Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PERSPECTIVE

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2025, **4**, 1124

The Citizen Data Science program at Dow†

Kyle Andrews,[a] Steven Arturo,[b] Matt Benedict,*[b] Birgit Braun,[c] Brian Clark,[a] Simon Cook,[d] Jaime Curtis-Fisk,[f] Fabio D'Ottaviano,[e] Tim Licquia,[d] Peter Margl,[a] Jonathan Moore,[g] Lynette Naler,[c] Parth Singh,‡[c] Alix Schmidt,[d] Anatoliy Sokolov,[a] John Talbert[d] and James Wade[f]

Received 2nd January 2025
Accepted 19th March 2025

DOI: 10.1039/d5dd00002e

rsc.li/digitaldiscovery

We present the Citizen Data Science (CDS) program, a data literacy program aimed at a Research and Development (R&D)/Technical Service and Development (TS&D) population from a heterogeneous background of traditional disciplines such as chemistry, materials science, engineering and others. The CDS program aims to facilitate the culture change required for maximizing researcher productivity and wellbeing by equipping every researcher with the skills to best manage, analyze, and communicate their data, enabling them to thrive in R&D/TS&D organizations that themselves are going through profound structural transformation induced by the pressures of digitalization. The Dow CDS program is going through its fourth year of implementation and improvement; we share the program and our learnings in the hope that they may be useful to other researchers in the materials development and adjacent spaces.

Introduction

Massive advances in digital technology over the past decade are revolutionizing the way research towards new materials is done both in academia and industry.[1] This creates enormous opportunities, but also puts pressure on organizations to adapt the organizations and processes to the new paradigm. In the chemicals and materials industries, companies that have in some cases organically grown over more than a century now need to embrace an entirely new way of doing research based on the ability to generate, process and learn from increasingly large amounts of data.[2] A perhaps unexpectedly difficult part of this process of "digital transformation" is preparing their workforce to make optimal use of the new enabling technologies.[3] The Citizen Data Science (CDS) program was conceived as a response to the challenge of digitally enabling the workforce in a chemicals and materials research organization with a 125+ year history. It is intended to provide the right skills in the right amount to the right person while respecting and working within the constraints posed by a modern research, development and technical support (R&D/TS&D) environment. The nature of an agile, fast-paced, and highly diversified R&D portfolio requires that data is quickly and efficiently collected, worked up, and analysed close to the source by researchers intimately knowledgeable with the physical/chemical meaning of the data. Under such conditions, it is not practical to outsource significant parts of the data flow to partnering information technology (IT) organizations simply due to the inefficiencies of hand-off, turn-around, and the necessary presence of physical/chemical subject matter experts at all stages of the data life cycle (Fig. 1).[4] CDS is designed to enable researchers to cope with this situation, in which they must not only be involved in the design of the experiments and the scientific analysis of the results, but also need to be deeply involved in a host of other "digital" activities far from their original area of expertise.

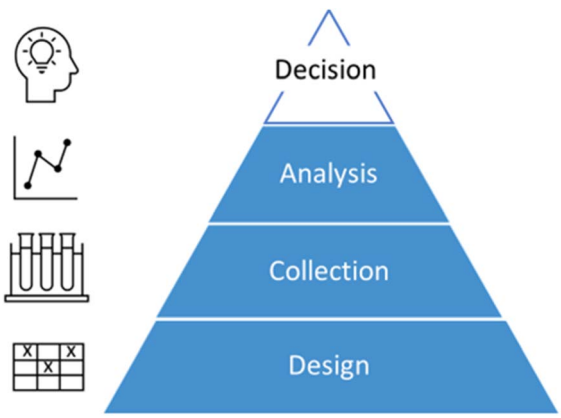

Fig. 1 The R&D data life cycle pyramid.

[a]*Dow Inc., 1776 Building, Midland, MI 48674, USA*
[b]*Dow Inc., 400 Arcola Road, Collegeville, PA 19426, USA. E-mail: MBenedict@dow.com*
[c]*Dow Inc., 240 Abner Jackson Parkway, Lake Jackson, TX 77566, USA*
[d]*Dow Inc., East End 715 E Main St, Midland, MI, 48667, USA*
[e]*Dow Inc., 220 Abner Jackson Parkway, Lake Jackson, TX 77566, USA*
[f]*Dow Inc., 1897 Building, Midland, MI 48674, USA*
[g]*Dow Inc., 1702 Building, Midland, MI 48674, USA*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d5dd00002e

‡ Current address: 680 W California Ave, Sunnyvale, CA 94086.

Digital literacy upskilling programs also have been defined in many other organizations, including companies across other industries like Merck[5] and Avery Dennison[6] and governmental organizations like the US Military.[7] Existing literacy programs vary in their approaches for who to upskill (*e.g.* whether they target everyone in the organization, or only those in specific roles), how to deliver and develop materials, and the specific skills they focus on. This paper focuses on the specific design of the program that has been designed for Dow's R&D and TS&D organization, which focuses on five skill pillars (Data stewardship, Visualization, Coding, Statistics, AI/machine learning) and seeks to be flexible enough to accommodate a wide variety of backgrounds and skillsets found in that organization.

Background and organization

Target audience

The Citizen Data Science (CDS) program at Dow was established with the aim of up-skilling R&D and TS&D employees (3000+ people) in data literacy and proficiency, and to enable them to engage with digital technologies: increasingly digitized laboratory infrastructure, newer and more complex techniques for exploiting scientific data sets, and increasingly fast-moving and voluminous data flows. The company's R&D organization includes experts across many areas of science including chemistry, materials science, physics, formulation science, analytical characterization, process engineering, and product quality. The audience for CDS is not primed evenly with respect to digital ability: some researchers have practiced digital skills such as coding previously (*e.g.*, while in school), while others may be highly skilled in a non-digital discipline but have no prior exposure to "digital practice".

Organizational environment

Job roles in R&D and TS&D are diverse and thus cannot all be met by a rigid, one-size-fits-all digital literacy curriculum. For instance, some roles may require the ability to comprehend and write code for data analysis, whereas others may require no ability to code but require an ability to quickly compose informative visuals from lab data, *etc.* Job roles are also mutable, and one individual's present needs may be completely different from the individual's needs a year from now. Therefore, there is a need for a digital curriculum that can be pulled from on an as-needed basis, enabling researchers to acquire the specific skills needed for their job profile or career aspirations exactly when they needed it.

We believe that the structure of CDS is general enough that it could be adapted to other research institutions, or to other functions beyond R&D. The competencies we have defined in the CDS program are focused specifically on digital specific skills that we have seen the most significant gaps and need for investment in our organization. Other institutions considering implementing CDS or any other digital literacy program should consider the level of investment support within their organization, time constraints, and existing competencies within their organizations before choosing their path forward. See the Implementation challenges and learnings section for a more detailed discussion.

Data culture

As part of its upskilling charter, the CDS program is tasked with instilling a "data culture" in which there is an institutional sense of the intrinsic value of data: namely that data is both the enabler as well as the product of research. Therefore, data must be made accessible democratically, protected and governed from the moment of its inception to its intended destiny, which is to drive decisions and motivate actions that create value for the company and the world.[8] Data creates value when it retains its context, and data contextualization is inextricably linked to proper data management to preserve that context (*i.e.*, data stewardship). Borrowing from the "Data Science Hierarchy of Needs",[9] CDS is designed to facilitate the process of driving decisions from scientific data (Fig. 1).

Structure of CDS

CDS is structured around the five skill pillars of data stewardship, coding, statistics, AI/ML, and visualization. The curricula for the five pillars are intentionally kept largely independent of each other, such that individuals are empowered to upskill to the required level with minimum dependencies on upskilling in the other skills.

Curriculum development guidelines

We have found it highly useful to establish a set of fundamental guidelines along which to structure the various parts of the CDS curriculum (see ESI†). These guidelines help to focus the individual components of the curriculum in such a way that they naturally support and reinforce each other, while providing clear criteria by which training creators can include or exclude content modules as may be required by the organizational context. Such a framework is particularly useful for screening out training content that is easily accessible and highly promoted on the internet but in fact is diametrically opposed to sustainable digital R&D practices.

The CDS curriculum is purposely designed to be flexible and not depend on specific tools or technologies to demonstrate skills. The reason for this is to keep the program relevant as technological changes occur in the digital space. Instead, the focus is on fundamental understanding of the best practices for use of digital tools as well as their limitations.

Skill levels

Each pillar has been defined at four individual achievement levels (1 – essential skills only to 4 – deep expert in the area) (Fig. 2). As an individual proceeds from lower to higher skills levels, both the scope of knowledge as well as its depth are increased; so, for instance, a Level 2 coding practitioner is able to write a program to accomplish a certain task, but a Level 3 coder will also be able to optimize runtime performance based on time/space complexity considerations, as well as be able to deploy their code *e.g. via* web app. The ESI† includes several

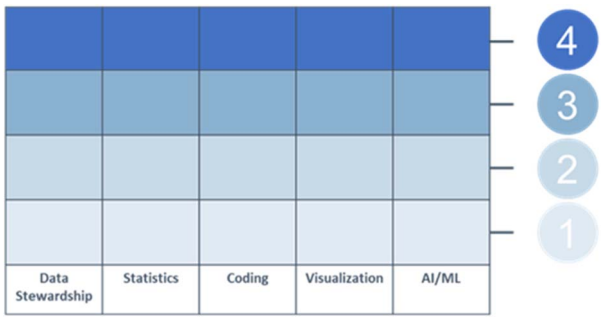

Fig. 2 The CDS skills matrix.

examples of potential profiles and skills that could match to specific roles in an organization.

Level 1 – essential: this level is designed for beginners and aims to provide researchers and supporting staff with the essential data science skills required for being part of a digital research ecosystem. Level 1 curricula are delivered *via* short classroom (virtually or in person) sessions; certification is based on attendance. Training events are given by researchers intimately familiar with the curriculum material, usually recruiting from Level 2+. Especially in strongly specialized environments (*e.g.*, analytical chemistry *vs.* engineering *vs.* materials science contexts), providing at least a limited number of references to the organizational context is important to help acceptance of the training. We have found that context is usually easy to accommodate in Level 1 training events (*e.g.* by demonstrating skills on specific datasets relevant to the attendees) and has disproportionately positive effects.

We found it beneficial if the trainer or a co-presenter originate from the local target organization because they speak the local "technical language" and can provide examples that illustrate contextualized use of best practices. For individuals who cannot attend the training, a self-paced learning platform is available on the company intranet, where certificates can be obtained after reviewing the materials and answering quizzes.

Level 2 – experienced: this level is for individuals who want to enhance their productivity or the quality of their research decisions within their project work by applying a skill level that can be acquired over several weeks or months by most researchers. Training content at this level is more in-depth and focuses on more complex analysis and data management techniques. Level 2 (and higher) training heavily leans on external options, which may cover multiple disciplines per venue. Depending on an organization's external partners and skills need at each level, different providers may be "plugged into" the curriculum to fill a particular need. Dow currently partners with Coursera, from which the CDS program picks individual courses suitable for a level and skill. At Dow, the CDS program has also vetted providers of web-based learning (*e.g.*, Dataquest) and has developed fruitful relationships with selected external vendors to deliver in-person courses several times per year for those who learn best in a cohort-based approach *versus* a self-paced approach. For instance, completion of a Posit Academy course results in Level 2 certification in Coding and Data Stewardship, and Level 1 certification in Visualization.

Generally, certification at Level 2 requires demonstrated ability, such as the completion of a training project or evidence of on-the-job practice.

Level 3 – champion: this level is achieved by those who have mastered application of the Level 2 skills such that their work products are leveraged to other projects or teams beyond their own, and/or they are recognized as a go-to resource for digital tools in the context of their research. Achieving level 3 certification requires strong skills which are usually acquired over months to years. The training content leading to this level of achievement is less directive and explicit than for Levels 1 and 2. Level 3 training – as currently implemented at Dow – is entirely self-driven due to the considerable diversity and variation between individual practitioners. It strongly relies on externally available resources, mentoring by experienced peers, and on-the-job practice. Level 2 certification confers enough subject matter knowledge to enable the trainee to accumulate the required knowledge and practice for Level 3 from publicly available courses and publications.

Level 4 – expertise leader: this level specifically caters to those who have advanced their skills and aim to play a leadership role in the data science field. Level 4 certification is attained by sustained practice and demonstration of skills at Level 3 over several years or through formal training (such as an appropriate degree). Level 4 certified individuals should be able to demonstrate that they can function as data science leaders and mentors to the organization in their aligned skill pillars.

Skill pillars

CDS recognizes five basic, data-centric skill categories. These categories collectively strive to provide a comprehensive grounding in data science, equipping individuals to effectively handle and interpret data. Progression in each category is structured by the program's different proficiency levels.

Data stewardship: this skill involves managing and organizing data effectively in the context of scientific research. Its focus is the use of effective data organization and storage strategies, maintaining data accessibility and veracity, and knowledge of raw data storage and data hierarchy principles. In terms of an equivalent curriculum, it is similar but not identical to data engineering and database design.

Coding: this track is for acquiring and improving coding skills which are fundamental in data science and research software engineering. The coding curriculum focuses strongly on the application of computer science to a practical research context. It is also designed to be a "catalytic" discipline for the other skills categories. The curriculum focuses on coding as relevant to data science in R&D/TS&D and is therefore centred deliberately on Python and R.§ It also purposely de-emphasizes scenarios that do not apply to a citizen practitioner context.

§ At the time of writing of this manuscript. This is a choice that is subject to changes in technology and based on a combination of factors: the suitability of the languages to the problem space, their vibrant ecosystems, and low barriers to entry and ability to accompany a practitioner from novice to high levels of accomplishment.

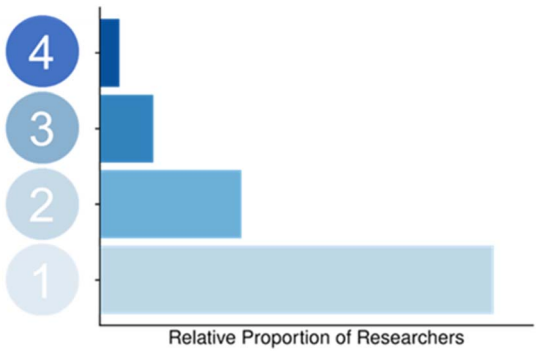

Fig. 3 Schematic, qualitative view of the envisioned skills distribution.

Statistics: this skill is about understanding and applying statistical concepts and methodologies in the R&D/TS&D context. By way of design-of-experiments, statistics is the basis for data generation and quantitative data analysis, leading to data-driven predictions and decisions.

AI/ML (artificial intelligence/machine learning): this track focuses on the application and understanding of AI and ML models. It is about learning how to create, implement, and interpret these models to derive meaningful insights from data.

Visualization: this skill involves the ability to visually represent data, an important aspect of data analysis and communicating data-driven findings. It covers knowledge of various visualization tools and techniques, from exploration through production dashboard delivery.

Distribution of skill levels

Organizational needs dictate the distribution of researchers across achievement levels. Dow's program implementation has not yet reached steady state. Therefore, we cannot give precise numbers regarding the proper target distribution. However, based on our current trajectory, we expect to achieve a steady state ratio of at least one Level 2 for ten Level 1 certificates. Level 3 practitioners and Level 4 practitioners for each skill are progressively more sparsely dispersed throughout an organization. This creates a pyramid in which the practitioners at Level 2+ act as advisors and mentors for lower levels and resources, and they have progressively higher degrees of skill and responsibility towards the top (Fig. 3).

Overview of the curricula

Data stewardship

We consider data stewardship the most foundational skill; it enables the other skills by allowing researchers to create data, handle it responsibly, and share it with others in ways that facilitate fidelity, transparency, and reproducibility.

The data stewardship discipline progression has the following skill requirements:

Level 1: at this foundational level, practitioners understand the basic concepts and the importance of data: structured data, rectangular ("Tidy") tables, proper use of identifiers (keys) and the relationships arising between bodies of data, and how to appropriately document and share data. The emphasis at this level is on being able to take part in the "R&D data economy" as an individual contributor. A practical example of this would be a researcher who compiles observational (*e.g.*, instrument) data into a table for their own analysis or as a feed into a larger data management effort.

Level 2: at this stage, an individual can fully implement the "Tidy Data" principles, as defined by Hadley Wickham.[10,11] They can create, clean, munge, and manage data as relational sets, which means they are conversant with pivot, joins, and query operations, using a platform of their choice (*e.g.*, SQL, PowerBI, R or Python, to name a few). The ability to code is optional at this level, as quite a few "point-and-click" tools exist that facilitate the required operations. A typical activity for a person at this level might be the aggregation of multiple instrument feeds into a cohesive set amenable to further exploration and analysis by a project team.

Level 3: a person at this stage can manage complex, heterogeneous data sets typical of larger research teams. In this context, the reproducibility of all data handling and analysis steps is important, as are well-documented processes – where possible, self-documenting. The person must be able to create performant, robust data pipelines connecting data sources with repositories and endpoints, appropriate to the project environment. To achieve this, knowledge of coding is required. At this level, a practitioner's ability approaches that of a professional data engineer within the specific environment of their research.

Level 4: a practitioner at this level has a proven track record of performing at Level 3 for a period of several years, delivering durable, performant solutions to their project team(s). They are expected to be proficient in all aspects of data stewardship and provide guidance to others as they work towards creating a data-centric culture in the organization. They should demonstrate leadership qualities, stay updated with the latest developments in the field, and contribute to the community of practice. An example of leadership in data stewardship would be identification of standard data models for analytical instruments (*e.g.*, Allotrope Foundation data standards[12]) and influencing the adoption of such standards across an organization.

Coding

In the Citizen Data Science program, the coding discipline is foundationally important as the enabler for advanced practice of all other skills. Progression has the following skill requirements:

Level 1: at this foundational level, researchers understand what code is and how it acts as an enabler in a modern R&D operation for efficient, transparent, and reproducible science. They understand what one can reasonably expect code to do. It is important to note that at this – introductory – level, there is no actual requirement for an ability to write or understand code. The focus is rather on being able to relate to coding practitioners and to follow data practices that enable computerized data flows, such as Tidy Data conventions.

Level 2: at this level, practitioners understand coding basics. They are expected to understand and be able to use

fundamental, code-based operations on data. The goal at this level is to develop a primary foundation in the use of a coding language such as R or Python in the context of scientific data management and analysis. A practitioner needs to be able to take data from the raw state to a finished analysis through visualizations and simple statistics. While interactive apps are not required, a basic knowledge of reporting through *e.g.*, a web portal is required.

Level 3: there is a substantial jump in the level of proficiency between Level 2 and this level; Level 3 demands a deeper understanding of code debugging, performance analysis, error handling, and the use of version control systems like Git in a team environment. Code quality, transparency and efficiency must be practiced, with a view towards broad leverage of developed code to others in the organization. Coders at this level need to understand code security, provide thorough documentation, and be able to deploy coded solutions to others, *e.g.*, through interactive web apps.

Level 4: individuals are expected to possess advanced coding skills, displaying proficiency across many areas of coding, and demonstrating the ability to lead and influence junior coders. They should also be capable of mentoring and advising others on coding best practices; staying updated with the latest developments in the field, including maintaining a connection to the larger professional information technology organization to align to corporate best practices; and contributing their expertise to the digital community at large.

Visualization

The visualization curriculum of CDS leans on commonly accepted best practices for visualization.[13] The visualization progression has the following skill requirements:

Level 1: practitioners learn basic principles and best practices of visualization. They understand the importance of clean, reliable data, and how certain data types naturally lend themselves to certain types of charts. They are aware of the common chart types relevant to R&D/TS&D and know how to create these charts in a commonly available tool of their choice (*e.g.*, Excel), using best practices. They know how to craft charts that efficiently convey the desired message emphatically but ethically, using heuristics that aid fast comprehension and avoidance of visualization anti-patterns.

Level 2: at this level, practitioners can craft complex visualizations with advanced plot types or combination of plot types resulting from difficult data sets and analyses, such as data resulting from data reduction techniques (*e.g.*, PCA). An additional difference from Level 1 is an ability to deploy visualizations as apps or dashboards to a broader audience in an appealing way that maximizes information content and comprehensibility. An example of this would be publishing a PowerBI dashboard to their workgroup.

Level 3: practitioners at Level 3 can create complex, interactive visualizations reproducibly through programmatic visualization pipelines. They document their pipelines thoroughly and understand the security implications of different data connection and publication methods. They understand how data needs to be structured in order to optimally build visualizations in their tools of choice and are able to manipulate data into that format using code or built-in transformation pipelines within the tools. They also advise others on how to improve their visualizations.

Level 4: practitioners demonstrate leadership in visualization. Based on their comprehensive knowledge of analytics tools and processes, they guide others in designing effective visualizations, choosing appropriate tools based on the visualization need, and propagating best visualization practices. They should have a robust understanding of audience-targeting with visualizations, and awareness of the latest trends in data visualization and available tools in the field.

Statistics

Within the CDS curriculum, the statistics pillar occupies a special position as it provides the mathematical backdrop for the entire practice of data science. In some programs, improper practice is encouraged by superficial "introductory" tutorials that gloss over difficult to understand mathematical background that is strictly necessary for the result to be valid ("cargo cult statistics").[14] To avoid instilling a deceptive sense of competence, the CDS statistics curriculum is crafted with a strong emphasis on dialog with professional statisticians with formal training (expertise leaders).

Progression in the statistics curriculum has the following skill requirements:

Level 1: aspiring level 1 practitioners are assumed to be already acquainted with basic statistical concepts and their practical applications; the training for this level focuses on the interplay between statistical techniques in an R&D/TS&D context. At this level, a practitioner should be equipped with the knowledge to have an informed dialog with a professional R&D statistician, based on an understanding of the relationship between measurement capabilities, design of experiments, and the outcomes of screening, prediction, and hypothesis testing.

Level 2: the experienced level focuses on the basics of industrial statistics: the training provides trainees with the skill sufficient for "personal use"—relatively simple analyses based on a blueprint provided by the coursework. It also aims to help those just looking to get a good grip on the foundational concepts in order to engage in informed conversations about the subject and to enhance synergy with statisticians. At this level, there is no promise of self-sufficiency. This is because the complexity of a rigorous design-of-experiments approach (DOE) in most practical applications in R&D today involves formal training in statistics.

Level 3: a researcher who reaches champion-level can demonstrate (a) a history of engagement with professional statisticians in project work and (b) that they are able to apply more advanced statistics in a subject that relates to their specific routine work. At Level 3, there is an increased degree of independence from professional statisticians, but not complete self-sufficiency.

Level 4: at this level, formal training in statistics is needed. An individual must have the experience and/or formal training

equivalent to an appropriate degree that would enable them to be employable in a statistics expertise center.

AI/ML

AI/ML is the "pinnacle" discipline of CDS; it leverages practically all other skills categories. Progression in this discipline has the following skills requirements:

Level 1: the goal of the Level 1 curriculum is to give researchers enough information to understand some basic concepts of AI and ML to enable them to relate to the concepts productively without unrealistic optimism nor unfounded concern. Level 1 training does not provide enough information to apply AI and ML to research projects but is geared towards enabling collaboration with AI and ML practitioners.

Level 2: researchers are introduced to multivariate analysis methods including partial least squares and principal component analysis, enabling researchers to analyse diverse datasets with the multivariate relationships often found in the chemical sciences. Practitioners are adept at identifying relevant mathematical concepts in algorithms, understanding model training as an optimization problem, and utilizing statistical concepts to describe data. They should be familiar with data preprocessing steps, differentiate between training, validation, and test sets, and employ methods like *k*-fold cross-validation for model testing. The ability to design model inputs and outputs, along with recognizing the need for varied modelling approaches for different problems, is crucial. They should be able to describe key machine learning models, comprehend regularization and hyperparameters, and grasp the impact of model type on interpretability. Additionally, they must define feature engineering, recognize the distinction between training and inference, and know the steps for model management.

Level 3: at this skill level competencies expand to include using mathematical concepts to explain model selection and performance in the application domain and selecting optimization algorithms tailored to the problem specifics. Because they must use statistical concepts to elucidate model performance and devise solutions, a CDS statistics level 2 certification is a prerequisite for this level. Practitioners are able to design virtual experiments to evaluate model performance and formulate assumptions with an understanding of their impact. A practitioner should identify model types based on the character of the data and modelling objectives, provide reasoning for ML method selection, use regularization and hyperparameter tuning, and apply interpretation methods suitable for the model and problem. They are expected to generate relevant features, specify models adhering to CDS coding track best practices appropriate for a Level 2 coder, and leverage machine learning operations (MLOps) tools for model management. This includes registering models, documenting with model cards, and incorporating model monitoring where appropriate.

Level 4: a top-level practitioner is recognized as a mentor and has influence beyond the particular workgroup or department they are a part of. They are acknowledged as subject matter experts in the field of AI/ML as applied to their chemistry or engineering expert domain. As part of this leadership role, Level 4 AI/ML researchers are responsible for keeping up with the literature on emerging AI/ML techniques external to Dow and/or the chemical/manufacturing industry, aligning these technology developments with the relevant problems within Dow, and proving both their technical feasibility and business value through project work (individually or through mentoring junior researchers). AI/ML expertise leaders must also advocate for the foundational infrastructure and resources needed to effectively accomplish AI/ML, such as data foundation and architecture, MLOps technology including model management and deployment platforms, and teams staffed with the data engineering and software development skills needed to be successful.

Results

Curriculum uptake and response

The CDS program is structured such that certification at Level 1 – essentials – is delivered to the organization on a group by group basis and (depending on group decisions) can be mandated ("push") to ensure a broad understanding of foundational digital principles. Certification at Level 2 and beyond is optional ("pull"). Thus far the Essential level curriculum has reached about half of Dow R&D personnel, with higher uptake (>90%) in certain functions of the organization.

One of the goals of the essential level curriculum is to help researchers understand the value of further upskilling. Thus far about 10% of individuals who have taken the essential level training have gone on to upskill and become certified at the experienced level or higher in at least one of the skill pillars.

A key piece of feedback that the team is actively working to address is that some trainees have found it difficult to relate the training to their own job roles, have trouble bridging the gap between levels (*e.g.* from essential to experienced), or find it too simplified to be useful. The CDS team is actively working to design more specific materials that give individuals an opportunity to use the skills hands on and become excited to apply them in their own work.

Impact of citizen data scientists

The Citizen Data Science program has been in development at Dow for four years. In order to measure the impact of the program, we have tracked the growth in use of Dow's data science workbench infrastructure (available since 2022). The data science workbench consists of a development environment for creating data pipelines, models, and documentation, as well as a platform for deploying interfaces to these to end users. Consistent growth in the number of developers, end users, and applications deployed using the workbench platform (Fig. 4) demonstrate that our community of trainees is actively practicing and that the community is growing both in size and in impact. It is difficult to quantify the exact impact of citizen data scientist efforts on productivity, but based on surveys of practitioners, a typical small deployed application may save 50–100 hours per year of manual work, which adds up to a substantial gain in efficiency across the entire R&D organization.

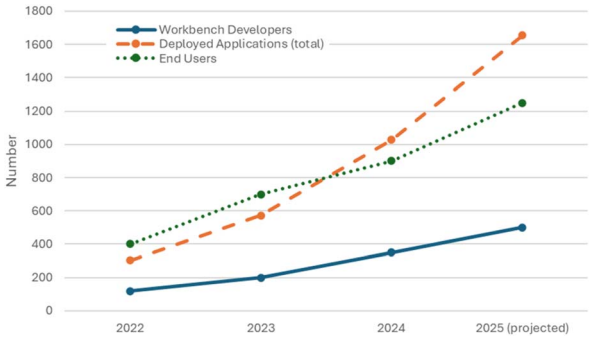

Fig. 4 Historical growth in the use of Dow's internal data science workbench infrastructure.

Implementation challenges and learnings

The implementation of the CDS program at Dow is ongoing, and we are still in the process of organizational learning. The curriculum details continue to be refined as we gather feedback from our organizations and identify additional learning partners. However, after four years of implementation and testing, we can report several learnings that will likely be applicable to other organizational contexts too.

Supporting infrastructure

Successful implementation of (citizen) data science rests on a foundation consisting of three major pieces of institutional and technological infrastructure:

(1) A guiding organizational strategy should be created that aligns and matches data science related efforts to the needs of the businesses and functions (a 'digital strategy' and the organization to support it). Success of the program depends on an organizational framework in which digital skills are valued and sought after. Digital skills must be an integral part of the project planning process and championed as an integral part of organizational growth strategy. It is also imperative that the organization include digital skill sets as part of hiring strategies and defines attractive and equitable career paths for upskilled employees.

(2) Appropriate data governance should exist to ensure that data remains secure but is also accessible and usable by data scientists. As the digital skill level of an organization rises, there is increased demand for access to data.[15] This increases the chance of accidental or malicious corruption or loss of the data. Appropriate implementation of a digital upskilling program requires a matching implementation of data and code ('model') governance processes that balance security requirements with an organization's need to utilize its data. While we cannot recommend what middle ground is appropriate for another organization and its data, we do note that ambiguity in this topic is a significant deterrent to progress; clear boundaries of what is and is not allowed will enable (citizen) data scientists to autonomously create value in a safe manner rather than become mired in bureaucracy.

Fig. 5 Successful implementation of a digital upskilling program requires foundational infrastructure ("data foundation") able to support the skills.

(3) A technology environment should be maintained that enables not only the highly visible "pinnacle" technologies (analytics, AI/ML) but the largely invisible foundational technologies (data collection, storage, search, handling, coding and model deployment platforms, Fig. 5) as well.¶ This includes well-defined and governed interfaces to foundational data systems into which citizen data scientists can tap, specific software or workspaces that citizen data scientists can use to develop and deploy their solutions, and well-defined guidance for citizen data scientists on how to use these tools appropriately.

In our opinion, the importance of maintaining a balanced investment in all three of these cannot be overstated. If any of these is absent, promoting a digital upskilling program like CDS will lead to wasted spending and disillusionment of aspiring practitioners because the newly acquired skills cannot be put into practice.

Change management

Even a perfect upskilling program will encounter some degree of organizational resistance. Resistance to digital transformation is by now well documented;[16] it should not be underestimated.

Since only the essential level curriculum is mandated within subsets of our organization, resistance to participation on the part of researchers is naturally limited to the initial step of the upskilling journey. We have found that three elements can be

¶ Such enabling data systems are, *e.g.*: chemical and/or materials registry systems, chemical/materials inventory systems, laboratory information management systems (LIMS), systems to store large amounts of raw data (a scientific data management system, SDMS).

implemented to reduce this initial resistance to Level 1 training: (1) contextualization of the material, (2) the local presence of practitioners and advocates, and (3) strong, visible leadership buy-in.

(1) Contextualization: providing workgroup-specific context, such as using example data sets or visualizations from a workgroup's natural context, helps break down barriers between trainers and trainees by creating a sense of shared background and purpose. We have found that context is usually easy to accommodate in Level 1 training events and has disproportionately positive effects.

(2) Local presence: trainees who are not sufficiently aware of the systemic effects that good (or bad) digital practices have can create local islands of "bad data" that may spread depending on the individual's local influence. For instance, a person in a traditional role whose responsibility is to create data but who never analyses it (such as, for instance, an instrument operator) may not understand why it is not good practice to share data as a screenshot from an instrument as opposed to a tidy data file. It is important that Level 1 training demonstrates, in a credible and relatable way, that the utility of the data is compromised if best practices are not followed, and that it is therefore important not to see oneself in a lone actor position where one can be exempt from being data literate or using best data practices. We believe that the presence of advocates for good data practices, who can counteract these tendencies at the local (*e.g.*, workgroup) level is an important factor in the success of a CDS implementation. Such advocates need not necessarily be managers; locally credible technical practitioners tend to be very effective advocates because they can lead by example.

(3) Leadership buy-in: emphatic, credible, and demonstrated buy-in and support on part of local leadership is essential in breaking down initial credibility barriers that the training encounters. Many employees suffer from corporate-initiative fatigue and are understandably sceptical towards new initiatives. This may be especially true in a generally overhyped field like digitalization. The presence of, and an introduction by, local leaders within training sessions is essential to overcoming employees' sense that yet another irrelevant initiative is foisted upon them. Leaders should emphasize why the training is relevant, and how it is aimed at removing friction from the data economy that ultimately will make everyone's job easier. The importance of this cannot be overstated; without demonstrated and sustained leadership buy-in at all levels of the organization, CDS will not work.

Expectation management

During this time of monthly breakthroughs in digital technology, mindshare is dominated by digital "shiny objects", *i.e.* technologies that have become hyped (such as AGI or quantum computing). There is a temptation to focus digital learning on these stand-out technologies without regard to the fact that these pinnacle technologies rest on an enormous foundation made up of good data practices paired with sound subject matter expertise in the scientific context surrounding the data (*e.g.*, materials chemistry). Likewise, one will often find a willingness to fund infrastructure or purchased services related to pinnacle technologies (*e.g.*, new ML platforms) at the cost of investing in the infrastructure and data curation that supports the new ML platform with data. In both cases, the foundation upon which the pinnacle technology/skill is built is not directly visible to the decision makers (Fig. 5). What is visible and attractive, both in terms of upskilling and infrastructure, is the tip of an iceberg that cannot be supported in the absence of exactly the kind of knowledge that CDS attempts to instil.

Proponents of a digital literacy program such as CDS must therefore pay scrupulous attention to educating leadership and learners at all levels about the hierarchy upon which pinnacle technologies such as large-scale AI are built, and that training in AI-for-everyone may sound appealing but is probably counterproductive in that it distracts researchers from acquiring vital, foundational skills that will be necessary if AI-for-everyone is to become reality for their company.

Inter-organizational communication

A large fraction of the digital practitioners in a company (coders, data engineers, *etc.*) traditionally reside within the IT organization. The practice of data science in an R&D context requires skills that are only superficially related to the skills profiles supplied by these digital-native organizations. Citizen Data Science for R&D/TS&D is a mixture of both digital skills (*e.g.*, coding) and deep subject matter expertise in chemistry, materials science, or engineering. The two realms (R&D/TS&D and IT) share little, if any, common vocabulary, and dialog across the interface is often laden with difficulties and misunderstandings. A digital upskilling program should have focal points in R&D who also understand and can interface with the IT organization to make sure that the needs of R&D/TS&D citizen data scientists are considered in policy making and infrastructure building. They in turn can credibly transmit the requirements imposed by the larger IT organization to the Citizen Data Science community.

Resourcing

Researchers who are assigned to develop, train, and organize CDS content need to take time away from other responsibilities, which can cause challenges in assigning resources to accomplish these tasks. The reality of corporate R&D/TS&D life is that attention is at a premium; the time an individual can devote to upskilling is limited in amount and may be available only in small instalments. Therefore, it is important that information at all skill levels is available and consumable in a way that places the least possible burden on the trainee.

The cost of digital upskilling is not negligible and should also be factored into the decision about adopting a digital upskilling program such as CDS. Trainees will in aggregate spend a significant amount of time in training. To illustrate, if a full course of Level 1 trainings takes seven hours per individual, an organization with 1000 researchers will spend an aggregate of 7000 hours in training to achieve a homogeneous adoption of Level 1 skills. Another readily available example

would be the cost to upskill an organization of 1000 researchers to a 10% permeation of Level 2 coding skills. Typically, a researcher will spend five hours per week for ~12 weeks to acquire Level 2 coding competency. The time cost per researcher is thus 60 hours, and $60 \times 1000 \times 0.1 = 6000$ hours for the entire organization. This does not include the expense of facilitating the course through externally available platforms/vendors.

Conclusions

Implementation of a digital upskilling program in a heritage chemicals and materials R&D and TS&D environment requires a comprehensive, strategic approach that allows individuals to acquire essential digital skills quickly while not burdening them with knowledge that is not essential to their role or career plans. In response, we have developed the "Citizen Data Science" (CDS) framework in which digital upskilling trajectories can be tailored to individual need from a matrix that offers five skill types (data stewardship, coding, visualization, statistics and artificial intelligence/machine learning) at four skill levels. We recommend that everyone in R&D/TS&D master the first level ("essentials") to acquire the basic "digital literacy" that is necessary to function in a digitally enabled organization. Progression beyond the first level can be fully tailored according to career requirements.

Any change to an organization will encounter some amount of resistance: CDS requires strong, sustained, and vocal leadership buy-in for successful implementation. A more skilled workforce will also put additional stress on the information technology infrastructure of the R&D/TS&D organization. Data systems must be prepared to accommodate this to shield the newly upskilled individuals from data starvation. Data access and security tend to become issues as the need to have free access to data collides with an industrial organization's need to protect its intellectual property. Successful implementation of the CDS program requires a guiding organizational strategy, appropriate data governance, and a supportive technology environment.

In order to reduce the need for digital upskilling programs in industrial settings, academic programs (not just for STEM degrees, but across a broad breadth of majors) can and should focus coursework on specific best practices and practical skills needed to be effective in a digital world. Practical skills like data management, visualization and other skills that form a foundation of data and digital literacy are important at all levels of modern industry, from higher management to team members[17] and historically, industry has had to fill the gaps between digital needs and digital skills of trainees coming from school.[18,19] Coursework targeting practical data literacy skills is starting to exist at many universities and further development of and attendance to such coursework should be encouraged.

Data availability

As this is a *Perspective* article, no primary research results, data, software, or code have been included.

Author contributions

Kyle Andrews, Steven Arturo, Matt Benedict, Birgit Braun, Brian Clark, Simon Cook, Jaime Curtis-Fisk, Fabio D'Ottaviano, Tim Licquia, Peter Margl, Jonathan Moore, Lynette Naler, Parth Singh, Alix Schmidt, Anatoliy Sokolov, John Talbert, James Wade: conceptualization, investigation, methodology. Matt Benedict, Peter Margl, Alix Schmidt, Anatoliy Sokolov, James Wade: writing-original draft, writing-review and editing.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We acknowledge Paul Sanderude, Clark Cummins, Nathan Wilmot and Dave Magley for helpful discussions and advice. The team is particularly grateful to Andre Argenton for his vision, leadership and support.

Notes and references

1 L. Goswami, M. K. Deka and M. Roy, *Adv. Eng. Mater.*, 2023, **25**, 13.
2 "Future of Jobs Report 2023", World Economic Forum, 2023, **https://www.weforum.org/publications/the-future-of-jobs-report-2023/**.
3 M. Fenlon and B. Fitzgerald, Reskilling – A solution for the digital skills gap, *PwC Business Higher Education Forum*, 2019, **https://www.bhef.com/publications/reskilling-solution-digital-skills-gap**.
4 S. Merkelbach, S. V. Enzberg, A. Kühn and R. Dumitrescu, Towards a Process Model to Enable Domain Experts to Become Citizen Data Scientists for Industrial Applications, *2022 IEEE 5th International Conference on Industrial Cyber-Physical Systems (ICPS)*, Coventry, United Kingdom, 2022, vol. 05, pp. 1–6.
5 *Merck KGaA on Training a Data-Capable Workforce*, ed. T. Cohen, Oxford Global, 2022, **https://oxfordglobal.com/discovery-development/resources/merck-kgaa-on-training-a-data-capable-workforce**.
6 Cultivating a Digital Learning Culture, Berkeley ExecEd, **https://executive.berkeley.edu/thought-leadership/case-studies/cultivating-digital-learning-culture**.
7 *Data Literacy Courses – U.S. Army Training and Doctrine Command*, US Army, **https://www.tradoc.army.mil/ocko/training-portal/data-literacy-courses/**.
8 M. D Wilkinson, M. Dumontier, I. J. J. Aalbersberg, G. Appleton, *et al.*, *Sci. Data*, 2016, **3**, 160018, **https://www.gofair.foundation/**.
9 M. Rogati, *The AI Hierarchy of Needs*, Hacker Noon, 2017, **https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007**.
10 H. Wickham, *J. Stat. Software*, 2014, **59**, 1.
11 H. Wickham, M. Çetinkaya-Rundel and G. Grolemund, *R for Data Science*, O'Reilly, 2nd edn, 2023.

12 Allotrope Foundation, **https://www.allotrope.com**.
13 C. O. Wilke, *Fundamentals of data visualization*, O'Reilly Media, 2019.
14 P. B. Stark and A. Saltelli, *Significance*, 2018, **15**, 40.
15 L. DalleMule and T. H. Davenport, *Harvard Business Review*, 2017, vol. 5, **https://hbr.org/2017/05/whats-your-data-strategy**.
16 A. B. Scholkmann, Resistance to (Digital) Change, in *Digital Transformation of Learning Organizations*, ed. D. Ifenthaler, S. Hofhues, M. Egloffstein and C. Helbig, Springer, Cham, 2021, DOI: **10.1007/978-3-030-55878-9_13**.
17 Numeracy: The New Literacy for a Data-Drenched Society, ACSD, 1999, **57**, 2.
18 J. Bersin and M. Zao-Sanders, Boost Your Team's Data Literacy, *Harvard Business Review*, 2022, vol. 2, **https://hbr.org/2020/02/boost-your-teams-data-literacy**.
19 E. C. Hilty, V. Vilski, S. Mishra, P. Condon and S. M. Gracia, Building an Inclusive Data Literacy Community, *Harvard Business Review*, 2025, vol. 1, DOI: **10.1162/99608f92.d622eaff**.