How I Met Your Data Science Team: A Tale of Effective Communication

Aayushi Roy, Deepthi Raghunandan
Department of Computer Science
University of Maryland
College Park, MD, USA
{aroy2530, draghun1}@umd.edu

Niklas Elmqvist
College of Information Studies
University of Maryland
College Park, MD, USA
elm@umd.edu

Leilani Battle
School of Computer Science & Engineering
University of Washington
Seattle, WA, USA
leibatt@cs.washington.edu

P6 Comp. Scientist Earth Science
I've had to adapt to how [the domain experts] work...
P2 Data Scientist Bioinformatics
If the director's in the meeting, it has to be PowerPoint, he does not want notebooks...
P3 Data Scientist Medical Tech
Sometimes, I'll have a Tableau page open, and then I can manipulate it live in the meeting.
P13 Data Scientist Aviation
I think most of the communication issues are really talking to stakeholders that are not as technical.
P1 Postdoc Bioinformatics
PI's don't really know how to use GitHub. They may not know how to use R. They just want to see graphs.
P10 Data Scientist Insurance
Very rarely would show [business people] the code...

COMMUNICATION COLLAGE. Communication between data scientists and domain experts in data science teams can be challenging. The above quotes (with source attribution as participant **PX**) are taken directly from our interview study involving such teams. (Figures generated by MidJourney v5.1.)

***Abstract*—Deriving actionable insights from data requires expertise in both data science as well as the specific application domain. This need for domain-specific knowledge often necessitates engaging an interdisciplinary team rather than an individual for most realistic data science problems in domains such as finance, biology, or drug discovery. This, in turn, requires effective collaboration between team members. This paper seeks to understand common themes in how such multi-disciplinary teams communicate to accomplish their analytical goals. We conduct an interview study with 15 professional data scientists working in small to large organizations in fields ranging from bioinformatics to accounting. Communication between individuals in these teams depends on their team structure and expertise. Data scientists specifically adapt their tools to communicate with team members with different types of domain knowledge. We discuss the strengths and weaknesses of these approaches for supporting communication in multi-disciplinary environments.**

***Index Terms*—data science, collaboration, communication, teams, interdisciplinary, qualitative evaluation, interview study.**

I. Introduction

Data science is intrinsically collaborative [36], [42]. The scale and complexity of real-world data science projects often necessitate coordinating tasks and communicating results among peers [29], [42]. Furthermore, the domain knowledge required to understand specialized datasets can surpass that of regular data scientists, requiring communication with subject matter experts [26]. For example, making hydrogen gas more affordable requires consistent communication between chemists and data scientists [45]. Similarly, data scientists must communicate with earth scientists to gain an understanding of how satellite imagery can be leveraged to detect trees [3]. While collaboration and communication among data scientists has been studied in-depth [27], [30], [36], [42], we see comparatively little work on understanding how data science tools may impact the way interdisciplinary teams collaborate on data science projects [6], [26], [38], [41].

We posit that **a data science tool's efficacy in inter-**

disciplinary contexts depends on its ability to facilitate communication between team members with disparate expertise. To test this hypothesis, we conducted interviews with 15 professionals who regularly collaborate on data science problems in diverse fields ranging from commerce to bioinformatics. Participants were asked about their workflow and team composition and the tools and processes they adapt to aid their communication. We transcribed and qualitatively coded these interviews—identifying key patterns and themes in team dynamics, tool usage, and communication strategies. While a few of our participants had domain expertise, all primarily self-identified as data science experts. This limits our understanding of the challenges domain experts and teams as a whole face in interdisciplinary communication. However, we are still able to gain valuable insight into interdisciplinary communication from the perspective of the data scientist.

Through our analysis, we contribute a framework for reasoning about communication among interdisciplinary teams based on the expertise and level of involvement of team members. Using our framework, we successfully map observed communication challenges from our interviews to mismatches between team composition and their chosen communication tools. For example, if data science experts use raw code to communicate results to domain experts, they will likely run into communication challenges. Potential miscommunications can be overcome by way of "intellectual bridges" that translate knowledge between the groups.

Based on our findings, we derive design recommendations for new and existing data science tools. For example, we find that more research is needed to help team members from different disciplines establish a common language for communicating analyses and results. We also recommend integrating in-situ presentation features into data science tools as a means to facilitate stronger and more regular communication across diverse data science teams.

In this paper, we contribute: (1) a qualitative study of 15 data scientists working on multi-disciplinary projects (Section III); (2) a framework to understand communication in interdisciplinary teams (Section IV); (3) and design recommendations for common data science tools to aid better communication in data science teams (Section V).

II. Related Work

There is a growing body of research on how data scientists can work together to develop a shared understanding of their data. We highlight the current understanding of collaboration in data science, review current strategies and tools within data science that enable collaboration, and establish a definition for effective collaboration.

A. Collaboration in Data Science

The practice of data science involves data integration, analysis, and interpretation [36]. Data integration involves collecting, cleaning, and formatting the data for inquiry and analysis. Data analysis is the process of deriving insights from data and ranges from being exploratory to directed [1]. Data interpretation involves finding a larger context for those insights or using them to facilitate decisions. These steps make up a data science workflow and are practiced iteratively.

Passi and Jackson [30] conduct an ethnographic study of corporate data science teams to find diverse expertise in those teams. As teams have a variety of members, they need to collaborate to maintain trust and manage issues. Zhang et al. [42] define five major roles in data science teams: communicators, domain experts, data scientists, managers, and researchers. Their user study finds that domain experts and researchers actively collaborate with data scientists across all phases of the data science workflow, highlighting the need to study these relationships to design tools that ease these types of collaborative efforts.

As data science is highly dependent on data and data comes from many interdisciplinary backgrounds, data science teams tend to include experts from multiple fields. Boukhelifa et al. [2] provide use cases where multiple data experts from various fields work together to provide deeper and better insights, which calls for collaborative tools to assist such experts. Mao et al. [26] interview biomedical researchers and data scientists and find that the biggest hurdle in interdisciplinary collaboration is finding common ground in terms of a question to solve or a mutually beneficial workflow. They also find that data scientists do not often engage the domain experts, as the data scientists found they did not need to understand the data to find a successful solution. These findings directly contrast the results of the study from Zhang et al. [42], who observe more active communication between the two parties. Kross and Guo [24] also study how data scientists actively collaborate with their clients over the lifetime of a project. They characterize this communication as a six-stage workflow that includes bidirectionally bridging the gap of knowledge between data scientists and domain experts. We focus on how, if at all, the gap of knowledge is bridged in our participants' teams.

1) Axes of Collaboration: Collaboration is broadly categorized on two axes: time when the collaboration occurs (synchronous and asynchronous) [4], [15], [23] and how closely individuals collaborate [28], [42]. Collaboration may overlap or occur in parallel to achieve a common objective.

Synchronous collaboration occurs when real-time editing is performed [4] with the primary goal of reducing communication costs and encouraging shared collaboration [36]. Asynchronous collaboration happens when members conjoin their efforts towards achieving a central goal [15]. Additionally, asynchronous collaboration occurs when resources are shared with team members and managers outside of meetings [4]—perhaps from different geographic locations and times [14].

Olson and Olson [28] introduce the concept of coupling in work to describe levels of collaboration. Tightly coupled work is interdependent and requires frequent and iterative communication between team members, whereas loosely coupled work can be performed more independently and requires less frequent communication [28]. This concept is echoed by Zhang et al. [42], who use language from *value sensitive* *design* [11] to refer to immediate team members as direct stakeholders. Collaborators who do not interact with the data directly but are still involved or impacted by the work are referred to as indirect stakeholders [42]. We use these concepts to frame collaboration in our participants' teams, particularly in terms of the level of involvement of collaborators.

It is necessary to understand the collaborative strategies of the team to build tools catering to collaboration. For example, Kang et al. [18] suggest that analysts work independently and only collaborate asynchronously when they fall into issues. Chung et al. [5] suggest that this may be because most collaboration is intrusive to the analyst's workflows.

B. Tools for Collaboration

Since collaboration occurs among various members of data science teams, there is a need for tools that will help communicate and coordinate within these teams. Current tools help team members engage each other and communicate using visualizations and analytical programming.

Computational notebooks are great mediums for collaboratively communicating insights and the analysis process because they exemplify the literate programming paradigm [22], where natural language annotations scaffold pivoting of data representations. However, most notebooks do not natively support collaboration, instead relying on source revision control systems such as Git for asynchronous collaboration. Some notebook platforms such as CoCalc and Google Colaboratory do support synchronous editing, but they were found to lead to inefficient data science collaboration [36]. A larger body of work focuses on easing communication with notebooks.

Data explorations are often "shaped" into a narrative to communicate the process and results [21], [32] in the notebook. When the audience is technical, the analyst may focus on retaining all details and laying them out in a comprehensible manner [21], [32]. When the audience is broader or non-technical, analysts may remove details that appear irrelevant and add more explanatory text—shifting the focus from the process to the analytical narrative [21], [25], [32]. This effort becomes particularly difficult as analysis becomes more complex [13], [19] or the nature of the work becomes more collaborative [20], [23], [36], [37]. Tools like Callisto [37], ToonNote [17], ForkIt [40], Code Gathering Tools [13], and Slide4N [39] focus on presenting solutions in this space.

The MIDST system by Crowston et al. [8] facilitates coordination between multiple analysts with the help of a shared code base and project management setting. Further, Ganji et al. [12] use visual cues to ease team discussions and coordination of the coding process to provide a streamlined coding pipeline for teams. Zhao et al. [43] and Crowston both represent an analyst's findings such as research articles, datasets, visualizations, and workflows of code as graph models to asynchronously collaborate with other teammates.

C. Effective Collaboration

Prior researchers have recognized consistent outcomes that indicate a successful collaborative effort in observing collaborative relationships and tools. We synthesize these outcomes to create a formal definition of effective collaboration, which we will use to evaluate existing data science tools.

Effective collaboration:

- minimizes the friction [10], [36] of collaborating,
- produces benefits for all participants in not only a final product [18], but also individual growth in terms of knowledge and skills [33], [36], and
- encourages the exploration and implementation of new ideas [9], [36].

Edwards et al. characterize the cost of communication as the time, energy, and attention spent communicating data or ideas across platforms [10]. Wang et al. [36] find that synchronous editing reduces communication costs by allowing individuals to explore and collaborate in a shared environment. Thus, minimizing the friction of collaborating entails reducing the number of steps or platforms needed to communicate an idea—whether through documentation or another person.

Benefits for the team consist of three main facets:

1) A final goal or product is achieved. Kang and Stasko [18] describe three types of collaboration: sharing, content, and function.
2) The collaboration boosts productivity. Teasley et al. [35] and Olson and Olson [28] collected productivity measures such as project length to determine how collaboration influences productivity.
3) Every team member achieves some level of individual growth, such as expanding their knowledge base. Schleyer et al. [33] assert that collaborations can be maintained only if all members benefit from participation, and that motivations such as individuals' needs for knowledge and relationships can influence collaboration.

Finally, the collaboration process should encourage team members to explore, expand upon, and implement new ideas towards the final product. This exploration can be an individual process or shared between team members [9], [36]. More specifically, Cummings and Kiesler [9] find that multidisciplinary collaboration promotes the innovation of ideas, and Wang et al. [36] find that synchronous editing encourages collaboration in a shared context.

III. Method

In this paper, we aimed to determine *how a data science tool's efficacy in interdisciplinary contexts depends on its ability to facilitate communication between team members with disparate expertise.*

To test this hypothesis, we interviewed 15 data professionals about their work environments and the tools they used. We evaluated a data science tool's efficacy by how participants' tool usage mapped to facets of effective collaboration. We define an "interdisciplinary context" as a project where at least two disciplines are represented by different members on the team. We acknowledge that all our participants primarily identified as data scientists, which limits our insight into how domain experts perceive such tools to be effective for communication. Below, we describe the details of the design and intuition for our interview study. This study was approved by our institution's IRB and preregistered on AsPredicted[1] after being piloted. Supplementary material, including the full set of interview questions and qualitative coding, are in our OSF repository[2].

A. Participants

We recruited participants primarily through our professional networks via email, social media, and word of mouth. To qualify for the study, participants had to work with data and perform data science work as part of their job description. We recruited 15 participants across 14 different workplaces and aimed for a diverse range of participants in terms of job profiles (data scientists (P2, P3, P5, P9, P10, P13), machine learning engineers (P4, P7), scientists (P11), etc). All but one participant was based in the United States. Table I provides an overview of the participant pool. An expanded version can be found in the supplemental.

TABLE I: **Participant demographics.** Gender, job title, and industry for each study participant.

P#	Gender	Job Title	Industry
P1	Male	Postdoctoral Associate	Bioinformatics
P2	Female	Data Scientist	Bioinformatics
P3	Male	Data Scientist	Medical Technology
P4	Male	Machine Learning Engineer	Commerce
P5	Male	Data Scientist	Business
P6	Male	Computational Scientist	Earth Science
P7	Male	Machine Learning Engineer	Finance
P8	Male	Machine Learning Scientist	Healthcare
P9	Male	Data Scientist	Commerce
P10	Male	Data Scientist	Insurance
P11	Male	Scientist	Bioinformatics
P12	Male	Computer Engineer	Earth Science
P13	Male	Data Scientist	Aviation
P14	Male	Professor	Earth Science
P15	Male	Machine Learning Researcher	Earth Science

B. Protocol

Each participant signed a consent form informing them that the study focused on collaboration in data science without revealing any of the interview questions. The participants also completed a demographic survey in which they responded to questions regarding their gender and occupation. The interviews were then conducted in a semi-structured format and were split into two sections.

a) Section 1: General Background: This section focused on establishing a general understanding of the participant's work environment in terms of the participant's responsibilities and the industry they worked in.

b) Section 2: Tools and Communication: This section delved into the research question, where we asked who participants communicated with and what tools they used in the contexts of both a long-term data project and a short-term data project. We chose to ask about these two contexts to understand how team communication may change depending on the length and scope of the project. Participants were also asked how effective they perceived collaboration with team members to be in regards to communicative and tool-related friction and collaborative benefits. In Table II, we present the main types of tools our participants utilized to communicate with their team members.

In Table III, we present some of the questions asked. Participants were interviewed for an average of 45 minutes over the Zoom video conferencing platform. Audio and video were recorded via Zoom and interviews were transcribed. Participants were compensated with a $20 gift card upon completion of the interview.

C. Data Analysis

We used grounded theory to further our comprehension of how analysts communicate with their teams and the tools used throughout their projects. The interview data was analyzed over two iterations.

a) General Coding: Two of the authors independently coded the same three interview transcripts to develop and reach agreement on the coding scheme in terms of length and breadth of codes. One author then used this coding scheme to code the remaining interviews. The codes were then aggregated and two of the authors looked for patterns relating to the interview goals. For example, several participants discussed using different tools and methods to communicate with domain experts versus data science experts.

We noted that participants tended not to collaborate when working on short-term projects, and the tools they used rarely changed between short-term and long-term projects. Given that our focus is on how teams collaborate and communicate, we chose to focus on participant responses in the context of long-term projects, where all participants collaborated with at least one other individual.

We also coded for effective collaboration by noting mentions of communicative or tool-related friction, benefits of collaboration, or exploration of new ideas. We found that all participants felt they benefited from collaborating with others and that collaboration encouraged the exploration of new ideas. However, friction was particularly indicative of how effective collaboration was between team members, and thus we evaluate tools on their impact on friction.

b) Coding for Communicative Relationships: Here we coded each interview for communicative relationships. We define a communicative relationship by the individuals involved, the purpose of the communication, the frequency of communication, and the tools used to communicate. We then looked for similarities in communicative relationships across participants. We noticed that the types of individuals involved in communicative relationships varied distinctly between participants. The purpose of communication and the tools used subsequently hinged on the individuals involved. For example, some participants communicated primarily with data science experts or domain experts, whereas other participants communicated with both. Thus, we grouped participants based

[1] https://aspredicted.org/yp3h8.pdf

[2] https://osf.io/2vzta/?view_only=01de3f6862bf4ca1916b2a5857c94ac8

TABLE II: **Participant tool usage.** Tools used by our 15 data scientist participants to communicate.

Tool Category	**Participants**	**Example Tools**
project management	P1, P2, P3, P6, P7, P8, P9, P10, P11, P12, P13, P14, P15	Slack, Microsoft Teams, JIRA, Gitlab, Github
development environment	P1, P2, P3, P4, P5, P6, P8, P9, P10, P11, P12, P13, P14, P15	Jupyter Notebook, Databricks, RStudio
documentation/presentation	P2, P3, P5, P6, P8, P10, P11, P12, P13, P14, P15	Google Docs, Microsoft Word, Confluence

TABLE III: **Sample Interview Questions.** Representative sample of interview questions asked of our 15 participants.

Main Idea	Sample Interview Questions
✥ Work Environment	- What kind of data do you usually collect and/or work with? - What are your responsibilities in the team? - How many team members do you have and how frequently do you work with them?
💬 Communication	- Who do you communicate with, collaborate with, or report to on this project? - What information did you share or communicate with each collaborator?
🛠 Tools	- What tools, processes, or documents did you use to achieve this communication? - Are there tools that you currently use that you feel fall short or do not fully meet your needs?

on similarities in team structure and who they communicated with.

IV. Results

This paper focuses on investigating the hypothesis that a data science tool's efficacy in interdisciplinary contexts depends on its ability to facilitate communication between team members with disparate expertise. Drawing from our examination of the communicative relationships of our participants, we present a framework for characterizing communication among interdisciplinary teams based on their composition, particularly in terms of expertise and level of involvement. We discuss the three team compositions we observed across participants and the role data science tools play in these teams.

A. Framework

Our framework consists of two main factors that influence how individuals communicate within data science teams: the *primary expertise* of the individuals involved in the communication and their *level of involvement*.

An individual's primary expertise delineates between data science expertise and expertise in the applicable domain. We observe a part-to-whole relationship between data science expertise and domain expertise, which can be represented as a Venn diagram. Team members fall into this Venn diagram, having either predominant knowledge in data science, the domain, or in both. We consider individuals who have knowledge of both data science and the domain "intellectual bridges" who can aid in crossing the barrier between the two fields. We found that teams with data science experts and domain experts but no intellectual bridges are indicative of dysfunction in team communication because the lack of common ground [6] inhibits effective communication. Meanwhile, teams comprised entirely of only data science or domain experts tend to avoid such dysfunction (although their versatility is reduced).

The second factor in the framework recognizes that team members are often involved in the project to varying degrees. For example, a data scientist may work daily with fellow data scientists to communicate progress and go back and forth on technical details, but only present high-level results to a client or leadership on a few occasions. We refer to team members who are worked with frequently and iteratively as "dedicated" collaborators, whereas team members who are interacted with less frequently are considered "casual" collaborators.

Depending on the project's requirements, these communicative relationships are mixed and matched to form a team.

B. Recurring Themes in Team Structure

We identify three distinct patterns in team structure across participants based on our framework:

1) **Team Structure A (The Token):** our participants were the sole data scientists in their team and worked with one or more domain experts as an intellectual bridge (Figure 1);
2) **Team Structure B (The Coven):** our participants had overlapping data science expertise with team members and partners, but did not actively interface with domain experts (Figure 2); and
3) **Team Structure C (The Two Cities):** our participants shared data science expertise with several other team members (data scientists, engineers, or computer scientists), and also worked with domain experts (Figure 3)

Below we describe these main types of collaborative relationships observed in each group and how these factors influenced data science tool usage.

1) Team Structure A – The Token: Not all of our participants had or sustained relationships with other data scientists (P1, P2). Some participants were the only ones in their teams (1-2 people) who performed data science work—essentially, they were the "token data scientist" collaborating with several domain experts. We consider these participants to be intellectual bridges who have knowledge in both data science and the pertinent domain, and use this strength to communicate analysis insights to domain experts.

These participants were generally from *research* backgrounds with dedicated team members who provided fund-

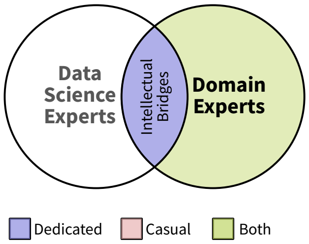

Fig. 1: Team A – The Token. These teams have a single token data scientist who, by necessity, takes on the role of an intellectual bridge to speak to a group of domain experts.

ing and domain expertise but did not have data science experience. The collaborative relationships with these team members were described as close and involving regular frequent (daily/weekly) and ad-hoc meetings to discuss process and progress. Participants also communicated with a wide audience of casual domain experts, such as funders and fellow researchers (P1), or leadership in the company (P2).

When collaborating with dedicated domain experts, participants avoided relaying more intermediate details about the data science process than necessary. P1 presented high-level overviews of his insights in the form of polished oral and visual presentations to his supervisor given that they were "more interested in results. And, you know, PIs don't really know how to use GitHub. They may not know how to use R. They just want to see graphs and maybe hear you kind of explain the analyses that were conducted and have a broad picture of what's going on." P2 sent computational notebooks over Slack as read-only HTML or showed them live via screenshare, as her collaborators felt uncomfortable working with the raw code. She found computational notebooks to be "very useful for working with people who are still learning how to code, or maybe just don't have that solid software development experience" (P2).

Presentations originally made for dedicated collaborators were adapted for a wider audience of casual collaborators. P1 stressed the need to ensure reproducible results for casual collaborators and took great pains to choose and customize open-source tools such as Jupyter Notebook and Singularity Container to reach a wider audience. P2 noted that the tools used to communicate results varied depending on whether the audience was dedicated or casual.

> *"If the director's in the meeting, it has to be PowerPoint, he does not want notebooks... if the director is not in the meeting, if it's more of just like literally the core team, then HTML notebooks."* (P2)

These participants considered their communicative relationships to be effective and relatively frictionless as a result of the tools available to them. They relied heavily on computational notebooks in conjunction with presentations and were able to act as an intellectual bridge to explain results to a broad audience of domain experts.

2) Team Structure B – The Coven: Another set of participants actively worked with one or more data scientists ("a coven of data scientists") towards the same data science objective (P3-P5). The participants' managers took on the role of engaging with the stakeholders or domain experts, who were often part of other internal teams in the company. Thus, these participants rarely presented directly to domain experts unless to report on a major result.

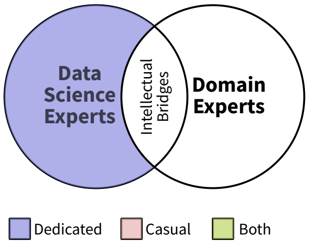

Fig. 2: Team B – The Coven. These teams mainly comprise specialized data scientists.

Participants primarily worked for larger companies and in medium to large teams of other data scientists. Their fellow data scientists were dedicated collaborators, and they communicated frequently through weekly and ad-hoc meetings. Individual objectives and tasks were often delegated to our participants during regular team meetings.

In regular meetings, participants were asked to share and present their process and progress to other data scientists with code, visualizations, or code reviews (P3, P4, P5). To allow his team members to interact with his code, P3 would "try to have like a dev branch where they can play with [the infrastructure] before it goes live. Sometimes, I'll have a Tableau page open, and then I can manipulate it live in the meeting." However, he noted the method of presentation depended on the audience—in the rare case of presenting to a Vice President, he would not create visualizations live to minimize room for error.

These participants communicated with their data science team members frequently outside of scheduled meetings. Communication was often spurred by major changes in code (P3, P4, P5), a request for information or help (P3, P4), or a wish to discuss new or interesting ideas (P4, P5). P4 communicated with his fellow data scientists to exchange knowledge or ask for help regarding the data.

> *"I confer with a teammate the most when it comes to talking about data... for example [I would ask], oh, can you find me this feature data? Where's this located? Or can you tell me if you have a PM in mind of a different team who might be able to ping me and something like that? So it's really about the data... the data is really what we all have in common."* (P4)

Code and data were hosted and shared on a common cloud platform shared across the organization (P3, P4, P5). Individual participants felt free to use any of their favorite tools to engage in iterative analysis (such as Jupyter Notebooks (P5)) but processed the code to meet organizational standards (P4, P5). Participants summarized their findings and insights in the form of a slideshow presentation or document (P5), which

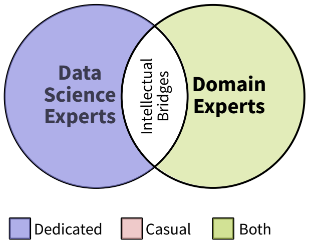

Fig. 3: TEAM C – THE TWO CITIES. These teams, who are often large, contain experts from both data science and the problem domain, but with no real bridges between them.

would get reworked by their supervisor or directly relayed to the stakeholders (P4, P5).

None of the participants in this group mentioned experiencing friction in communicating with their team members, potentially because all of them had expertise in data science and thus did not feel a need to change their tools or processes significantly to support collaboration. For example, since his team members were familiar with Git, P3 could easily add a development branch to receive feedback on his infrastructure rather than using a separate tool. It follows that tools like Git and shared repositories seemed to be most useful in aiding effective communication, particularly to enable a shared understanding of the current state of a project.

3) Team Structure C – The Two Cities: Yet other teams involved both domain experts and data scientists, but no intellectual bridges (P6-P15). As a result, there was very little or strained collaboration between the two parties: essentially a "tale of two cities" of data scientists and domain experts.

These participants tended to work in government agencies or mid-sized companies with designated data science teams. They collaborated often with other data scientists on their team and collaborated with both casual and dedicated domain experts. Notably, a few participants (P8, P13, P15) collaborated as or more closely with domain experts than with their fellow data scientists. These participants had a large and diverse group of dedicated domain expert collaborators, and they communicated frequently to exchange pertinent domain knowledge.

Participants regularly communicated progress and insights with their team of data scientists via presentations (P10, P12, P13, P14), shared documentation (P8, P11, P12, P13), project management tools (P7, P8, P9), messaging platforms (P6-P15), or meetings (P6-P15). All but one participant (P13) presented their process via code shared on online repositories such as GitHub, GitLab, Box, or CNVRG. Data science teammates were frequently updated on process details via code presentations, or data visualizations generated via integrated development environments like Jupyter Notebook, RStudio, or Visual Studio. This type of collaborative relationship was maintained mostly when required—to aid validation, evaluation, and process improvement (P13).

Participants in this team structure directly communicated with domain experts via presentations (P8, P10, P12, P13, P15), shared documentation (P7, P8, P15), dashboards (P7, P9), visualizations (P6, P7, P10, P12), and screen-sharing computational notebooks in meetings (P6, P7, P12, P13, P14). Between formal meetings, participants would often collaborate with domain experts on an as-needed basis, using tools like Slack, Microsoft Teams, or Skype to maintain a shared understanding and exchange domain knowledge. Similar to teams in 'The Token' structure, these teams tended to avoid directly sharing intermediate code or analysis and used tools familiar to the domain experts when it was necessary. P6 mentioned that the domain experts he worked with did not like using Git, and he thus directly emailed computational notebooks to minimize friction in tool differences.

> *"I've had to adapt to how they work... instead of telling them, hey, clone this repo, open up, you know, go to this branch and open up this file, they get very uncomfortable about that, they just want you to send them a file."* (P6)

Participants in this group faced the most friction in maintaining effective communicative relationships in their interdisciplinary teams. P7 and P13 indicated that establishing a common language was one of the most difficult parts of their collaborative efforts, likely due to the lack of an intellectual bridge in their teams.

> *"I think most of the communication issues are really talking to stakeholders that are not as technical, but like it's important that they have to understand what I'm trying to say and it's a little for hard for them, for me to understand what they're trying to say, so I think the non-technical to technical communication is most important... within the technical space it's actually easier."* (P13)

P7 and P8 valued computational notebooks for quick iteration but recognized that they do not lend themselves to collaboration and documentation. P8 noted "sharing notebooks [with other data scientists], often is not the best way to persist code in a meaningful manner... for sharing those results with non-technical folks, we will often resort to wikis because they're easier to read and much easier to manage." While communicating with domain experts increases the number of tools data scientists need to use, outcomes such as more interpretable documentation benefit everyone on the team.

V. DISCUSSION

In this section, we highlight opportunities to develop new data science tools and features in light of our findings. We also discuss study limitations and directions for future work.

A. Supporting Asynchronous Communication

Interdisciplinary teams rely on asynchronous and synchronous mechanisms to share findings and build consensus among team members. However, our findings reveal that data science tools—if they even directly support collaboration in the first place—are designed primarily for synchronous communication, limiting their ability to support effective collaboration among diverse teams. We observed that several participants (P1, P2, P6, P7, P12, P13, P14) discussed information stored in tools such as Jupyter Notebook or RStudio through synchronous meetings, despite these tools having the capability to be used asynchronously. This is especially problematic for teams distributed across space and time.

For example, our participants expressed concern over sharing intermediate code and notebooks with non-data scientists outside of meetings, since these artifacts are rarely designed to be interpreted and manipulated by non-expert users. Instead, participants preferred to share non-interactive HTML reports exported from notebooks and forgo using Git to share analyses. Based on these findings, we encourage the community to *consider how data science tools can be used by non-data scientists in an asynchronous context*. One opportunity we observe is to design the interface to match user expertise. For example, notebook environments could have "exploration" versus "explanation" modes based on whether the notebook will be used to advance a data exploration task or present results to non-data-scientists [31], [32]. When configuring the explanation mode, data science experts could designate certain parameters to be varied such that a domain expert can substitute different values and see how the analysis results change, without needing to focus on how the rest of the notebook works. This would be similar to the limited interactive functionality exposed by *interactive articles* [44] such as Distill[3] and Idyll [7].

B. Preserving Synchronous Contexts Outside of Meetings

Even when they used data science tools effectively during meetings in synchronous collaboration, our participants struggled to extract and preserve valuable insights from these sessions. In other words, current data science tools are effective for building shared context during synchronous communication but ineffective for carrying this context forward into the next iteration of a data science project. Meetings are ethereal and fleeting, and the group dynamics are difficult to capture.

Our participants identified two specific aspects of synchronous context that they seek to preserve: (1) annotations created by other team members using presentation tools, and (2) comments and notes made about specific parts of a notebook or other artifact. We believe the fundamental issue is *the inability to hand off context between tools*, such as mapping whiteboard annotations on a presentation screen back to the original computational notebook environment being annotated. It would be interesting to design features to enable these handoffs such as by designing notebook environments to include in-situ presentation features or even enabling direct integration between presentation tools and notebook environments.

A secondary issue we observe is *a lack of structure around managing annotations and insights*. We observe ample API coverage and built-in support for managing data, models, and code within data science tools. In contrast, we observe little if any structure provided for managing team requirements, comments, and insights within data science tools. This issue is exacerbated in a collaborative context, where team members may contribute new information in informal ways such as through presentation annotations. While this issue has been studied specifically in visualization contexts [16], we have yet to see viable solutions in the broader context of data science.

C. Limitations

While our study has unearthed new findings about the role of communication in real-world data science teams, our work is still limited in many ways. For one thing, our perspective for all of these interviews was centered on the data scientist rather than the domain experts on the teams, thus potentially tinting our findings with a data science bias. Furthermore, we only ever spoke to a single member of each team; involving additional members, potentially in a group discussion, may have provided additional insight into team communication patterns. Our rationale for this choice is that our study is conducted through the lens of data science as well as the tools data scientists surround themselves with.

As with any qualitative study, our work is also limited in scope by the sheer volume of interview transcripts from 15 participants. At the same time, 15 data science teams is not a big number, and it is likely that our unique sample is not fully representative of data science teams everywhere. While we took care to select participants from a range of academic, industrial, and governmental backgrounds as well as team sizes and problem domains, there is only so much diversity that can be found in a pool of 15 participants.

VI. Conclusion and Future Work

We have presented results from an interview study involving 15 professional data scientists drawn from multidisciplinary teams tasked with data science projects within academia, industry, and government. Our study focused on the role of tools in facilitating (or hindering) communication between different members and contingents in such teams. Our findings revolve mostly around the impact of team structure on team communication as well as team communication dysfunction. Drawing on these findings, we identify a few design inspirations for the next generation of data science tools, including improved support for asynchronous collaboration, in-situ presentation and provenance, and embedded communication channels.

We see several avenues for future work. Implementing our design recommendations in future data science tools would elevate team-based data science beyond the current two extremes of collaboration either through GitHub or through collaborative editing. If our work has shown the importance of "intellectual bridges" connecting the two cities of data science and problem domain, then a future priority should be to investigate how to build tool interventions that can serve in this stead in lieu of a human bridge. For example, applying Star and Griesemer's concept of boundary objects [34] as a framework to design tools that simultaneously serve both domain experts and data science experts. And finally, we see a need for expanding and generalizing the interview study protocol that we followed in this paper. In particular, understanding the full picture of effective communication requires including all members of a team, not just the data scientists.

[3] http://distill.pub/

Acknowledgments

We thank the anonymous data scientists who gracefully agreed to take part in our interview study.

References

[1] S. Alspaugh, N. Zokaei, A. Liu, C. Jin, and M. A. Hearst. Futzing and moseying: Interviews with professional data analysts on exploration practices. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):22–31, 2018.

[2] N. Boukhelifa, A. Bezerianos, I. C. Trelea, N. M. Perrot, and E. Lutton. An exploratory study on visual exploration of model simulations by multiple types of experts. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–14, New York, NY, USA, 2019. ACM.

[3] M. Brandt, C. J. Tucker, A. Kariryaa, K. Rasmussen, C. Abel, J. Small, J. Chave, L. V. Rasmussen, P. Hiernaux, A. A. Diouf, et al. An unexpectedly large count of trees in the West African Sahara and Sahel. *Nature*, 587(7832):78–82, 2020.

[4] A. Camisetty, C. Chandurkar, M. Sun, and D. Koop. Enhancing web-based analytics applications through provenance. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):131–141, 2018.

[5] H. Chung, S. Yang, N. Massjouni, C. Andrews, R. Kanna, and C. North. VizCept: Supporting synchronous collaboration for constructing visualizations in intelligence analysis. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 107–114, Los Alamitos, CA, USA, 2010. IEEE.

[6] H. H. Clark and S. E. Brennan. Grounding in communication. In L. B. Resnick, J. M. Levine, and S. D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, Washington, DC, USA, 1991.

[7] M. Conlen and J. Heer. Idyll: A markup language for authoring and publishing interactive articles on the web. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, pages 977–989, New York, NY, USA, 2018. ACM.

[8] K. Crowston, J. S. Saltz, A. Rezgui, Y. Hegde, and S. You. MIDST: A system to support stigmergic coordination in data-science teams. In *Conference Companion Publication of the ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 5–8, 2019.

[9] J. N. Cummings and S. Kiesler. Collaborative research across disciplinary and organizational boundaries. *Social Studies of Science*, 35(5):703–722, 2005.

[10] P. N. Edwards, M. S. Mayernik, A. L. Batcheller, G. C. Bowker, and C. L. Borgman. Science friction: Data, metadata, and collaboration. *Social Studies of Science*, 41(5):667–690, 2011.

[11] B. Friedman, P. H. Kahn, A. Borning, and A. Huldtgren. Value sensitive design and information systems. *Early Engagement and New Technologies: Opening up the Laboratory*, pages 55–95, 2013.

[12] A. Ganji, M. Orand, and D. W. McDonald. Ease on down the code: Complex collaborative qualitative coding simplified with 'Code Wizard'. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–24, 2018.

[13] A. Head, F. Hohman, T. Barik, S. M. Drucker, and R. DeLine. Managing messes in computational notebooks. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 270:1–270:12, New York, NY, USA, 2019. ACM.

[14] J. Heer and M. Agrawala. Design considerations for collaborative visual analytics. *Information visualization*, 7(1):49–62, 2008.

[15] P. Isenberg, N. Elmqvist, J. Scholtz, D. Cernea, K.-L. Ma, and H. Hagen. Collaborative visualization: Definition, challenges, and research agenda. *Information Visualization*, 10(4):310–326, 2011.

[16] S. Kandel, A. Paepcke, J. M. Hellerstein, and J. Heer. Enterprise data analysis and visualization: An interview study. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2917–2926, 2012.

[17] D. Kang, T. Ho, N. Marquardt, B. Mutlu, and A. Bianchi. Toonnote: Improving communication in computational notebooks using interactive data comics. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–14, New York, NY, USA, 2021. ACM.

[18] Y.-a. Kang and J. Stasko. Characterizing the intelligence analysis process: Informing visual analytics design through a longitudinal field study. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, pages 21–30, Los Alamitos, CA, USA, 2011. IEEE.

[19] M. B. Kery, B. E. John, P. O'Flaherty, A. Horvath, and B. A. Myers. Towards effective foraging by data scientists to find past analysis choices. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–13, New York, NY, USA, 2019. ACM.

[20] M. B. Kery and B. A. Myers. Exploring exploratory programming. In *Proceedings of the IEEE Symposium on Visual Languages and Human-Centric Computing*, pages 25–29, Los Alamitos, CA, USA, 2017. IEEE Computer Society.

[21] M. B. Kery, M. Radensky, M. Arya, B. E. John, and B. A. Myers. The story in the notebook: Exploratory data science using a literate programming tool. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 174:1–174:11, New York, NY, USA, 2018. ACM.

[22] D. E. Knuth. Literate programming. *The Computer Journal*, 27(2):97–111, 1984.

[23] L. Koesten, E. Kacprzak, J. Tennison, and E. Simperl. Collaborative practices with structured data: Do tools support what users need? In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 100:1–100:14, New York, NY, USA, 2019. ACM.

[24] S. Kross and P. Guo. Orienting, framing, bridging, magic, and counseling: How data scientists navigate the outer loop of client collaborations in industry and academia. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–28, 2021.

[25] Y. Liu, T. Althoff, and J. Heer. Paths explored, paths omitted, paths obscured: Decision points & selective reporting in end-to-end data analysis. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–14, New York, NY, USA, 2020. ACM.

[26] Y. Mao, D. Wang, M. Muller, K. R. Varshney, I. Baldini, C. Dugan, and A. Mojsilović. How data scientists work together with domain experts in scientific collaborations: To find the right answer or to ask the right question? *Proceedings of the ACM on Human-Computer Interaction*, 3(GROUP):1–23, 2019.

[27] M. Muller, I. Lange, D. Wang, D. Piorkowski, J. Tsay, Q. V. Liao, C. Dugan, and T. Erickson. How data science workers work with data: Discovery, capture, curation, design, creation. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–15, New York, NY, USA, 2019. ACM.

[28] G. M. Olson and J. S. Olson. Distance matters. *Human-Computer Interaction*, 15(2-3):139–178, 2000.

[29] R. Y. Pang, R. Wang, J. Nelson, and L. Battle. How do data science workers communicate intermediate results? In *Proceedings of the IEEE Symposium on Visualization in Data Science*, pages 46–54, Los Alamitos, CA, USA, 2022. IEEE.

[30] S. Passi and S. J. Jackson. Trust in data science: collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–28, 2018.

[31] D. Raghunandan, A. Roy, S. Shi, N. Elmqvist, and L. Battle. Code code evolution: Understanding how people change data science notebooks over time. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[32] A. Rule, A. Tabard, and J. D. Hollan. Exploration and explanation in computational notebooks. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–12, New York, NY, USA, 2018. ACM.

[33] T. Schleyer, B. S. Butler, M. Song, and H. Spallek. Conceptualizing and advancing research networking systems. *ACM Transactions on Computer-Human Interaction*, 19(1):1–26, 2012.

[34] S. L. Star and J. R. Griesemer. Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in berkeley's museum of vertebrate zoology, 1907-39. *Social studies of science*, 19(3):387–420, 1989.

[35] S. Teasley, L. Covi, M. S. Krishnan, and J. S. Olson. How does radical collocation help a team succeed? In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pages 339–346, New York, NY, USA, 2000. ACM.

[36] A. Y. Wang, A. Mittal, C. Brooks, and S. Oney. How data scientists use computational notebooks for real-time collaboration. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–30, 2019.

[37] A. Y. Wang, Z. Wu, C. Brooks, and S. Oney. Callisto: Capturing the "why" by connecting conversations with computational narratives. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–13, New York, NY, USA, 2020. ACM.

[38] D. Wang, Q. V. Liao, Y. Zhang, U. Khurana, H. Samulowitz, S. Park, M. Muller, and L. Amini. How much automation does a data scientist want? *arXiv preprint arXiv:2101.03970*, 2021.

[39] F. Wang, X. Liu, O. Liu, A. Neshati, T. Ma, M. Zhu, and J. Zhao. Slide4N: Creating presentation slides from computational notebooks with human-ai collaboration. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–18, New York, NY, USA, 2023. ACM.

[40] N. Weinman, S. M. Drucker, T. Barik, and R. DeLine. Fork it: Supporting stateful alternatives in computational notebooks. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–12, New York, NY, USA, 2021. ACM.

[41] Y. L. Wong, K. Madhavan, and N. Elmqvist. Towards characterizing domain experts as a user group. In *Proceedings of the IEEE Workshop on Evaluation and Beyond - Methodological Approaches for Visualization*, 2018.

[42] A. X. Zhang, M. Muller, and D. Wang. How do data science workers collaborate? roles, workflows, and tools. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW):1–23, 2020.

[43] J. Zhao, M. Glueck, P. Isenberg, F. Chevalier, and A. Khan. Supporting handoff in asynchronous collaborative sensemaking using knowledge-transfer graphs. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):340–350, 2017.

[44] Z. Zhao and N. Elmqvist. The stories we tell about data: Surveying data-driven storytelling using visualization. *IEEE Computer Graphics and Applications*, 2023.

[45] C. L. Zitnick, L. Chanussot, A. Das, S. Goyal, J. Heras-Domingo, C. Ho, W. Hu, T. Lavril, A. Palizhati, M. Riviere, et al. An introduction to electrocatalyst design using machine learning for renewable energy storage. *arXiv preprint arXiv:2010.09435*, 2020.