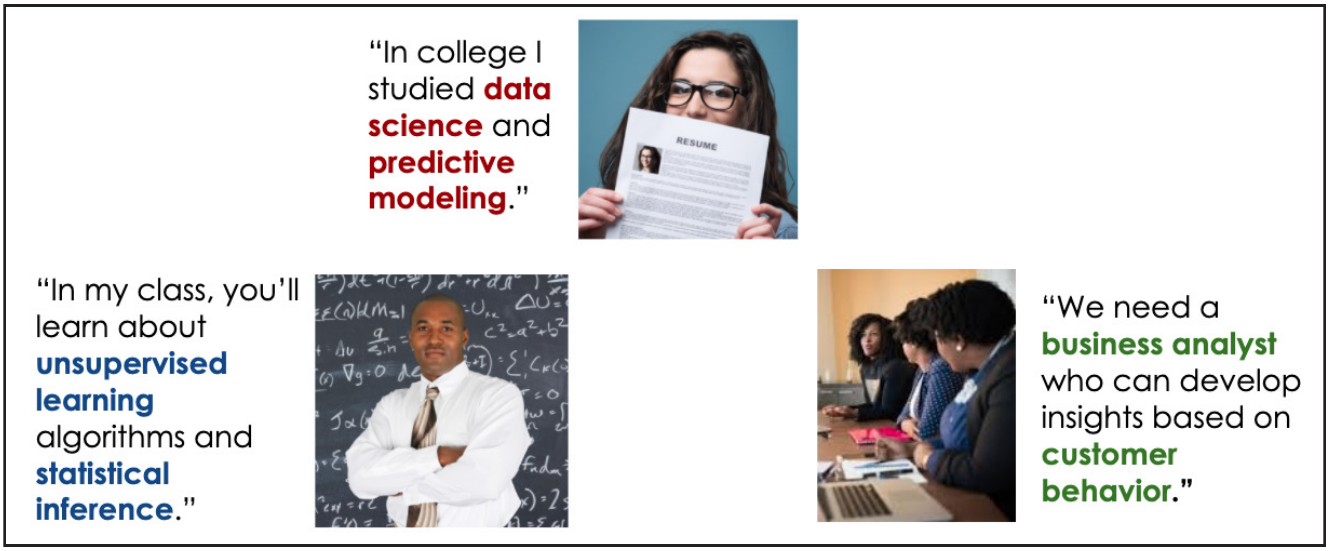
[9] US Chamber of Commerce Foundation, “T3 Innovation Network Work Group 3 Report: Developing and Analyzing Competencies,” US Chamber of Commerce Foundation, 2018, https://www.uschamberfoundation.org/sites/default/files/Work%20Group%203%20Final%20Report_July%202018.pdf.

[10] U.S. Chamber of Commerce Foundation, “Landscape Analysis White Paper,” US Chamber of Commerce Foundation, 2020, https://www.uschamberfoundation.org/sites/default/files/2020_WorkforceLandscapeAnalysis_FINAL_5.3.20.pdf.

[11] Samantha McLaren, “Candidates’ Soft Skills are Notoriously Hard to Assess, But Following These 6 Steps Will Help,” Linkedin, 2019, https://business.linkedin.com/talent-solutions/blog/recruiting-strategy/2019/soft-skills-are-hard-to-assess-but-these-6-steps-can-help.

[12] Julie Avrane-Chopard and Jaime Potter, “Are hard and soft skills rewarded equally?,” McKinsey & Company, 2019, https://www.mckinsey.com/business-functions/organization/our-insights/the-organization-blog/are-hard-and-soft-skills-rewarded-equally.

Each group uses different vocabularies, styles of documentation, and formats to indicate skills and competencies, which is highlighted in the example below.



"In college I studied **data science** and **predictive modeling**."

"In my class, you'll learn about **unsupervised learning** algorithms and **statistical inference**."

"We need a **business analyst** who can develop insights based on **customer behavior**."

The U.S. Chamber of Commerce Foundation's report "Landscape Analysis: Improving Signaling of In-Demand Skills," describes how the signaling problem affects the skills gap.

> This means that learners may be acquiring the competencies and credentials they believe matter and that education, training, and credentialing providers may be designing programs to teach them, but all parties are responding to signals that have been diluted or distorted. And, to compound the challenge, it may have taken so long for the signal to travel down the chain that, by the time it reaches the learner, employer needs have changed. As a result, supply and demand are mismatched: learners have competencies they cannot sell in the labor market, and employers are unable to fill open positions.[13]

Skill and competency data originates in four primary types: job descriptions/postings, resumes and curricula vitae (CVs), competency frameworks and classification systems, and course syllabi and curriculum descriptions. Much of the data needed to combat the skills gap already exists in federal and state data repositories, professional associations, or private technology platforms.[14] The challenge is that the data is largely proprietary and/or non-machine-actionable, making it incredibly expensive and difficult to create innovative, data-driven competency-based technologies and practices. Fortunately, the foundational technology and standards required to convert the data and make it publicly available at scale already exists. To move from the current state of skill and competency data silos to an open and actionable data ecosystem will require a critical mass of data providers and users to make the shift.

[13] US Chamber of Commerce Foundation," Landscape Analysis White Paper," US Chamber of Commerce Foundation, 2020, https://www.uschamberfoundation.org/sites/default/files/2020_WorkforceLandscapeAnalysis_FINAL_5.3.20.pdf.

[14] Examples of these skills/competencies sources include Federal data repositories (e.g. O*NET), national repositories (e.g., Credential Registry, IMS Global CASE, and D2L ASN), State data repositories, professional associations (e.g. NICE Cybersecurity Competencies), technical platforms for creating competency registries (e.g. Open Salt), and private technology platforms (e.g. LinkedIn, Workday, Indeed, etc.).

Current Ecosystem Gaps

While private sector actors have dedicated resources to machine learning to solve these problems and policymakers have pledged their support, significant roadblocks remain. Even as data becomes more structured,[15] we anticipate the following gaps will still need to be addressed. Gaps in data cause the classic "garbage in, garbage out" phenomenon. The best choice for improving the effectiveness of data-based solutions is to improve the data before investing in improving the algorithm. A community which can develop around the creation and use of this data will strengthen the quantity and quality of data produced, its rate of use, and its impact. In parallel, we see needs around the following areas of algorithmic development in order to improve the accuracy and applicability of the science.

DATA GAPS

- **Outcomes data**
  - » **Description:** Job performance data, employment status, wages, job satisfaction, grades, and job placement which can be tied to artifacts such as learning and employment records, resumes, and job postings.
  - » **Problem:** This data exists in employment and earnings records, unemployment insurance (UI) wage data, and other public and private-sector datasets of individual-level data. Highly sensitive, it must be properly aggregated and responsibly shared in order to be used. Most of this data is proprietary and/or unstructured.
  - » **Potential:** Outcomes data, when connected to competencies, would allow for validation, comparison, leveling, and calibration of competency-related algorithms and business processes. It is useful for answering questions like "what skills do I need to obtain a family-sustaining wage in my town?" or "which programs certify the most students in computer numerical control and what skills and competencies do they teach?" A large, representative, anonymized, and regularly-updated dataset would dramatically accelerate progress for both researchers and algorithmic learning, and would provide an avenue for determining the fit between job requirements and applicants.

- **Labeled skill and competency data**
  - » **Description:** "Tagged" unstructured or structured data. Tagging can be binary, (identifying whether or not a given word or phrase constitutes a competency) or multiclass (defining what skill or competency a given body of text refers to).
  - » **Problem:** Tagging requires expertise that may involve manual work or review by subject matter experts. Tagged competency data (e.g., a full resume with the skills and competencies labeled) is rarely openly available.
  - » **Potential:** This data is invaluable for training supervised learning algorithms so they can learn to pick out which text from documents contains skills and competencies and identify them.

[15] Structure provides a kind of predictable scaffold for the underlying data, giving the computer a set of easy-to-follow rules for how the data is to be read and worked with. Structured data is comprised of clearly defined data types whose pattern makes them easily machine-searchable.

- **Machine-actionable competency frameworks**
  - » **Description:** These frameworks facilitate easier communication between employers, employees, and learners within a given domain or across domains by providing lists of skills and/or competencies.
  - » **Problem:** The T3 Network's Competency Data Collaborative work on the Competency Framework Extraction Module (CFEM) finds that most industry-specific frameworks are available only in PDF format. More research is needed to improve the automatic conversion from PDF to a machine-actionable format. For competency frameworks which are not in PDF or image formats, a solution that works for every document is theoretically possible but highly ambitious and must draw on a large body of training data that is not available.
  - » **Potential:** We recommend that the ecosystem start by focusing on using documents which are in Microsoft Word, excel, csv, text, or other non-PDF/image formats for automated conversion efforts. Conversion of each new style (the kinds of information and its order and location in the documents) that the CFEM encounters requires one-time analysis of each style to produce a new algorithm which can be used on all the subsequent documents it identifies as that style. The CFEM then outputs the data in a standardized format which is ready to be shared on the OCF Collab if the owner would like to make the information searchable from the network with usage restrictions of their choosing.

- **Translational, hierarchical, and level relationships**
  - » **Definition:** Data about the relationships between different skills, including hierarchical (e.g., skills subsumed under higher-level competencies), level (e.g., high degree of skill versus a lower one), and translational (e.g., competencies related in complex ways but are not perfectly analogous).
  - » **Problem:** While progress has been made in extraction of competency phrases from large corpuses of text, such extraction often lacks the structure expected by users of traditionally constructed competency frameworks. Moreover, competency translation across frameworks can cause information loss if the structure of the source framework is not preserved or mapped to the target framework.
  - » **Potential:** Labeled relationship data and improved algorithms for inferring these types of relationships would assist in algorithmic development to overcome these challenges.

- **Equity data**
  - » **Definition:** Demographic and outcomes data that can be associated with skill and competency information.
  - » **Problem:** It has long been a stated goal of competency-based hiring to open up new opportunities to disadvantaged communities and nontraditional learners, but demographic information may not exist or may be disconnected from skill and competency information and outcomes.
  - » **Potential:** In order to validate the hypothesis and to ensure that employers and educators are not inadvertently creating disparate impacts as they change their business practices, it would be valuable to have anonymized longitudinal worker/learner data which includes key demographic elements. Though progress can be made by removing potentially biased language from job postings, demographic information is critical to exploring disparate impact.

- **Canonical data sets for model training and benchmarking.**
    - **Definition:** In many other machine learning contexts, one or more data sets has emerged as the industry-standard way to validate and compare proposed algorithmic improvements. The MNIST dataset, for example, is commonly used to test different machine learning techniques.[16]
    - **Problem:** Within the skill and competency data ecosystem, the appropriate data sets may already exist, but have not been designated by academic researchers.
    - **Potential:** Designating and propagating a canonical dataset is useful for sharpening problem statements and lowering the barrier to entry for new algorithm development.

COMMUNITY GAPS

- **Cross-sector communities of practice.** Communities of practice exist among education providers and their vendors, and similarly among employers and their human resource information (HRIS)/ application tracking system (ATS) vendors. There are also valuable skill- and competency-related efforts underway among academic researchers, labor market information (LMI) organizations, and government statistical agencies. However, it is rare to find venues for these separate communities to interact with each other, or even to interact among themselves (both domestically and internationally). For progress to continue, it is important that a practitioner-focused, sustainable, cross-sector network emerges beyond this work. These communities should promote communications and education around using skill and competency data, including how to digitize, share, and source information around common use cases.
- **Research hub.** This may include a web forum, publication, or conference that can serve to communicate ideas and highlight emerging problems. Current efforts along these lines include the LER Resource Hub and the T3 Network Resource Hub (expected in 2021), both of which could continue to grow, develop, and promote best practices for competency data with the benefit of additional participation and investment.
- **Unsubstantiated value propositions.** The theory of the value is robust, but it would shine with more demonstrations and testimonials from each stakeholder group, particularly enabled by proof-of-concept pilots.
- **Ways to find skill and competency data.** The Open Competency Framework Collaborative (OCF Collab) is an open member trust network to make skill and competency frameworks readily available to humans and machines (through agreed upon search, retrieval and retention rules). Using federated search and format interchange, it allows users to search across the data of all the members in the Collaborative from any member registry's interface and receive the data in the native format of the interface they are using.

ALGORITHMIC GAPS

- **Competency extraction.** The context-sensitive extraction of words or phrases representing competencies from a body of raw text, like a resume or job posting, and matching it against a finite list or hierarchy of competencies. Automating this task will lower the barrier to entry into competency-based learning and hiring for institutions with legacy data. This work is largely mature but needs to be brought to market. One key area of improvement for this work lies in identifying soft skills as opposed to hard skills given the more heterogeneous wording used to describe soft skills
- **Inference of hierarchical relationships.** Tweet-length phrases describing skills and competencies,

[16] Deep AI, "MNIST Dataset," September 22, 2019, https://deepai.org/dataset/mnist.

in isolation, are insufficient to support many of the business processes which educators and employers seek to modernize. Some notion of hierarchy or relatedness within a given set of competency statements is essential for matching, discovery, and classification. This includes being able to easily shift up across levels of abstraction. If given a specific competency description, inference algorithms should be able to identify the general class of competencies to which it belongs. It also includes distinguishing between levels of competencies; for example, an entry level job might require basic spreadsheet skills, while analyst positions might require a more advanced understanding of the same concepts.

- **Inference of translational relationships.** High fidelity translation across multiple competency frameworks is a necessary requirement flowing from the T3 Network as adoption of a single standard framework is undesirable across industries, stakeholder groups, and in an open marketplace that competes on talent. Translation should preserve hierarchical relationships to the extent that it is possible.

HOW TO GET THE MOST OUT OF SKILLS AND COMPETENCIES

True Interoperability Means Translation Between Format and Meaning

Today there is a great deal of interest and growing momentum to promote a competency-based talent marketplace, optimizing transactions in the labor market while promoting a more equitable society. Education providers are increasingly focused on competency development and attainment and are promoting a wide variety of alternative pathways and credentialing opportunities that are designed to prepare all learners for current careers as well as the future of work. Employers are also improving their hiring processes and ability to hire based on competencies rather than proxies such as degrees.[17]

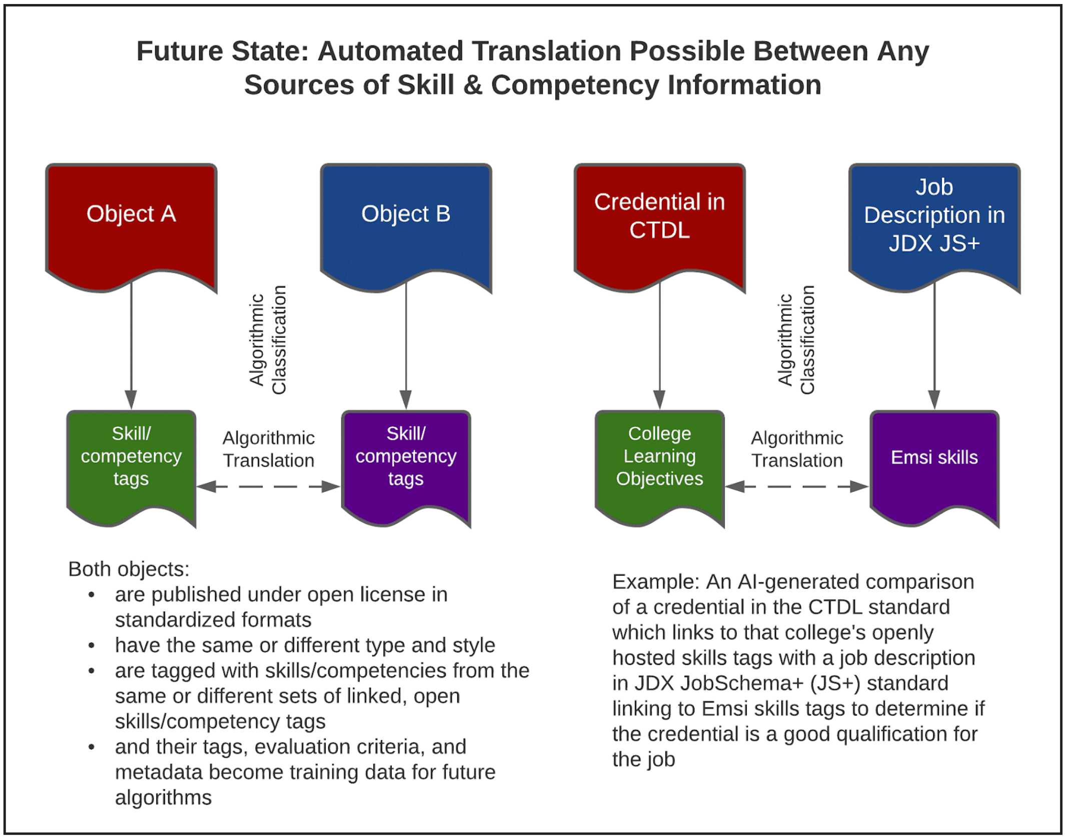
Competency-based learning and hiring optimizes the granularity of skills and competencies to help unbundle education courses, job positions, upskilling, and reskilling opportunities. For decades many organizations, both public and private, have wrestled with the idea of promoting a “common language” for skills and occupations. Efforts like the National Skills Standards Board (NSSB), a public-private coalition tasked by law to build a national system of skills and credentials, and the Secretary of Labor’s Commission on Acquired Needed Skills (SCANS) worked to develop a standardized skills taxonomy in 1994. The NSSB morphed into a voluntary partnership and SCANS was deactivated within a few years[18] as both efforts struggled to secure buy-in and support from the community. In order to succeed, this type of initiative needed to engage in a complex and time-consuming consensus-building process for defining competencies across the public and private sectors. In addition, rapid changes in the economy and changing job requirements make a singular competency taxonomy difficult to obtain, update, and maintain.

[17] Mark Lobosco, “4 Trends Changing the Way You Hire and Retain Talent in 2020,” LinkedIn, 2020, https://business.linkedin.com/talent-solutions/blog/trends-and-research/2020/global-talent-trends-2020.

[18] Frank Leibold, “Failing to close the global skills gap was a critical national failure: here’s why it failed,” Linkedin, 2019, https://www.linkedin.com/pulse/why-cant-american-higher-education-help-nation-close-its-leibold/.

Employers in the United States participate in an unstructured labor market in which employers have wide latitude in terms of their ability to set their hiring requirements—namely the skills, credentials, and experience needed to be considered qualified for a position. What exists today is not a "common language," but a diverse and highly complex landscape with multiple "languages" or styles that have been developed in silos. Beyond employers, industry and professional associations, education providers, technology vendors, social media platforms, and real-time labor market information (LMI) providers have developed their own competency languages based on client data or information aggregated from the Web. The federal government also maintains competency taxonomies that are used for a wide variety of planning and public policy purposes, including the U.S. Department of Labor's O*NET and the U.S. Department of Commerce's National Institute of Standards and Technology (NIST) cybersecurity skills framework. Industry and professional organizations, as well as certification organizations and state governments, also have developed skill and competency frameworks. The labor market is now awash with competency data, but much of it is not accessible, usable, or interpretable. True interoperability is not just format—it's meaning.

Humans are quite capable of identifying competencies and making informed guesses about the relationships between them, particularly within domains in which they are experts. However, the primary motivation for the T3 Innovation Network's focus on machine-actionable competency data is the potential of machine learning and artificial intelligence to allow a competency-based talent ecosystem to operate across diverse domains, stakeholders, and education and career pathways at scale, beyond the current labor capacities of advisors and experts. The below diagram illustrates matching between different sources of information from Object A and Object B in a way that is scalable and sustainable because the bulk of the work is automated, and the process produces training data which is openly licensed and linked to context.



Future State: Automated Translation Possible Between Any Sources of Skill & Competency Information

Both objects:

- are published under open license in standardized formats
- have the same or different type and style
- are tagged with skills/competencies from the same or different sets of linked, open skills/competency tags
- and their tags, evaluation criteria, and metadata become training data for future algorithms

Example: An AI-generated comparison of a credential in the CTDL standard which links to that college's openly hosted skills tags with a job description in JDX JobSchema+ (JS+) standard linking to Emsi skills tags to determine if the credential is a good qualification for the job

Definitions

Object

Any instance of document containing skill or competency data (i.e., a resume) from a source (i.e., a workforce agency).

Object Type:

The category object (i.e., a resume, competency framework, course, microcredential, and badge) which include both assertions and skill/competency definitions.

Style:

The manner of presenting the information (i.e., word choice, sentence structure, order of information, and more). There are infinite styles. The colors indicate that each object and set of tags may be using a different style and thus different words.

Data Standard:

Data standards are the rules by which data are organized, described, formatted, transmitted, and made available for different uses.

Skill/competency tags:

A set of only skills and/or competencies which can be extracted from an object or referenced from a competency framework.

Evaluation criteria:

Description of how a skill/competency item is assesses (i.e., a rubric).

Essential Elements for an Open Competency Data Network

We have an opportunity to build the first-ever open and distributed competency data infrastructure for the talent marketplace. This data infrastructure can create a living, linked network of skill and competency frameworks that can be rendered accessible and usable to key stakeholders: employers seeking to hire and upskill, educators seeking to align learning to careers, and learners and workers seeking to connect their skills to available opportunities.

The recommendations below aim to realize four visions of a near-term reality:

- Education providers and employers can easily access **data about applicants' prior learning and experience** in a structured, easily searchable way that allows applicants' information to be easily compared and analyzed.
- Education providers and employers can use AI to easily extract data from their curricula or training, enabling the rapid, efficient, and accurate propagation of **structured skill and competency data**.
- Education providers and employers can integrate data across internal systems and documents, ensuring **a common understanding of skills and competencies** across course catalogs, syllabi, hiring requirements, and instructional materials.
- Employers are empowered to use structured competency data as part of their hiring, training, and upskilling processes, enabling them to **determine gaps and strengths** in order to drive better hiring, talent sourcing, and training decisions.

Essential elements for an open skill and competency data network are as follows:

- **Public-private consensus on using open data standards** (including data models, formats and terms of access and use) and supporting technologies that render competencies frameworks accessible and usable on the web in a variety of open data formats. This is being addressed by the T3 Network CFEM and OCF Collab.
- **A distributed network of registries** or "nodes" where open, linked skills and competencies can be searched and discovered. This is being addressed by the T3 Network OCF Collab.
- **AI and machine learning technologies** that can be trained on machine-actionable competency data across multiple ontologies and frameworks that results in improved translation services. Additional research and work is needed in this area.
- **Applications that can access and draw upon open competency frameworks** for publishing learning outcomes, hiring requirements, and learning achievements. Additional research and work is needed in this area.
- **A dynamic feedback loop** to inform owners of skill and competency ontologies and frameworks of their use and how they can be edited, modified, or otherwise improved. Additional research and work are needed in this area.

The end result is a distributed network of the naturally-occurring diverse "languages" or vocabularies that can be accessed and translated for a wide variety of end-use applications (e.g., developing skill-based job postings for employers, competency-based learning outcome statements for educators, and education and career navigation tools for learners). This outcome would make it unnecessary to continue pressing for an infeasible consensus on a "common language," and would enable stakeholders to focus on proliferating languages that address a wide variety of value propositions. This makes all the underlying data accessible and usable to the wider talent marketplace without requiring behavior changes in the vocabularies used by the market.

DATA MATURITY MODEL

Organizations looking to participate more fully in the open competency ecosystem can benefit from an understanding of the data formats and how they affect data maturity. Much of the skill and competency data in the marketplace is digitized. However, the majority of this data is unstructured. This means it is readable and understandable by humans, but not by machines. Examples of unstructured data include course catalogs, LinkedIn biographies, or skill descriptions on resumes. Unstructured data is often in "natural language," or everyday writing. In order for this data to be used, it needs to be put in a format that computers can work with. This conversion can occur either by manual data entry or by algorithms that apply structure to unstructured text. Once competency data is structured, modern AI algorithms can be put to use in interpreting the data.

The model advances from the top left to the bottom right. While all stakeholders should aim to advance their data maturity, not all data is either appropriate or a priority to advance, based on the use case undertaken by the organization(s). Identifying where each dataset, or even element, fits on the model is a good way to take stock. Making skill and competency data publicly available (row 4) in any state of usefulness can still benefit the whole ecosystem. This model draws from both Tim

Skill and Competency Data Maturity Model

		Usefulness			
"This skill and competency information is..."		**A. Unstructured** Non-machine-readable data. This data may be digital and human-readable, but a machine would have great difficulty processing it.	**B. Structured or Semi-Structured** Data in rows and columns. A machine can read it if you tell it where to look for the information. CSV, XML, and JSON formats are semi-structured and relational databases are structured.	**C. Standardized** Data is not only structured, but formatted in a predefined standard so machines can easily locate information and change the data's format if needed.	**D. Contextualized** Standardized, structured data with permanent unique resource identifiers (URIs), links to descriptions of the original work or educational context (such as a rubric), and relationships to other skills and competencies.
Openness	**1. Department**	1A Human readable with department-specific access	1B In a custom format with department-specific access	1C Interoperable with department-specific access	1D Contextualized with department-specific access
	2. Organization	2A Human readable and internal only	2B In a custom format and internal only	2C Interoperable and internal only	2D Contextualized and internal only
	3. Micro Ecosystem	3A Human readable and shared in a collaboration	3B In a custom format and shared in a collaboration	3C Interoperable and shared in a collaboration	3D Contextualized and shared in a collaboration
	4. Macro Ecosystem / Public	4A Human readable and open	4B In a custom format and open	4C Interoperable and open	4D Contextualized and open

Level	Example
1A Human-readable with department-specific access	Images of competency information, like a scanned textbook page in JPEG, which has not been shared outside the department
1B In a custom format with department-specific access	Employee records with skill and competency data stored in JSON
1C Interoperable with department-specific access	Learning objectives in CASE format which have not yet completed a review process to be approved
1D Contextualized with department-specific access	A Learning & Development team's training programs are not integrated with the HR team's employee records
2A Human readable and internal only	Paper or Word doc copies of industry standards which are not shared outside the organization
2B In a custom format and internal only	A table in a relational database (i.e. SQL) containing skills
2C Interoperable and internal only	Career technical course information shared over a campus area network in the Common Educational Data Standards (CEDS) format
2D Contextualized and internal only	Standardized military competency framework with a rubric which links to other frameworks in a progression, considered too sensitive to release
3A Human readable and shared in a collaboration	Job requirements in PDF which are shared between a collective of employers and education providers
3B In a custom format and shared in a collaboration	Three training providers share and compare program data from separate excel sheets
3C Interoperable and shared in a collaboration	Jobs data in schema.org JobPosting format shared to a data collaborative and aggregrated for labor market info
3D Contextualized and shared in a collaboration	Learners grant a community-based organization access to their records which contain skills, competencies, and assertions in the Open Badge 2.0 standard
4A Human readable and open	Course catalogs in PDFs which are publicly accessible
4B In a custom format and open	An excel sheet of the U.S. Dept. of Labor Competency Clearinghouse models
4C Interoperable and open	Credentials published in the Credential Registry
4D Contextualized and open	This is the vision of Web 3.0. A badge hosted publicly by an individual which contains evidence of learning and skills that link to other skills

Berner Lee's five-star linked data model[19] and the OCF Collab's Five Star Scale,[20] with emphasis on adding context and relationships to the data, encouraging the open licensing of data at any point in the model, and providing a level on the model for all maturity levels that exist in the market. This maturity model should be used in concert with more comprehensive data maturity models that inform the advancement of an organization's data and processes across self-reinforcing facets of data security, governance, capacity, culture, and collection. Data maturity should also be pursued in conjunction with user experience.

A feature of column D, linked data, is when "every identifier is a URI, using standard lists (see vocabulary) of identifiers where possible, and where datasets include links to reference datasets of the same objects. A key aim is to make data integration automatic, even for large datasets." This data is highly valuable to algorithms. It allows competencies to not only have metadata, but also to be linked to other data points for rich context. Context is important because the knowledge, skills, and abilities (KSAs) required to fulfill any particular job can be described at various levels of abstraction, and the tasks of the jobs themselves can be described with different levels of granularity. Two jobs with similar tasks may require wildly different competencies and the same competency may be necessary in two dramatically different jobs. The language used to refer to the same underlying competencies may vary significantly. Conversely, similar language can be used to refer to diverse underlying competencies.

What You Can Do

All stakeholders should see the use case categories in Appendix A for a range of ways skill and competency translation and analysis can benefit them and the talent marketplace as a whole. An integrated perspective is the most powerful approach to increasing the visibility of competency-based learning and hiring. In the following stakeholder-specific recommendations, we suggest that data gaps are best filled by stakeholders with a vested interest in ecosystem development; community gaps by foundations and funders; and algorithmic gaps by academic researchers and private sector practitioners.

FOR RESEARCHERS

- Probe the possibilities for useful metadata attached to skills and competencies.
    - » Investigate and identify how to incorporate **skill levels** into competency standards' metadata. More than knowing what competencies applicants or employees have, employers and employees alike can benefit from structured, verifiable data about different levels of competencies.
    - » Conduct research into **the relationship between programs, courses, individual competencies, and labor market outcomes**. Though the competency data ecosystem is still developing, existing data products can be used to investigate the relationship between skills and outcomes. Individuals willing to consent to sharing their Learning and Employment Records (LERs) with researchers can provide self-reported and verifiable credentials as longitudinal data.[21]
    - » Investigate avenues for **attribution and licensing**. As the competency data ecosystem

[19] Timothy Holborn, "What is 5 Star Linked Data?," W3C Community, 2014, https://www.w3.org/community/webize/2014/01/17/what-is-5-star-linked-data/.

[20] Stuart Sutton, "5 Star Skill and Competency Data Scale", November 13, 2020, https://docs.google.com/document/d/12j1R02L7M1uPjOEp5O5j4421XqT7DODuT18xqzY0QuY/edit?usp=sharing.

[21] The T3 Innovation Network, "LER Resource Hub," 2020, https://lerhub.org/.

develops, it will become increasingly difficult to track attribution requirements and usage restrictions as interoperable competency data moves between platforms.
  - » Include **timestamps**, or some form of recognition, for acknowledging when the competency was developed, used, and/or verified.
  - » Conduct studies on the effects of **source** of the skills or competencies on the outcomes.

- Investigate risks in this space, including:
  - » **Limits of AI:** discerning when human subject matter expertise is needed.
  - » **Bias in AI:** processes and auditors to hold algorithms accountable.
  - » **Underutilization:** what are the demand drivers and blockers?
  - » **Ethics:** pre- and post-mortems to identify and plan for risks and failures.

FOR ADVOCATES, GOVERNMENT AGENCIES, AND PHILANTHROPY

- Build and fund systems that can **safeguard against bias** and adverse impact in competency data and the tools that leverage it. Without active stewardship, competency tools can lead to inequitable outcomes.
- Engage and encourage existing **advocacy groups** like the Open Skills Network and the T3 Network to encourage active development of competency data and tools. Continued growth of the competency data ecosystem is necessary for its survival.
- Fund the development of **open-source tools** that make use of competency data. The open-source ecosystem and the world of proprietary software should be synergistic. In order to build a world where competency data is fully utilized, both types of software are necessary.
- Advocate for the use of **self-sovereign identity principles**. Credentials must provide a way for the holder to prove conclusively that they are its legitimate holder without compromising the holder's privacy.
- Incorporate machine-actionable competency frameworks and open licensing into **funding requirements**, as demonstrated in the CARES Act. Translation algorithms however, need not be shared utilities.
- Recognize **quality assurance groups**. Continuous bias review is vital in order to insure algorithmic interpretation of competency data does not reproduce existing biases or introduce new ones. As competency data begins to be used more widely, philanthropists, advocates, and policymakers should encourage third-party assessments of the potential for bias in new systems. A prior era of innovation resulted in the now-universal use of the FICO score as a measure of creditworthiness. Despite their ostensible status as neutral algorithmic measures, however, credit scoring systems may underpredict the creditworthiness of members of protected classes simply as a function of the data used to create them. Independent, in-depth review of credit scoring systems has been vital to identifying and addressing these biases.[22] Credible groups who can evaluate algorithms could be cataloged for easy perusal and selection by decision makers.
- Investigate improvements in **outcomes data collection**. The T3 Network has supported data standards organizations (like HR Open Standards Consortium) to develop and promote standards for maintaining employment and earnings records data on workers. These standards are intended

[22] Dowse Rustin IV, Neil Grayson, and Kiersty DeGroote, "Pricing without discrimination: Alternative student loan pricing, income-share agreements, and the Equal Credit Opportunity Act," American Enterprise Institute, 2017, https://www.aei.org/research-products/report/pricing-without-discrimination-alternative-student-loan-pricing-income-share-agreements-and-the-equal-credit-opportunity-act/.

to reduce the time and cost for employers and their HR system partners in providing data to federal and state governments for a wide variety of public and private sector uses. For example, these standards could be used to enhance state Unemployment Insurance (UI) wage records and improve federal statistical data collections to provide better information on the labor market and return on investment for education and training programs. Outcomes data should include both positive and negative outcomes (attained job, did not attain job) to help validate algorithms and competency frameworks.

FOR EMPLOYERS, MILITARY, EDUCATION, AND LABOR MARKET INFORMATION PROVIDERS

- Increase the **quantity and quality of skill and competency** data. For example, draw on product descriptions and manuals to source a more granular level of skill than a statement such as "familiarity with excel."
- Convert your data into **machine-actionable formats** as detailed in the data maturity model. See the LER Resource Hub[23] for open source data sources and standards for storing your data and open tools like the Competency Extraction Module (CFEM).
- Create a **shared pool of raw and labeled data**. Identify what skills and competency data currently exist at your organization and convene internal stakeholders in order to align on goals and arrive at a vision of how this data can improve existing processes. Several existing collaborations exist to serve this purpose:
    - » The Open Competency Framework Collaborative (OCF Collab) is a member trust network exposing competency and skill frameworks to humans and machines through agreed upon search, retrieval and retention rules. Each member establishes clear usage and access rights for their frameworks.
    - » The Open Skills Network (OSN) is currently developing an open tool, the Open Skills Management Toolset, to enable authoring, editing, sourcing, and publishing competency frameworks.[24]
    - » Collaboratives for jobs data such as the National Labor Exchange (NLx) provide raw and semi-structured job postings,[25] and are worth contributing to both as a job posting channel and for AI training purposes.
- Ask your **technology vendors** to participate (i.e., Human Resource Information Systems (HRIS), Student Information Systems (SIS), Learning Management Systems (LMS)) in the above processes. Although many vendors see the value of a competency-based approach, they may need demand from customers before implementing it. As technically sophisticated players in the competency space, employer participants in the T3 Network are well-positioned to provide this validation.

FOR INDIVIDUALS (LEARNERS AND WORKERS)

- Ask your education provider if they will be **tagging their curriculum with skills** so that you can have more insight into the program offerings. Ask for digital credentials that can help make your job application more competitive.
- Ask how your institution's career services stays connected to the labor market and measures

[23] The T3 Innovation Network, "LER Resource Hub," 2020, https://lerhub.org/.

[24] The Open Skills Network, "Open Skills Network," 2020, https://www.openskillsnetwork.org/.

[25] The National Labor Exchange, "National Labor Exchange," 2020, https://usnlx.com/.

graduate outcomes. How does your institution ensure that employer **competency requirements are aligned with curriculum learning outcomes**?

- **Start collecting digital credentials** to demonstrate what you can do and explore "wallets" or online profiles for storing them. Request features that would improve your experience. Skills, competencies, and credentials which are "portable" and "stackable" allow you to easily transfer them from one institution to another and see how you can build on top of them instead of repeating content you've already done.
- Ask your employer how your skill and competency information is collected and stored. If such information is being captured, work with your employer to **pilot ways for skill and competency data to be used** to enhance performance evaluations, hiring and upskilling opportunities, and how you can contribute to, access, and request changes to your learning and employment record, such as by implementing a learning and employment record to improve career planning and management.

FOR VENDORS DEVELOPING ALGORITHMS

- Develop metrics for **algorithmic benchmarking** in order to be able to judge and define success for new algorithms in the competency space.
    - » For **competency tagging, definition, disambiguation, and inference of hierarchical relationships**, algorithms can be benchmarked against human taggers on both accuracy and time. Since machines will outperform on time, the primary benchmark to meet is the human average for classification accuracy. In other words, inter-rater reliability (IRR) between human and machine should be statistically indistinguishable from IRR between human taggers.
    - » Metrics for benchmarking **extraction algorithms** should measure whether these algorithms outperform a naive text search of unstructured text. They should yield a greater number of individually extracted competencies than a naive text search, without sacrificing accuracy.
    - » Metrics for judging the success of algorithms for **inference of hierarchical relationships** should allow these algorithms to be benchmarked against existing taxonomies. For instance, these algorithms should be able to take an unstructured corpus of job descriptions within a given occupational category and replicate in large part the structure of the existing O*NET content model.
- Seek **quality assurance frameworks** and groups who are offering to certify or provide assurance to the public about the implications of the algorithms.
- Algorithm developers can create or join an **open source community** around algorithm development, if they choose.
- Leverage emerging **open source software**, standards, and competency frameworks.

Collaborate

Skill and competency analysis and translation involves many stakeholders, even within a single organization. To find a sustainable solution, interorganizational collaboration will help reduce redundancy and accelerate progress. The T3 Network's Phase 3 will continue to bring all the stakeholder groups mentioned in this report together for participative learning and problem solving. The future T3 Network Resource Hub can serve as a learning center to help efforts form a minimum viable coalition that includes voices that will contribute to program and product sustainability, discover and collaborate on vendor-neutral, principle-aligned use cases and processes, discuss relevant regulation, and share the results of community efforts. Demonstrating ROI on investments in skill and competency analysis will influence more pilots and more available data.

CONCLUSION

A growing community of policymakers and researchers see skills and competencies as essential to modernizing the talent marketplace. Each of the three central players—employers, educators, and workers/learners—think and talk about this in their own way and have their own long-established artifacts and ways of doing business. Job postings, syllabi, and resumes are all loosely-structured documents with their own distinctive vocabularies, each with loose relationships to government reporting standards. Market incentives for adoption of competency-based approaches have been weak and inhibited by established interests and collective action problems. Even the technical meaning of the words "skill" and "competency" are the topic of significant debate.

In 2019, the T3 Innovation Network set the goal of supporting open source infrastructure for creating and using competency data and catalyzing the development of competency translation and analysis tools. Through the network's grant-funded work, as well as the ongoing efforts of its members, these goals are now in sight. However, competency-based learning and hiring is still the exception rather than the rule.

Skill and competency authors and publishers can render existing skill and competency data into machine-actionable formats online, making them accessible and translatable (using AI and machine learning) for workers, learners, and employers. This conversion can supplement, augment, or even replace their own proprietary or unstructured languages. We expect this to improve efficiency in all skill and competency-related tasks. This would build the first-ever open and distributed skill and competency data infrastructure to underlie systems across sectors. This infrastructure does not impose a single set of standards and approaches but grows a community of practice empowered with a rich set of data and algorithmic resources. This document identifies the key use cases and data, community, and algorithmic gaps and next steps highlighted by T3 Network members.

Skills and competencies form the foundation of a more adaptive and resilient labor market. Implementing the recommendations in this report will provide the labor market with the potential for an inclusive digital transformation. Skills and competencies as the new currency of the talent marketplace could result in reduced friction, improved efficiency, and more equitable outcomes across stakeholders.

ACKNOWLEDGEMENTS

This paper was authored by Danielle Saunders, Matt Lerner, Tom Plagge, and Matt Gee at Brighthive. Jillian Gilburne and Samantha Levy at BrightHive provided edits. Additional contributions to this paper were provided by Joshua Westfall of the U.S. Chamber of Commerce Foundation, Robert Sheets of the George Washington University, and Stuart Sutton of Sutton & Associates Consulting. We would also like to thank the T3 Innovation Network's Competency Advisory Group for their input.

APPENDIX A: USE CASE CATEGORIES

These example use cases show the kinds of work that can be benefitted by this report's recommendations. The requirements column references the algorithms from Appendix B, dependencies on other use cases, and other expected needs to accomplish the use case. Classification can be replaced by extraction or definition algorithms if needed. All use case categories could add on the use of leveled skill and competency data as the technology develops to increase the precision of the data. Any of the use case categories involving individual-level data could be enhanced by demographic data to monitor disparate impacts.

No.	Category	Requirements	Ecosystem and Stakeholder Benefits
1	Education & Training		
1a	Education providers can extract data about applicants' prior learning and experience (such as from prior learning assessments (PLAs)) in order to facilitate the transfer or enrollment process and secure credit for students.	Competency classification algorithm; labelled competency data.	Produces structured competency data for use in LERs, in matching to job postings.
1b	Education providers can use AI to tag courses, programs, pathways, learning resources, and credentials to skills to improve student and academic services.	Competency classification and translation algorithm; labelled competency data.	Produces structured competency data to align curricula to labor market demand and construct LERs; improves information for consumer choice, increasing market efficiency; integrated, governed data between internal systems and documents, (i.e. across course catalogs, syllabi, instructional materials); may reduce operating costs and/or improve educational offerings.
1c	Teachers and learners can find learning resources that are aligned to a certain set of competencies.	Relies on 1b. Competency translation; labelled translational data.	Improved efficiency of education marketplace; utility for alternative and home education providers who are sourcing their own curriculum.
1d	Education providers can work closely with assessment providers to compare skills needs with assessment products or create new assessment products without laborious manual comparisons.	Relies on 1b. Competency classification, translation; labelled translational data.	Improved efficiency for filling positions; closer matches between employee skills and position needs.

1e	Education providers can facilitate mentoring relationships based on skills desired and skills held by members of the learning community.	Requires community members skill/competency profiles which benefit from 1a. Competency translation.	Gallup found a strong relationship between caring mentors in college and outcomes after college, that students need help finding mentors, and that racial minorities access mentoring significantly less.[26]
1f	Advanced: Education providers can analyze student records, curricular data, and/or labor market data together to help determine student pathways, enhance career services, improve their understanding of the capabilities of incoming students and source institutions, and facilitate the matriculation of transfer students.	Relies on 1b. Competency translation; outcomes data; cross-sector communities of practice.	Improved responsiveness of training providers to labor market needs; faster retraining and upskilling for job seekers; tighter feedback loop between student outcomes and training.
2	Hiring & employee upskilling		
2a	Employers can extract skills/competencies from data such as profiles, resumes, competency frameworks, LERs, and portfolios to obtain structured data on applicants and current employees to create or add to employee records.	Competency classification; labelled competency data.	Reduced search time; better-qualified applicants; more equitable opportunities for similarly skilled job seekers.
2b	Employers are empowered to use structured skills data as a part of their hiring, training, and upskilling processes. Employers use current employees' or applicants' structured data to conduct or contract out an analysis of employee knowledge, skills, abilities, etc. to determine gaps and strengths to make hiring, talent sourcing, and training decisions.	2a must occur first. Competency classification and translation algorithm.	More informed hiring processes; improved efficiency and less waste in spending on hiring and professional development; more accurate performance evaluation. (Evaluating the impact of learning and tracking skills gaps are among the top priorities of Learning and Development professionals in 2020, according to LinkedIn[27]).
2c	Employers can create new or convert existing apprenticeship postings in a structured format and build on that format to submit data for registering their apprenticeships with DoL. Employers are enabled to easily share this data with state or private career pathways sites,	Competency classification; labelled competency data.	Improved connectivity between apprenticeship standards and available programs; faster validation of program options and more efficient inclusion in career pathways platforms.

[26] Leo Lambert, "The Importance of Helping Students Find Mentors in College," Gallup, 2018, https://news.gallup.com/opinion/gallup/245048/importance-helping-students-find-mentors-college.aspx.

[27] LinkedIn Learning, "4th Annual 2020 Workplace Learning Report," LinkedIn Learning, 2020, https://learning.linkedin.com/content/dam/me/learning/resources/pdfs/LinkedIn-Learning-2020-Workplace-Learning-Report.pdf?utm_campaign=LDC_Autoresponder_Content_Download_Dynamic_Global_EN_v3&utm_medium=email&utm_source=Eloqua&veh=LDC_Autoresponder_Content_Download_Dynamic_Global_EN_v3&eqid=CLNKD000362481942&elqTrack=true.

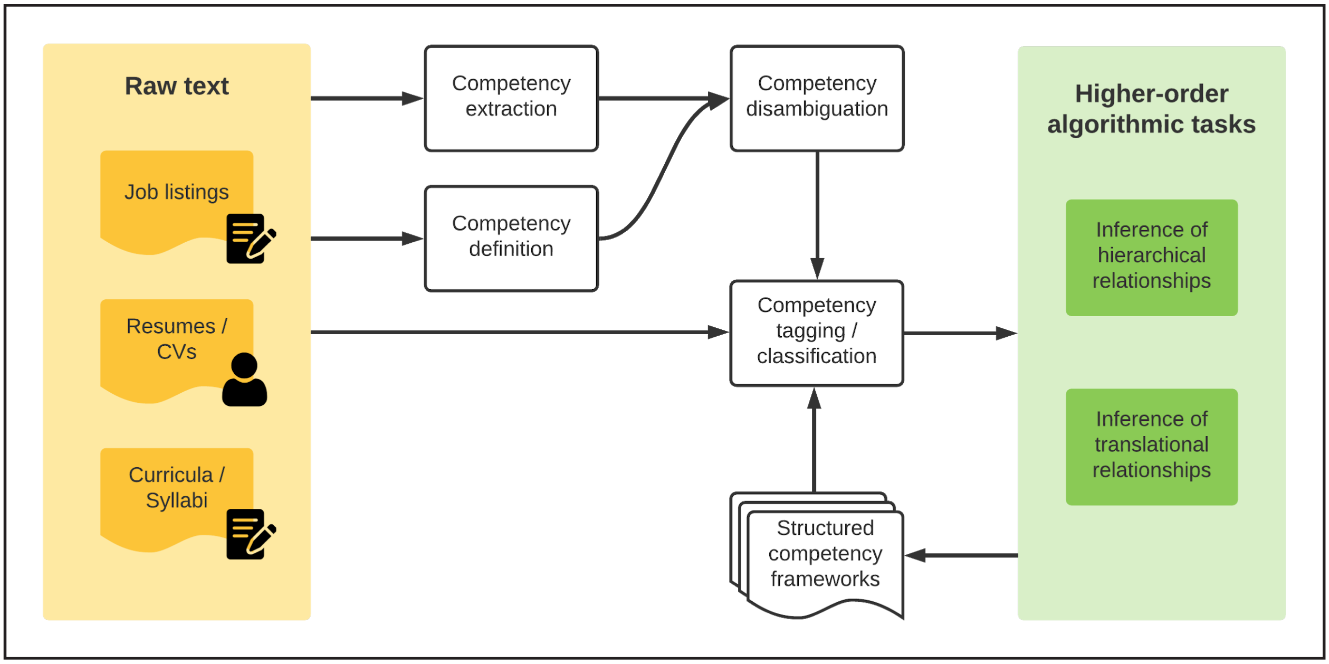
	government agencies, and talent sourcing providers.		
2d	Employers obtain education/training program/credential data that enables them to understand program ROI and make decisions about which training programs to invest in for their employees to upskill by comparing skills across programs.	Competency classification; outcomes data; labelled competency data.	Resources flow more efficiently to effective training programs; effective training programs propagate more widely.
2e	Employers are able to work closely with assessment providers to compare skills needs with assessment products or create new assessment products that enable them to rigorously assess potential employee skills and career readiness or current employee skill needs.	Competency classification and translation; labelled competency data.	More equitable hiring as a result of a focus on skills, not resumes; better matches between applicants and positions; reduced search time.
2f	Employers and employer collaboratives can provide information about hiring needs (i.e. tagging jobs to skills) to education providers to allow them to design programs that effectively serve students' desire to gainful employment.	Competency translation; outcomes data; cross-sector communities of practice.	More informed training program design; lower costs for program development and resulting lower program costs; improved pathways; lower training costs for employers.
3	Government and workforce agencies including labor market information		
3a	Labor market information providers, fueled by the influx of structured job posting data, are able to build out their analytics (i.e. tagging occupations with skills.)	Dependent on many of the above. Competency classification; labelled competency data.	More robust labor market information; a greater diversity of descriptive statistics to inform policymakers; supplement existing competency frameworks and processes for skill and competency discovery—this would improve the data that nearly all labor market stakeholders make use of.
3b	Policymakers and support organizations such as workforce boards have access to program, employment, & wage data, data they need to evaluate programs, advise learners and workers, and drive their decision-making.	Competency translation; outcomes data; cross-sector communities of practice.	Program evaluation becomes more rigorous and efficient; policymakers are better informed about the most effective training pathways.
4	Military		
4a	Credentialing pathways are created to address known translation gaps in military and civilian competency information.	Competency classification; labelled competency data.	Veterans transition more easily to civilian jobs; military recruitment for specialized positions sees improved efficiency.

4b	Military personnel can be matched more effectively to careers and possibly job postings based on their records.	Competency levelling algorithm; labelled competency and level data.	Reduced search time; improved outcomes for separating veterans.
4c	The labor-intensive mapping of military work tasks to military training competencies is automated.	Competency classification; labelled competency data and level data.	Reduced costs for military training program design; improved outcomes for training program completers.
5	Education & Career Pathways		
5a	Advanced: Navigation/pathway apps and services can ingest individual records, individual preferences, education provider data, military, and labor market information, and outcomes data to output suggested pathways to goals.	Competency classification, translation; outcomes data; cross-sector communities of practice.	Career recommendations better aligned with both labor market needs and individual preferences; improved worker and employer satisfaction with navigation services.

APPENDIX B: RELEVANT ALGORITHMIC TASKS

This is a quick reference for technical or semi-technical readers who need an overview of the ecosystem's most pressing technical needs. The following algorithmic tasks take data about skills and competencies in unstructured, semi-structured, structured, or standardized sources, process it, and use it to unearth complex relationships between different skills and competencies and the text that describes them. These algorithms do not pertain to undigitized formats. In the data flow, note that the raw text sources marked with the pen and paper are sources in which stakeholders define skills and competencies for the ecosystem or a subset of it and the source marked with the person icon (resumes and CVs) is an individual's assertions of skills and competencies that they have.

Data flow


Raw text
Job listings
Resumes / CVs
Curricula / Syllabi
Competency extraction
Competency definition
Competency disambiguation
Competency tagging / classification
Structured competency frameworks
Higher-order algorithmic tasks
Inference of hierarchical relationships
Inference of translational relationships

Competency extraction

Competency extraction is the extraction of words or phrases representing skills or competencies from a body of raw text, like a resume or job listing, and matching it either deterministically or against a finite list of predetermined skills. Understanding context is key to avoiding false positives and other errors.

> Example: "A successful candidate for this position will have at least five years of experience working in the field of intellectual property law. Candidates must be excited about the prospect of excavating actionable insights from troves of patent applications and the development of a new and varied skill set."

In this example, a skills extraction algorithm must recognize that the legal skill described is "intellectual property law" not the different but equally plausible "property law," both of which may exist in a predefined list of competencies. In addition, a good algorithm must identify that while "excavating" would fit as a skill in job listings for archaeologists, construction workers, or geologists, and that while "applications" and "development" might make sense as software engineer competencies, none of those terms represent skills or competencies in the above description. The simplest versions of this task involve direct or fuzzy matching of strings, for example, finding words related to the practice of law in a document. Word stemming software may be required in order to extract "baking" from both "bakery" and "bakes." This is the simplest algorithmic task we list. It can only extract what it knows to look for from a list of competencies - like giving someone a list of words and telling them to mark every time a word appears in a text.

Potential datasets:

- Raw text corpus (i.e. list of job listings, curricula, individual skills profiles, resumes)
- List of identified competencies

Candidate algorithms:

- Perfect string matching is the simplest way to perform this task. Raw text is searched for exact string matches against a list of predefined skills. This method will only capture exactly what it is looking for and may underperform other models.
- Fuzzy matching using measures like string distance to identify text that may be close enough to the text in an existing list to warrant the conclusion that the items are identical. Though there are many algorithms to accomplish this, it is likely that stemming raw text is a necessary preprocessing step before attempting fuzzy matching on competencies.

Competency definition

Similar to competency extraction, competency definition involves recognizing previously unidentified skills and competencies in individual raw text items or in a large corpus of texts. Effectively achieving this task with raw text in natural language requires a system with the capacity to make sense of sentence structure. In the case of data organized graphically, as with resumes, competency models, or syllabi, competency definition techniques may involve optical character recognition, or deterministic models of document organization.

> **Example: "Principal functions of this position involve developing preconstruction checklists, preparing bid documents and soliciting bids, and developing a final construction budget."**

A simple but effective system for competency definition might take the above text and extract "developing preconstruction checklists," "preparing bid documents," "soliciting bids," and "developing construction budget." More sophisticated competency definition systems might look at a large corpus of similar listings and, using a clustering algorithm, identify that these items form a distinct overarching competency (the "construction management" competency).

Potential datasets:

- Raw text corpus
- WordNet, Word2Vec/Doc2Vec embeddings, or some other pre-trained or deterministic NLP framework
- A trained dataset containing words or phrases pre-tagged as competencies/non-competencies

Candidate algorithms:

Algorithms used to extract novel competencies from raw text will generally need to go through two distinct steps. First, they will need to preprocess the text, tagging parts of speech and analyzing sentence structure (i.e. using NLTK or the Stanford POS tagger). Next, they will need to develop a model for what types of words or phrases constitute a skill. In order to achieve this, they will need a training set that identifies words—along with their context, part of speech, and place in the sentence structure—as either competencies or not. A trained model will take new text, analyze its sentence structure, and then use the trained model to classify words or phrases as competencies. Because of the complexity of this data, deep learning is perhaps best-suited to the task.

- ELMo is a "deep contextualized word representation." In contrast to more traditional word embeddings such as word2vec, ELMo word vectors take into account where a word falls in a sentence structure. In addition, ELMo is character- instead of word-based, which allows for greater flexibility with respect to misspellings, different spellings, and different forms of the same word. ELMo is trained with a Long Short Term Memory architecture.

- ULMFiT, like ELMo, uses an LSTM. Unlike ELMo, it uses more traditional word-based tokenization. It is a transfer learning method that makes a claim to generalizability that may not necessarily require training the model on a large, domain-specific dataset. This approach can be advantageous in the competency definitions space, since it potentially enables the extraction of words describing never-before-seen competencies in unusual contexts, even when the training set is irrelevant to those competencies and contexts.
- Either of the aforementioned representations will be used in conjunction with a classification algorithm like the ones described below in the tagging and classification section.

Competency disambiguation

Closely related to definition and classification, competency disambiguation is the task of recognizing when two competencies are identical or nearly identical. Examples of such pairs of groupings might be “car mechanic” and “automotive repair” or “landscaping” and “gardening.” Competency disambiguation is complicated by the necessity of avoiding false positives. For example, though “hydrology”/”hydrography” is a pair of closely related skills, inferring that they are identical would be too liberal an approach to disambiguation in many use cases.

Potential datasets:

- Raw text corpus. The development of a competency disambiguation algorithm can proceed directly from an untagged raw text corpus, from a tagged raw text corpus, or from a previously taxonomized graph of competencies.
- Pre-trained word embeddings resulting from an algorithm with some capacity to capture context, such as BERT

Candidate algorithms:

Competency disambiguation is fundamentally an entity resolution or record linkage problem.

- The popular Python module Dedupe determines if two entities are identical by determining the “distance” between sets of records—the similarity of these records across multiple different variables—and performing hierarchical clustering to identify groups of probable duplicates. In combination with limited manual training, this method allows users to deploy a generalizable deduplication framework but customize it to their own data. This method can, at its simplest, rely on string similarity (or “string distance”) as a base metric for identifying possible duplicates.
- The Data Tamer system also requires manual training and the clustering of entities along similar characteristics, but also leverages an additional step that can be particularly useful in competency disambiguation: the use of a graph structure to recognize instances of transitivity across duplicates (and a Naive Bayes classifier for flagging these duplicates). This could be useful in cases where conceptually identical competencies are described in different languages, but where they share critical features with some third, common competency.

- Some entity resolution frameworks are designed explicitly for use with knowledge bases or ontologies like those upon which some skills frameworks are built. One such method formulates the multi-type entity resolution task as a graph summarization problem in which "super nodes" represent clusters of vertices on the original graph.

Tagging and classification

Competency tagging and classification involves the deployment of algorithms to tag skills and competencies in raw text. Though similar to competency extraction and definition, tagging and classification is a more general "bucketing" problem. In general, tagging and classification systems take a body of raw text as input and produce a set of tags as output. The most common way this task is approached is as a supervised learning problem: a machine learning model is trained using a corpus of pre-tagged documents. The structure of the statistical relationship identified by this model is determined by the choice of algorithm used, and by the kind and extent of preprocessing applied to the text. The tags that result from applying this pre-trained model to new text may not be actually present in the new raw text, but result instead from a statistical relationship between a given competency and the occurrence of various other items in the text.

> **Example: "This course will teach students to identify and exploit vulnerabilities in common web server software. Students will learn about buffer overflow, distributed denial-of-service, and SQL injection, among other common attacks and exploits."**

A serviceable competency tagger might tag this posting with "cybersecurity" or "information security." Though the course description describes attacks and vulnerabilities, a tagging system should be sophisticated enough not to characterize it as a military science course and should be sufficiently well-trained not to mistake the single mention of SQL as evidence that the course described is a database design class.

Potential datasets:

- Machine-labeled text corpus (from the output of extraction and/or definition then disambiguation tasks) or human-labeled text corpus. In general, classification tasks benefit from manual labeling by humans, though this manual work does not necessarily require legions of workers classifying texts one by one. For example, LinkedIn users write their own profile summaries and also select their own skill tags; the alignment between these skill tags and profile texts could become a training set.

Candidate algorithms:

Classification and regression are the two broad subcategories of supervised machine learning. Competency classification is a special case of classification. There are dozens of algorithms applicable to this task. Some common ones include the following:

- Support vector machines are a type of classifier in which features of documents (for example, the occurrence of certain words in a job posting) are represented in a multidimensional vector space. The SVM distinguishes between categories by using training data to draw boundaries ("hyperplanes") between groups of documents in a fashion that maximizes the separation between them.

- Neural networks are algorithms organized into artificial “neurons” which attempt inference based off of the proposed functioning of the human brain. Long short-term memory (LSTM) algorithms in particular have shown promise as neural net-based approaches to text classification.
- Random forest classifiers are decision-tree based algorithms well-known for their versatility and high performance in data science competitions. These algorithms work by estimating a number of decision trees for the same classification problem, then selecting the modal predicted class for new data.

All of these models can use either simple vector representations of documents or more complex word embeddings, which represent words, sentences or documents as vectors of real numbers that also represent their semantic relationships to other words.

Inference of hierarchical relationships

The inference of hierarchical relationships from competency information is a challenging algorithmic task that typically requires a large amount of data. This task is complicated by the reality that lower-level competencies may fall hierarchically under multiple upper-level competencies. For example, the lower-level skill “board-breaking” likely falls under both the upper-level “Karate” and “Kung Fu” competencies.

Potential datasets:

- Raw text corpus. As with the development of a competency disambiguation algorithm, the development of a model to infer hierarchical relationships can proceed directly from an untagged raw text corpus, from a tagged raw text corpus, or from a previously taxonomized graph of competencies.
- Pre-trained word embeddings resulting from an algorithm with some capacity to capture context, such as BERT or DSSM.

Candidate algorithms:

- The dominant strategy for inferring these relationships is hierarchical clustering, which can be used in conjunction with word embeddings in order to learn conceptual hierarchies, such as those that flow from higher-order competencies to lower-order skills. Many variants of this strategy exist, and have shown promise for topic taxonomy construction. In general, these models are rule-based procedures that proceed either “agglomeratively” or “divisively”: they either start with one large cluster and divide it based on the distance between vectors or start with as many clusters as documents and combine them based on a similar criterion.
- A related set of models uses a probabilistic framework to achieve the same task; models that take this approach make implicit assumptions about the distributions of words and topics within a concept space and fit a statistical model in order to make predictions about new data.

Inference of translational relationships

The inference of translational or partial-knowledge relationships involves the identification of connections between competencies that are more than simply hierarchical. This could involve recognition that two competencies are related: for example, being able to throw a baseball and being able to throw a football are related even though there is relatively little overlap between professional football players and professional baseball players.

Systems that use competency data may also need to identify when skills are substitutable or partially substitutable, or when existing competencies are transferable to tasks that don't explicitly call for them. Algorithms that are able to infer this type of relationship might be able to identify that the competencies required to play lead guitar are partially substitutable for those required to play bass guitar, or that experience as a waiter may also serve as preparation to be a restaurant host.

Potential datasets:

- Raw text corpus. As with the previous two tasks, algorithms to infer translational relationships can rely on raw text alone (starting from scratch) or can build on models developed in less complex algorithmic tasks.

Candidate algorithms:

- Distance metrics include cosine similarity, Euclidean distance, and word mover's distance, which was developed explicitly for use with word2vec embeddings. When a sophisticated embedding is employed, distance metrics are intuitively interpretable as semantic similarity. However, the task of inferring translational relationships between skills likely requires greater sophistication than the simple calculation of similarity.
- Word2vec embeddings famously allow word vectors to be added and subtracted in ways that meaningfully reflect the underlying concepts (e.g. "king - man + woman = queen"). This has relevance for similar translational relationships with respect to skills and competencies (e.g. "python + statistics = data science" or "inDesign + CSS = graphic design").
- Tensor Product Decomposition Networks learn tensor product representations that approximate existing vector encodings. Early work in this area has focused on encouraging TPDNs to learn compositional representations such as the examples given for word2vec above.

Inference of proficiency level / evaluation criteria

Transcripts, test results, rubrics, and some job postings, performance reviews, hiring processes, individual profiles, and resumes pair skills and competencies with an associated "level" to provide more granularity about where on a scale performance falls or should fall. Examples include anything from "advanced" to "4" (out of 10), to letter grades, to lengthy descriptions. There may also be weights on each item or group of items. For example, a job description which lists a group of competencies on customer service as 30% of the job.

Potential datasets:

- Learning and Employment Records (LERs) will offer some students and job seekers the opportunity to present employers and educators with verifiable records of their expertise, many of which will contain skills and competencies with associated levels.
- Test results will enable some (but not all) companies to gauge applicants' suitability by assessing their performance on tasks similar to those they would do on the job. For example, Toptal allows aspiring software engineers to take a coding test; if they pass, top tech companies see their resumes. Video games designed as behavioral assessment for job screening can generate a high volume of quantitative performance data on hard and soft skills, with the potential to reduce the bias of self-reporting.[28]
- Rubrics with descriptions of what performance looks like for each level of a skill/competency or set of skills and competencies.

Candidate algorithms: The ability to compare skills and competencies with associated levels adds a layer of complexity to all of the above algorithmic tasks and deserves additional research.

[28] Barbara Marder and Frida Polli, "Want the perfect job in a fair and diverse workplace? There's an app for that," World Economic Forum, 2016, https://www.weforum.org/agenda/2016/09/perfect-job-unbiased-workplace-mercer-pymetrics-app/.

U.S. CHAMBER OF COMMERCE FOUNDATION