MIT Emerging Talent

January - December 2025

Certificate in **Computer** and **Data Science**

Collaborative Data Science Projects (CDSP)

Welcome to the collaborative Data Science projects for Emerging Talent! With this ***required*** **collaborative project** you will explore how data science methodologies interact with domain expertise to create new understanding, and how this understanding can be translated into real-world solutions.

To focus on skills that cannot be learned individually, workshops and deliverable assessments will emphasize non-technical aspects of collaborative Data Science such as: **research questions, team dynamics, task division, appropriate skepticism, and communication with non-technical audiences.** Our workshops will focus on higher-level, more human skills; you must take ownership of your own technical learning. We will be here to support you with any questions and to help you find resources, but you must take initiative to identify the specific skills you need to learn and to ask for help when necessary. A good starting point is The Art of Data Science, a great resource for navigating the CDSP without getting lost in code.

The project will be broken into 6 milestones, each focusing on a different aspect of collaborative Data Science:

- **0 - Cross-Cultural Collaboration:** The zeroth milestone covers frameworks of cross-cultural communication, design and innovation. In this milestone you will meet your group mates and prepare your group repository.
- **1 - Problem Identification:** In the first milestone you will do an initial domain study and frame an actionable research question in your project domain, and within your groups' constraints.
- **2 - Data Collection:** In the second milestone you will decide how to model your problem domain in data and what data is relevant to your research question. Finally, you will collect, clean, document and host a data set to study.
- **3 - Data Analysis:** In the third milestone you will focus on finding the appropriate analysis techniques for the question and data you have. You may find that a simple bar graph or statistical analysis is enough to answer your questions!
- **4 - Communicating Results:** In the fourth milestone of the project you will translate your findings into a clear message and prepare a communication strategy to reach a well-defined target audience capable of acting on your advice.
- **5 - Final Presentation:** In the final milestone, you will prepare a presentation for your friends, family, Emerging Talent partners and potential employers. Your presentation will cover your communication artifact, your project's evolution, what you learned, and next steps you imagine for your work. For some groups this will be a great end to their projects, for others it may be the first pitch for their new initiative.

Practical Information

Participation in the collaborative Data Science Projects is a *required element* of the Experiential Learning Opportunities and is a requirement to be able to graduate from the Certificate program. If you are not able to fully participate in the collaborative projects for reasons out of your control, you need to email emergingtalent@mit.edu to make your case, and we will try to find a solution together.

Important Dates

Group Formation & Project Launch:

1. **Submit the Group Formation Survey by** May 15, 2025
2. **Groups will be announced on** May 22, 2025
3. **Project Kickoff:** May 27, 2025
    a. Weekly Workshops will be held since then until end of August.

Deliverable Deadlines:

- **Milestone 0, Cross-Cultural Collaboration:** June 02
- **Milestone 1, Problem Identification:** June 16
- **Milestone 2, Data Collection:** June 30
- **Milestone 3, Data Analysis:** July 21
- **Milestone 4, Communicating Results:** August 11
- **Milestone 5, Final Presentation:** August 25

Final Presentation Event: *TBD*

Expected Participation

Participation in the projects is required for graduation. You are expected to commit a minimum of 8 hours per week to your group project. Most of this time is flexible, asynchronous work that you can schedule within your group. Below is an estimate of how much time you will spend on different aspects of your project each week, actual times will vary week-by-week:

- **1.5+ hours of workshop (1-2 per week).** Organized by the ET team covering concepts related to your projects. Attendance for each group is required, **we expect *at least 2 members* of each group to be present in each workshop**. We understand that sometimes there are conflicts out of your control. If you cannot attend a workshop please let us know ahead of time and coordinate with your group so they can fill you in afterwards.
- **1.5 hours of team meetings.** Your group will decide the best way to divide this time across the week, and the best use of your meetings. Don't forget to plan agendas ahead of time!
- **5+ hours of asynchronous work.** This includes completing your tasks, communicating with team members, and self-study.

Project Topics

We will form groups for you based on your domains of interest. As a group you will identify and tackle a *new* problem in your chosen domain. **Some possible domains include Economic Inclusion, Education or Healthcare.** To support this process we will provide workshops on ideation and innovation, as well as two weeks dedicated to researching and defining a problem. Approaching any open-ended project with a solution already in mind is not a recipe for success. Approaching a Data Science project with a Software Development solution in mind is even less productive.

Group Formation

>> Project Groups, 2025 (coming soon) <<

We will form new groups based on your responses to the Sign up Form which is due May 15. When forming groups our first consideration will be your domains of interest; we want you all to be working on a project you are passionate about. Next we'll consider your technical and professional experiences; we want to make balanced groups so everyone can take turns leading and learning. We will consider demographics like gender, language and nationality to guarantee balanced groups; succeeding in diverse teams is a skill that can be learned and practiced. Finally we'll consider your availability and time zones; we want to make sure each group has some overlap to help schedule meetings.

You do not need to be an expert in a domain to choose it for your project, you just need to be curious and ready to learn. One goal of these projects is for you to practice exploring a field and identifying new challenges. It can even help to be a newcomer!

Milestone Workshops

We will host a workshop on the first Monday of the first 5 project milestones. These workshops will explain what is expected in the coming milestone and will provide you with a set of mental models and questions you can use to guide your work towards the next deliverables:

- **Cross-Cultural Collaboration:** *May 27*
- **A "Big Picture" of Data Science:** *June 03*
- **Domain and Data:** *June 17*
- **Learning from Data:** *July 01*
- **Communicating Results** *July 22*

In addition to these 5 workshops, we will host workshops covering skills or concepts helpful in the execution of your deliverables. These workshops will include topics such as Innovation & Design Processes, and Python Notebooks. We will share the dates and titles of these workshops through the Emerging Talent events calendar.

Self-Study Plans

You are responsible for your personal learning throughout these projects! You are accountable to your group and yourself for identifying which skills you need to learn and for learning them well enough to complete your tasks. We recommend you create a clear study plan with learning objectives, selected resources, and deadlines linked to project tasks or milestones.

While not required, it can help to share a link to your study plan so your group members, technical mentors, and the Emerging Talent team can support you. Support might include suggesting changes to your learning objectives, sharing links to great resources, or organizing shared study sessions.

Submitting Deliverables

Over the course of this project, your group will build a GitHub repository containing not only your data analysis scripts, but a record of the process you followed to reach your final presentation including: ideation, research, data collection, task division, code review, and research summary. This all-in-one package will be a strong addition to your professional portfolio, demonstrating your professional skills beyond programming.

At the end of each milestone each group will submit a form linking to your repository and a few questions about your work. We will check your deliverables by directly reviewing your group's repository, so be sure to keep it well organized and documented with a clear README! If we cannot easily find your deliverables we will mark them as missing. Depending on the format of a deliverable, it does not need to be stored directly in your repository but we still ask you to include a link in your repository. For example, if your deliverable is a slide show, we will ask you to link to your slide show from the repo. This may sound like extra work, but the goal is to make sure a future recruiter, collaborator or potential client can access everything about your project in one place.

Support and Further Questions

If you have questions about what to expect during the project, or need extra support at any time during the project, please reach out to emergingtaleng@mit.edu. If we can't answer your questions directly we'll be happy to schedule a call and reach a solution together!

Project Assessment

You will be based more on the *processes* you follow than the *products* you deliver. This means your group communications, your workflow discipline, your project documentation and your milestone retrospectives will have ***substantially*** more weight than the complexity of your data analysis. Below are some criteria we will use for your assessment:

Group Assessment

- Deliverables are on time and complete
- GitHub repositories are well-maintained and well-documented
- Communicating proactively with ET team
- Other evidence of successful collaboration (github activity, slack history, …)
- Technical decisions are inclusive of all group members' level and experience
- Final presentation is an accurate reflection of your group's experience
- Workshop attendance planning - Was your group's attendance communicated *before* each workshop? Did you communicate any last-minute changes to this plan?

Individual Assessment

- Owned tasks are well-managed on team project board
- Individual study plan was proactive and well-organized
- Evidence of effective communication and conflict resolution
- Workshop attendance planning - Did you either attend each workshop or communicate that you would not be joining?

Successfully completing CDSP is a required component for graduating with the MIT Emerging Talent Certificate in Computer and Data Science.

Learning Support and Resources

Your group will be responsible for keeping a healthy dynamic, for scoping your project, submitting deliverables on time, and for getting the support you need when you need it. We will always be available to support you in any way possible, but will not be looking over your shoulder to know what you need if you don't ask. Your group will learn the most and have the highest chance of success if you take full advantage of these supports and resources we will make available.

Project Guidance

You will have access to project guides who can help you with everything related to project management. This includes: group dynamics, task management, project scoping, defining constraints, and planning meetings. Before reaching out to a project guide for assistance you must have a list of specific questions you would like help with, including how you have already tried to resolve the problems on your own.

Domain Expertise

We will have domain experts available for consultation at key points in the project. Domain experts can help you with any non-technical questions about your project's scope, help guide you when choosing research questions or selecting data sources, and can help at the end of the project when you are exploring how to act on your research findings.

Technical Mentorship

When you have technical questions related to your individual tasks, or when you need to make an important technical decision as a group, you can reach out to one of our technical mentors. They can help with anything related to programming, data management, or data analysis. You can also reach out to a technical mentor individually if you need help with your individual study plan or have been stuck for too long on one problem.

Learner Resources

Finally, we will maintain a collection of suggested resources for your independent study. Because we will not offer technical workshops throughout the project, it is your responsibility to identify the skills you need to learn and to structure your own study plans. You can always reach out to a technical mentor for help deciding what you should study at any point in the project.

Project Milestones

Your project has 6 milestones, each milestone has its own learning objectives and deliverables.

May 27 - June 2	*June 3 - June 16*	*June 17 - June 30*	*July 1 - July 21*	*July 22 - August 11*	*August 12 - August 24*
Cross-Cultural Collaboration	Problem Identification	Data Collection	Data Analysis	Communicating Results	Final Presentation

Milestone 0: Cross-Cultural Collaboration

May 27 - June 2

Workshops in this milestone of the project will introduce the foundations of cross-cultural communication, design and innovation methodologies.

Workshop on May 27: **Cross-Cultural Collaboration**

Learning objectives include:	**Deliverables** *due June 2:*
● Frameworks and methods for group communication and collaboration. ● Understanding the value of discussing and defining project constraints. ● Innovation and design processes.	1. A repository set up with: a project board, branch protections, pull request template, and 6 milestones (named after the CDSP milestones) 2. Collaboration documents: 1) Group norms 2) Constraints 3) Communication plan 4) Learning goals 3. Write contributor guidelines in CONTRIBUTING.md 4. Meeting agendas and minutes. 5. A retrospective for this milestone. 6. A labeled Git tag for this milestone created before the deadline - we will review your deliverables based on the tagged commit. 7. Complete the milestone survey

Milestone 1: Problem Identification

June 3 - June 16

The Data Science part of the project begins with your group clearly stating a problem you have encountered in your lives. To frame your problem as an actionable research question within your constraints, you will complete an initial domain study and identify some key stakeholders with the ability to act in this domain.

Workshop on June 3: **A "Big Picture" of Data Science**

Learning objectives include:	**Deliverables** *due June 16*
● Understand how to balance divergent and convergent thinking. ● Appreciate the importance of domain expertise in data science. ● Identify important problems in a domain that are accessible within your constraints. ● Learn how to state a clear research question around your problem that can be answered using data science.	1. * A problem-statement based on your personal experiences 2. A thorough background review of your research domain in the `0_domain_research` folder of your repository 3. * A summary of your group's understanding of the problem domain. (Bonus points for using systems thinking!). 4. * An actionable research question informed by the realities of your problem domain and your constraints. 5. Maintain your planning documents (group norms, learning goals, constraints, a communication plan, …) 6. A retrospective for this milestone. 7. A labeled Git tag for this milestone created before the deadline - we will review your deliverables based on the tagged commit. 8. Complete the milestone survey. * Written in your repository's README.md.

Milestone 2: Data Collection

June 17 - June 30

After your group has clearly defined your research objectives, you will decide how to model your problem domain with data taking into account considerations such as: the advice of domain experts, what data is available & how it was collected, which data features are most useful, and what data is not available.

Workshop on June 17: **Domain and Data**

Learning objectives include:	**Deliverables** *due June 30*
● Understand the strengths and weaknesses of modeling the world using data. ● Learn to study a domain to understand which data is interesting and relevant for a given problem ● Investigate the data available in a domain to understand what is available, what is missing, what you could realistically collect yourself, and possible flaws in the data. ● Collect, clean, organize and document a data set so it is easy to study.	1. A non-technical explanation of how you chose to model your domain, and possible flaws in this approach, in your README. (visuals are helpful!) 2. Documentation for your data set describing where it came from, how it's structured, possible flaws, and how someone can recreate it. Think broadly: qualitative data is still data! 3. All of your data collection and cleaning scripts so someone can replicate your final data set given your raw data sources. Includes scripts for separating your data set into training and validation data, if applicable. 4. A public hosting of your prepared data set. This can be directly in your repository, or a link to i. 5. A labeled Git tag for this milestone created before the deadline - we will review your deliverables based on the tagged commit. 6. Complete the milestone survey. 7. A retrospective for this milestone (group & individual)

Milestone 3: Data Analysis

July 1 - July 21

You are not expected to use the most advanced algorithms or machine learning techniques, if anything we encourage you not to. Instead, you should focus on *matching* your analysis technique to the questions you're asking, the data you have available, and your group's constraints. You may find that a bar graph or basic analysis is enough to answer your questions!

Workshop on July 1: **Learning from Data**

Learning objectives include:	**Deliverables** *due July 21*
● Understand the limitations of data analysis, and how to ask questions that data analysis *can* answer. ● Use analysis techniques that are appropriate for your question, your domain, your data and constraints. ● Understand how to identify, address, and communicate sources of uncertainty in your data analysis. ● Being prepared to accept undesirable results, or even no results.	1. A non-technical explanation of your findings, including your levels of certainty and possible sources of error in your analysis. This should include visualizations. 2. A technical description of your analysis & results including explanations of why you used the techniques you did, possible flaws in your analysis, and possible alternative approaches. 3. All of the scripts and documentation necessary for someone to replicate your analysis using your data set. 4. Complete the milestone survey 5. A labeled Git tag for this milestone created before the deadline - we will review your deliverables based on the tagged commit. 6. A retrospective for this milestone (group & individual)

PS. Always ask yourself; "How could I be wrong?"

Milestone 4: Communicating Results

July 22 - August 11

After analyzing their data, you will prepare a communication strategy to reach a well-defined target audience. You decide how to do this. It could be: simple actions an individual could take, a startup pitch, policy advice, an academic paper, or something we never could expect!

In this final milestone of the project you will explore challenges in science communication such as: constructively discussing uncertainty, choosing the appropriate medium of communication, translating statistical findings into actionable knowledge, and much more.

Workshop July 22: **Communicating Results**

Learning objectives include:	**Deliverables** *due August 11*
● Communicating research results to a non-technical audience. ● Conveying the limits of your results and any sources of uncertainty.. ● Translating new understanding of a problem into actionable knowledge. ● Tailoring your communication to fit a well-defined user/audience, accounting for their capabilities and constraints. ● Communicating your results and their implications to your user as quickly and effectively as possible.	1. A document describing your target audience, their capabilities and constraints, how you intend to reach them, what you would like them to learn, and how you hope they will act. (Personas can be helpful!) 2. A communication artifact to achieve the goals in your document. Let your imaginations run with this! It could be a website, a powerpoint, a printed leaflet, a WhatsApp campaign … anything, as long as you can justify why it's the best way to reach them. 3. Complete the milestone survey 4. A labeled Git tag for this milestone created before the deadline - we will review your deliverables based on the tagged commit. 5. A retrospective for this milestone (group & individual)

Milestone 5: Final Presentation Event

August 12 - August 25

To close the project, there will be a presentation session open to friends, family and ET partners where each group shares their final communication strategy, the process they went through to arrive at this, and what they learned along the way. For some groups this will be a great end to their projects, for others it may be the first pitch for their new initiative.

Learning objectives include:	**Deliverables** *due August 25*
● Summarizing your learnings into a concise and clear presentation. ● Honestly assessing your group's performance, the good and the bad. ● Individual presentation skills.	1. A 2.5-minute pitch of your research & communication strategy. (see workshop slides for more details) ○ Remember to apply the models of communication we covered in Communicating Results! 2. A link to your presentation in your repository README. 3. Complete the milestone survey. 4. A labeled Git tag for this milestone created before the deadline - we will review your deliverables based on the tagged commit. 5. A retrospective for this milestone (group & individual).