WORKERA

a deeplearning.ai company

The AI Human Capital Playbook

Companies everywhere are building AI teams, but it's still unclear what aspiring machine learning engineers, data scientists, and software engineers should focus on when applying for AI jobs. This report walks you through different types of organization, different roles within them, the tasks you'll work on, and the skills recruiters are looking for in each role. It is the result of two large-scale studies of the supply of and business demand for AI talent. We're here to provide mentorship and help you find a job that suits your skills, experience, and aspirations.

CONTENTS

This report is a work in progress and is being provided to the public for information purposes. Because it is a work in progress, there are parts that are either missing or will be progressively revised as our team learns more about the supply of and the demand for AI talent.

We welcome your comments and feedback. Please send any comments and/or questions to Kian Katanforoosh (kian@workera.ai).

- The Workera team

PART I

AI Organizations

We interviewed 100+ data science and machine learning leaders from companies such as Airbnb, Amazon, Earnin, Facebook, Google, Landing.ai, Lyft, Proofpoint, and Upstart to understand the roles, tasks, and skills that make up a corporate AI organization (this is opposed to an academic setting, where often the AI project development lifecycle differs).

Data Science vs. Machine Learning Organizations

Based on our research, we've identified two types of AI organizations: the data science organization and the machine learning organization.

- The data science organization focuses on making scientific decisions, and its business goal is usually to help businesses run more effectively. The output of the organization is often a set of actionable insights and its workflow includes (at a high-level) collecting data, analyzing them, and suggesting hypotheses/actions.

- The machine learning organization focuses on automating tasks, and the business goal is to decrease operational costs or to scale a product. Automation is often the output, and the workflow includes (at a high-level) collecting data, training models, and deploying them.

NOTES

- Companies can have both data science and machine learning organizations. Some companies have hybrid organizations working toward both the data science and machine learning organizations' business goals.
- AI organizations can be centralized or decentralized. Centralized AI organizations group AI scientists and engineers to support non-AI teams, while decentralized AI organizations are scattered in different business units throughout a company.

PART 2

The AI project development lifecycle: Tasks and Skills

From our research, most AI organizations' work divides into five tasks: data engineering, modeling, deployment, business analysis, and AI infrastructure. Together, these tasks make an AI project development lifecycle. Each task requires specific skills and can be the focus of multiple roles.

First, we'll discuss the differences between machine learning (ML) and data science (DS) projects development lifecycle. Then, we'll answer the following questions:

- What are the goals of each task?
- What skill set is necessary to perform well in a given task?
- Given individuals with different skill sets, who should focus on which task?

Overview of the AI Projects Development Lifecycle

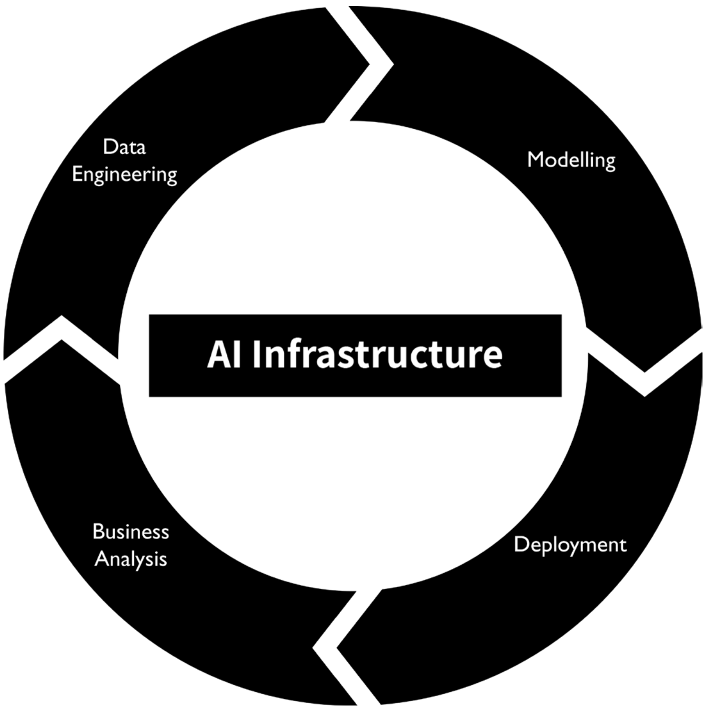
Here is a quick summary of ML and DS projects development lifecycle.

The ML project development lifecycle

An ML project starts with (i) data on which you (ii) fit models that will later be (iii) deployed into production. A deployed model (iv) needs to be monitored and its performance compared to the business goals. The (v) AI infrastructure is necessary to support all tasks described above (i, ii, iii and iv).

The DS project development lifecycle

A DS project starts with (i) data on which you can (ii) fit a model. These models and (iii) other data analyses help you make actionable business decisions. The (iv) AI infrastructure is necessary to support all tasks described above (i, ii and iii).



ML and DS projects can be carried-out by the same organization. Thus, we summarized both the ML and DS projects development lifecycle in a template called the AI project development lifecycle. A visual representation of the AI project development lifecycle is presented on the right column.

Now let's consider the tasks in an AI project one by one. We will illustrate each task with concrete examples, and identify the necessary technical skills to carry them out.

Data Engineering

Data engineering aims to provide the necessary data to achieve the modeling or business analysis task. Most of the time, data engineering is done using database query languages (such as SQL) and object-oriented programming languages (such as Python, C++ or Java). Big data tools (such as Hadoop or Hive) are also commonly used.

Data engineering work includes:

Subtask	Examples	Technical Skills Involved
Defining the data requirements	Creating a data model Defining the features of high-quality data Defining the covariates to be collected to achieve a desired functionality Providing feedback regarding the clarity and completeness of data requirements	Machine Learning Business Acumen Software Engineering
Collecting data	Setting up a Mechanical Turk Collecting data by manually taking images of cats Coding a javascript tracker on a website to collect user data Scraping the Web, and if necessary, synchronizing data located in different sources	Machine Learning Software Engineering
Labelling data	Drawing bounding boxes on images Building an automated labelling pipeline on a Mechanical Turk. Writing a labelling tutorial for workers Evaluate the labelling performance of workers Relabelling mislabelled data	Machine Learning
Inspecting and cleaning data	Replacing all non-usable structured data records by NaN using a Python library (e.g. pandas) Reformatting a data set (e.g. converting everything to jpeg and squaring all images) Cleaning a text dataset (e.g,. removing special characters)	Machine Learning Algorithmic Coding
Augmenting data	Writing a python script using skimage to rotate, warp, translate, or blur images Using test-time augmentation to reduce the variance of an algorithm Synthesizing speech by overlaying distinct audio signals	Machine Learning Algorithmic Coding
Moving data and building data pipelines	Writing a script to allow online learning for a model Designing an ETL system Writing a script to preprocess training data and send it as input to a model automatically Writing a script to record model predictions in a database	Domain-specific (e.g,. Data Query) languages
Querying data	Pulling data from a database	Domain-specific (e.g., Data Query) languages
Tracing data	Keeping track of the data sources Setting-up a data version control system	Software Engineering

Modeling

Modeling involves prototyping models to exploit patterns found in data to predict outcomes, identify business risks and opportunities, or determine cause-and-effect relationships. Modeling is usually programmed in Python, R, Matlab, C++, Java, etc. (though the dominant languages are Python and R). To understand modeling, it helps to have strong foundations in mathematics, data science, and machine learning. Deep learning skills are required by some organizations, depending on their product focus. Deep learning often empowers products leveraging computer vision, natural language processing, or speech recognition.

Modeling work includes:

Subtask	Examples	Technical Skills Involved
Training machine learning models	Using one of the following methods: Linear Regression, Logistic Regression, Decision Trees, Random Forest, XGBoost, Support Vector Machines, K-means, K-Nearest Neighbors, Neural Networks, Principal Component Analysis, Naive Bayes Classifier, Lasso/Ridge regression, etc.	Machine Learning Algorithmic Coding Mathematics Data Science
Fitting probabilistic or statistical models	Testing hypotheses via data experiments Applying a dimensionality reduction on a dataset to facilitate training of a model or gather insights	Data Science Algorithmic Coding Mathematics
Training deep learning models	Using deep learning for a domain-specific application such as object classification, detection, segmentation, text summarization, machine translation, speech recognition, etc. Extensively tuning hyperparameters involved in neural network optimization	Deep Learning Algorithmic Coding Mathematics Data Science
Accelerating training	Setting-up code to train your model on multiple machines in parallel	Domain-specific languages (e.g., CUDA) Algorithmic Coding
Defining evaluation metrics (usually also involves a data product manager)	Choosing F1-score to evaluate a machine learning model's performance on a classification task Implementing evaluation metrics such as accuracy, precision, recall, intersection over union, mean average precision (mAP), etc.	Machine Learning Algorithmic Coding Mathematics
Speeding-up prediction time	Applying techniques such as pruning, quantization or compression to reduce the memory requirements Running inference speed vs. accuracy experiments on a model	Machine Learning Algorithmic Coding
Iterate over the virtuous cycle of machine learning projects: Idea, Code, Experiment.	Translating a business problem into a machine learning problem. For instance, depending on the quality and quantity of accessible data, a better solution to the problem might come from an end-to-end or a pipeline network Experiencing the three-step cycle of ideating with your team, coding to set up experiments, analyzing results	Machine Learning Business Acumen
Searching hyperparameters	Organizing time effectively to run a maximum number of experiments in the shortest time period Setting-up hyperparameter search experiments using tools such as AutoML	Machine Learning Algorithmic Coding
Keeping up with the state-of-the-art	Reading research papers Watching conference lectures or attending conferences	Research Mathematics Data Science Machine Learning

Deployment

Deployment includes all of the activities that make a model available for use. Given a data stream (from the data engineering task) and a model (from the modeling task), individuals in charge of deployment will package and test models before pushing them to production environments. Deployment activities require the ability to write production code, including strong back-end engineering skills and understanding of cloud technologies.

Deployment work includes:

Subtask	Examples	Technical Skills Involved
Converting prototyped code into production code	Refactoring an entire repository's code Minimizing duplicate code Writing clean code to improve readability and consistency, for example, by following the PEP8 guidelines in Python	Software Engineering
Setting up a cloud environment to deploy the model	Mastering cloud tools and infrastructure provided by Amazon AWS, Microsoft Azure, Google Cloud, etc. Preparing files (usually model architecture and parameters) required for deployment	Software Engineering
Branching	Design a branching workflow. Using development, staging and production branches Participating in or leading code reviews	Software Engineering
Improving response times and saving bandwidth	Setting up load-balancing requirements with engineers in charge of AI Infrastructure	Software Engineering
Encrypting files that store model parameters, architecture and data	Understanding encryption at a high level and leveraging existing functions	Software Engineering
Building APIs for an application to use a model	Setting up HTTP RESTful API services to facilitate communications between software components Setting up authorization and authentication to access the API	Software Engineering
Retraining machine learning models (lifelong learning)	Monitoring changes in data distribution and staging model updates	Software Engineering Machine Learning
Fitting models on a resource-constrained device	Pruning or quantizing a model so it fits memory requirements Deploying a model on a mobile device using TensorFlow	Software Engineering Machine Learning

Business Analysis

Business analysis includes analytics, business activities related to communicating with clients and colleagues, thought leadership, and marketing. Working on business analysis requires a foundational understanding of mathematics and data science for analytics. It also requires strong communication skills and business acumen. Programming languages such as R or Python can be helpful, although many tasks can be carried out in a spreadsheet.

Business analysis work includes:

Subtask	Examples	Technical Skills Involved
Building data visualizations	Visualizing high-dimensional data in lower dimensions using methods such as PCA or t-SNE Building and presenting graphs produced using Tableau, ggplot or matplotlib Building visualizations in Javascript, HTML and CSS	Domain-specific programming languages Data Science Mathematics Business Acumen
Building dashboards for Business Intelligence	Writing a script that periodically notifies business leaders of trends in the data	Domain-specific programming
Presenting technical work to clients or colleagues	Preparing presentations (e.g., PowerPoints decks) Communicating effectively with team members Giving technical talks to present research outcomes	Communication Business Acumen
Translating statistics into actionable business insights	Making marketing decisions based on analysis of various sources	Data Science Business Acumen
Analyzing datasets	Plotting a correlation matrix to analyze covariates Computing statistical variables such as mean, variance, mode, etc. Segmenting customers into groups	Data Science Algorithmic Coding Mathematics
Running experiments to evaluate deployed models	Working with the deployment team to evaluate the business performance of a deployed model Helping the deployment team make decisions Translating model performance into business outcomes (e.g., revenue)	Data Science Algorithmic Coding
Running A/B tests	Optimizing web pages with A/B tests Evaluating systems in production	Data Science Algorithmic Coding Business Acumen

AI Infrastructure

AI infrastructure aims to facilitate data engineering, modeling, and deployment by building and maintaining reliable, fast, secure, and scalable software systems. Working on AI infrastructure requires strong and broad software engineering skills.

AI infrastructure work includes:

Subtask	Examples	Technical Skills Involved
Making software design decisions	Reducing latency by locating a model close to data	Software Engineering
Building distributed storage and database systems	Building the databases (SQL, NoSQL, MySQL, Cassandra, etc.) that will store data and facilitating access by other team members	Software Engineering Domain-specific (e.g,. Data Query) languages
Designing for scale	Adding GPU compute or storage as needed	Software Engineering
Maintaining software infrastructure	Managing software upgrades such as Python 2's end of life on 01/01/2020, and driving stability through automated monitoring and alerting	Software Engineering
Networking	Controlling access to all infrastructure elements	Software Engineering
Securing data and models	Building security features allowing for production deployments into regulated organizations, satisfying the needs for privacy and security	Software Engineering
Writing tests	Writing unit and functional tests for multiple components across tasks of the AI project lifecycle	Software Engineering
Carrying out various software tasks	Building a labeling software for a client, or key tools such as A/B testing frameworks or analysis environments	Software Engineering

PART 3

The roles of an AI team

There is no standard for roles in AI teams. Besides, the lack of information about the supply of AI talent makes it difficult for hiring managers to set reasonable job requirements that correlate with on-the-job performance. To bridge this gap, we assessed the skills of thousands of individuals aspiring to work in AI organizations and analyzed hundreds of job descriptions for AI roles.

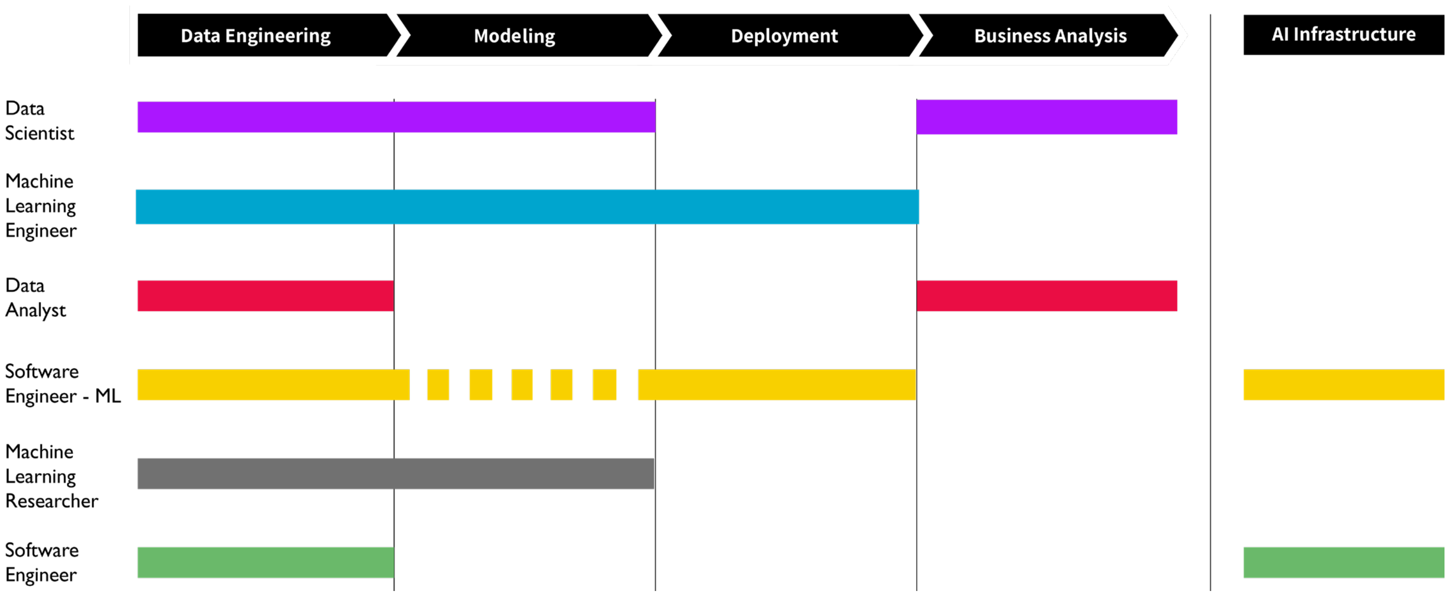
In part II, we defined the tasks carried out by AI teams. In this section, we'll introduce the different roles of an AI team, their skill sets, and the tasks they focus on. We hope that learning about these roles will help you find a career track and prioritize your learnings.

The Six Roles of an AI Team

We identified six technical roles with distinct skill sets and focus areas. Each of these roles contributes to a number of tasks in the AI development cycle.

All roles undertake (to a certain extent) the data engineering task. That's because data engineering is usually a necessary step to enhance modeling, deployment, and business analysis.

For each role, we list the tasks it may carry out and the skills necessary to achieve those tasks.

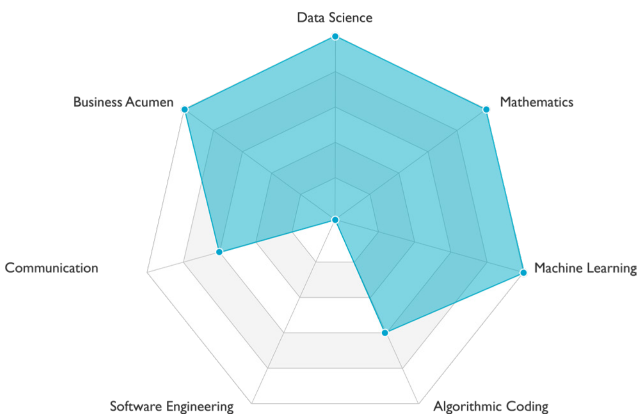


Note:

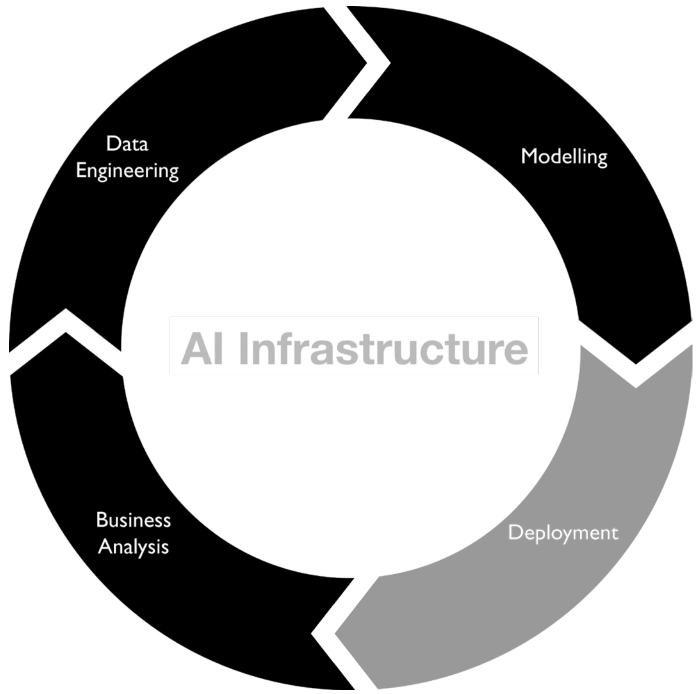
- The dotted line indicates a less signifant involvement with the task at hand. A Software Engineer - ML uses out-of-the-box methods to carry out the modeling task while an MLE, MLR or DS is able to customize models.

Data Scientist

SKILL PROFILE



TASKS



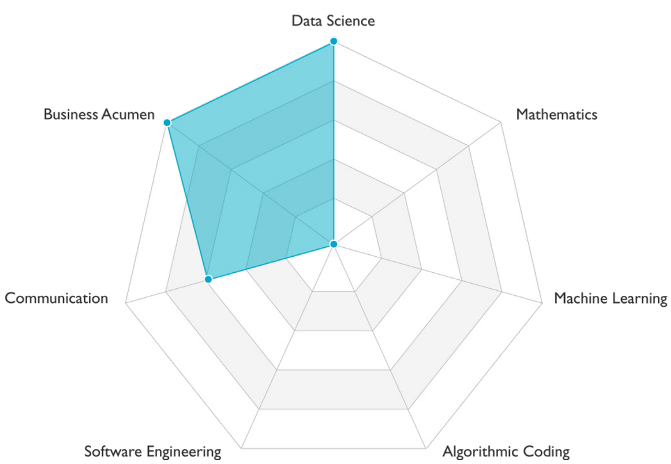
TOOLS DATA SCIENTISTS USE

- Modeling in Python using packages such as numpy, scikit-learn, TensorFlow, PyTorch, etc.
- Data Engineering in Python and/or SQL (or other domain-specific query languages).
- Business Analysis in Python, R, other domain-specific tools such as Tableau or Excel, and presentation software applications such as PowerPoint or Keynote.
- Collaboration and Workflow using a version control system (e.g., Git, Subversion, Mercurial, etc.) along with a Command Line Interface (CLI) (e.g., UNIX) and an Integrated Development Environment (IDE) (e.g., Jupyter Notebook, Sublime, etc.).

- Communication skills are usually required, but the level depends on the team.
- Terminology: Companies may refer to this position as data scientist, data analyst, machine learning engineer, research scientist, statistician, quantitative analyst, full-stack data scientist, and other titles.

Data Analyst

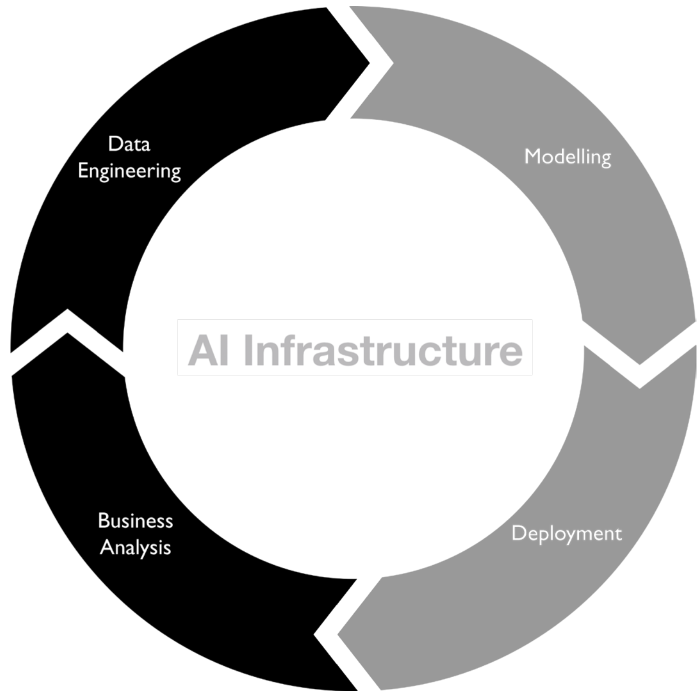
SKILL PROFILE



TOOLS DATA ANALYSTS USE

- Data Engineering in Python and/or SQL (or other domain-specific query languages)
- Business Analysis in Python, R, other domain-specific tools such as Tableau or Excel, presentation software applications such as PowerPoint or Keynote, and external software services for A/B testing

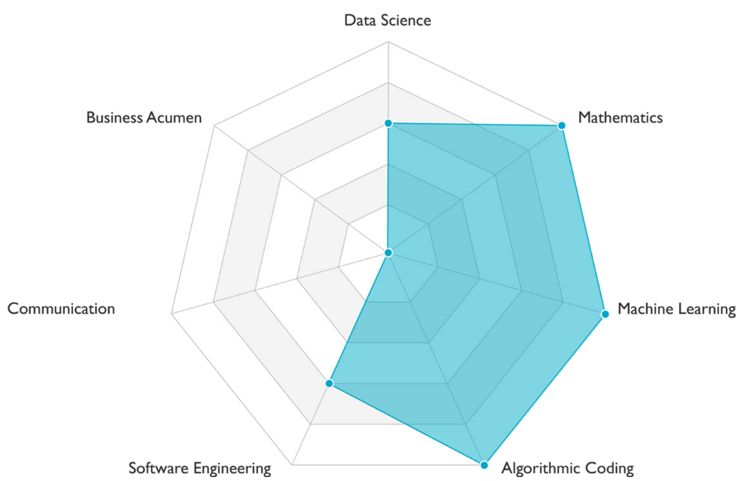
TASKS



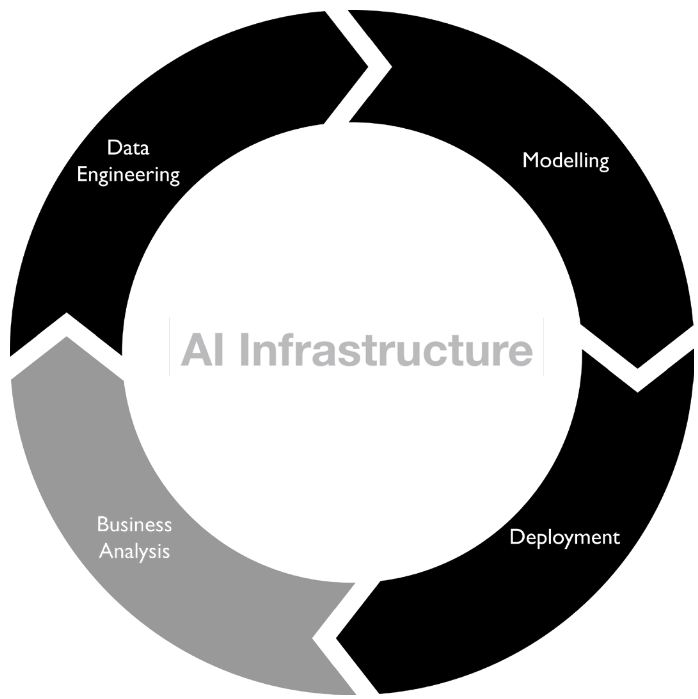
- Our definition of a Data Analyst is specific to an AI organization. It is different from what is usually referred to as a Business Analyst. The latter is less quantitative and focuses on creating data pipelines, cleaning data, and analyzing it. Data Analysts are accomplished in query languages such as SQL and commonly use spreadsheet software tools but don't need Algorithmic Coding skills.
- Communication skills are usually required, but the level depends on the team.
- Terminology: Companies may refer to this position as data scientist, research scientist, business analyst, risk analyst, marketing analyst, and other titles.

Machine Learning Engineer

SKILL PROFILE



TASKS



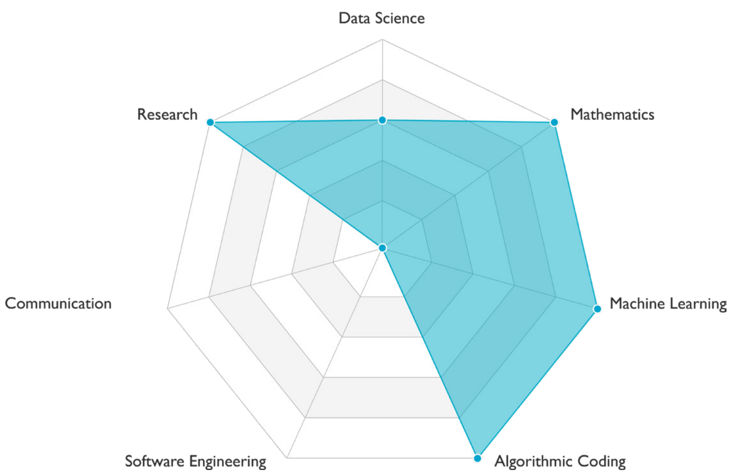
TOOLS MACHINE LEARNING ENGINEERS USE

- Data Engineering in Python and/or SQL (or other domain-specific query languages)
- Modeling in Python using packages such as numpy, scikit-learn, TensorFlow, PyTorch, etc.
- Deployment using an object-oriented programming language (e.g., Python, Java, C++, etc.) and cloud technologies such as AWS, GCP, Azure, etc.
- Collaboration and Workflow using a version control system (e.g., Git, Subversion, Mercurial, etc.), a Command Line Interface (CLI) (e.g., UNIX), an Integrated Development Environment (IDE) (e.g., Jupyter Notebook, Sublime, etc.) and an issue tracking product (e.g., JIRA)

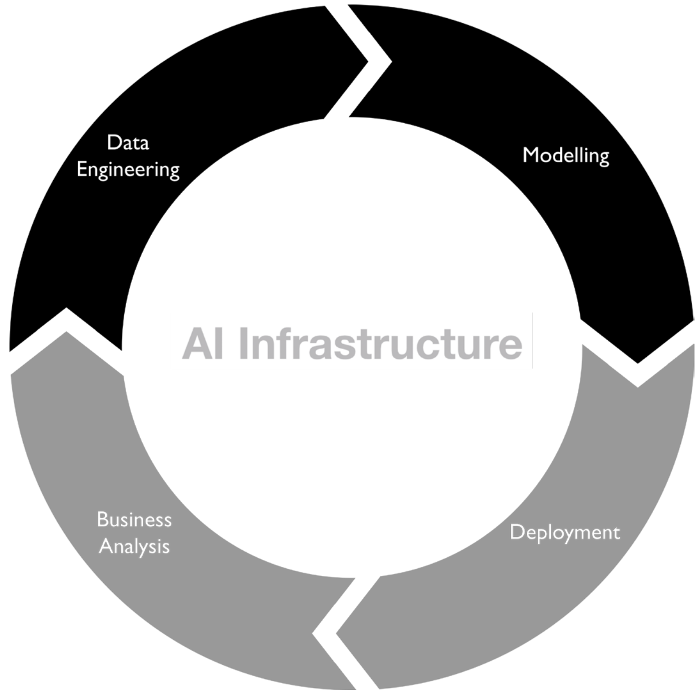
- There is a variant of the Machine Learning Engineer, called the Deep Learning Engineer, that requires deep learning knowledge in addition to the skills profile above. These engineers focus on applications usually powered by deep learning. Examples include speech recognition, natural language processing and computer vision. Hence, they need skills specific to deep learning projects such as understanding and using various neural network architectures (fully-connected networks, CNNs, RNNs, etc.).
- Although it depends on the team, communication skills and business acumen aren't usually strong requirements.
- Companies may refer to this position as: machine learning engineer, software engineer - machine learning, software engineer, data scientist, algorithm engineer, research scientist, research engineer, full-stack data scientist, and other titles.

Machine Learning Researcher

SKILL PROFILE



TASKS



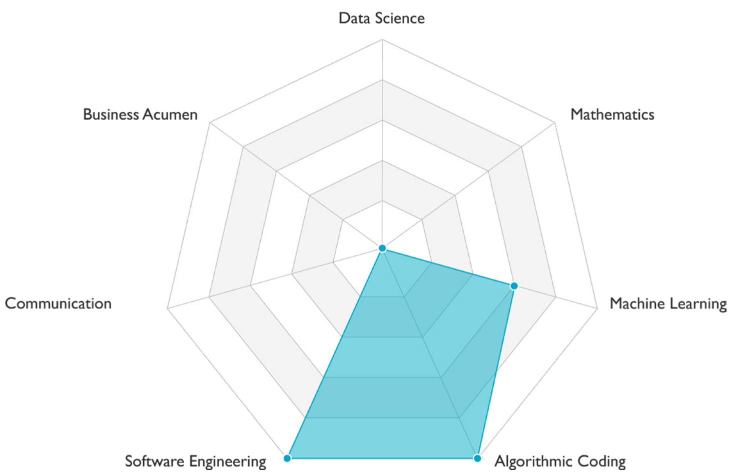
TOOLS MACHINE LEARNING RESEARCHERS USE

- Data Engineering in Python and/or SQL (or other domain-specific query languages)
- Modeling in Python using packages such as numpy, scikit-learn, TensorFlow, PyTorch, etc.
- Collaboration and Workflow using a version control system (e.g., Git, Subversion, Mercurial, etc.), a Command Line Interface (CLI) (e.g. UNIX), an Integrated Development Environment (IDE) (e.g., Jupyter Notebook, Sublime, etc.) and an issue tracking product (e.g., JIRA)
- Research by following updates via channels such as Twitter, Reddit, word of mouth, Arxiv, and various conferences (e.g. NeurIPS, ICLR, ICML, CVPR, ACM, etc.)

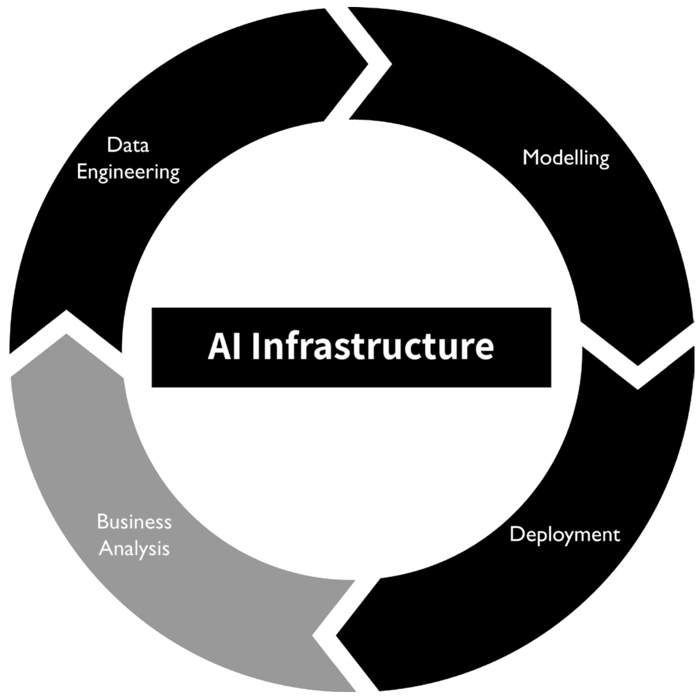
- There is a variant of the Machine Learning Researcher, called the Deep Learning Researcher, that requires deep learning knowledge in addition to the skills profile above. These engineers focus on applications usually powered by deep learning. Examples include speech recognition, NLP and computer vision. Hence, they need skills specific to deep learning projects such as understanding and using various neural network architectures (fully-connected networks, CNNs, RNNs, etc.).
- Although not represented on the graph above, some machine learning researchers focus on deployment (e.g., life-long learning, model memory optimization for edge deployment, etc.) or AI infrastructure (e.g., distributed training, scheduling, experiment, and resource management).
- Although it depends on the team, communication skills aren't usually a strong requirement.
- Companies may refer to this position as: machine learning researcher, research scientist, research engineer, data scientist, and other titles.

Software Engineer - Machine Learning

SKILL PROFILE



TASKS



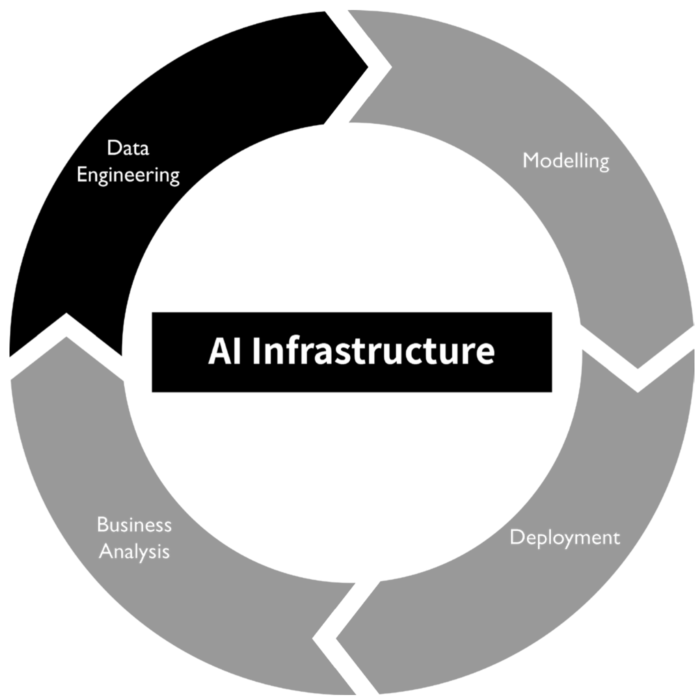
TOOLS SOFTWARE ENGINEER - MACHINE LEARNING USE

- Modeling in Python using packages such as numpy, scikit-learn, TensorFlow, PyTorch, etc.
- Data Engineering in Python and/or SQL (or other domain-specific query languages).
- Deployment and AI infrastructure using an object-oriented programming language (e.g., Python, Java, C++, etc.) and cloud technologies such as AWS, GCP, Azure, etc.
- Collaboration and Workflow using a version control system (e.g. Git, Subversion, Mercurial, etc.), a Command Line Interface (CLI) (e.g., UNIX), an Integrated Development Environment (IDE) (e.g., Jupyter Notebook, Sublime, etc.) and an issue tracking product (e.g., JIRA).

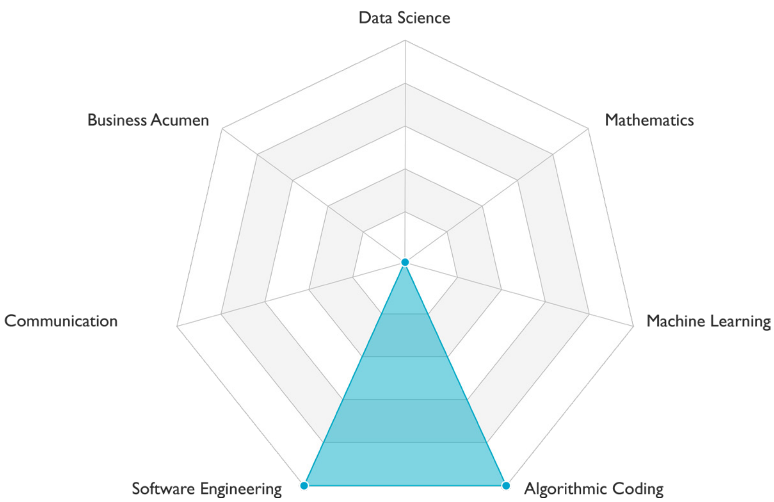
- There is a variant of the Software Engineer - Machine Learning, called the Software Engineer - Deep Learning, that requires deep learning knowledge in addition to the skills profile above. These engineers focus on applications usually powered by deep learning. Examples include speech recognition, natural language processing and computer vision. Hence, they need skills specific to deep learning projects such as understanding and using various neural network architectures (fully-connected networks, CNNs, RNNs, etc.).
- Although it depends on the team, communication skills and business acumen aren't usually strong requirements.
- Companies may refer to this role as: machine learning engineer, software engineer, full-stack data scientist, and other titles.

Software Engineer

TASKS



SKILL PROFILE



TOOLS SOFTWARE ENGINEERS USE

- Data Engineering in Python and/or SQL (or other domain-specific query languages).
- AI infrastructure using an object-oriented programming language (e.g. Python, Java, C++, etc.) and cloud technologies such as AWS, GCP, Azure, etc.
- Collaboration and Workflow using a version control system (e.g. Git, Subversion, Mercurial, etc.), a Command Line Interface (CLI) (e.g. UNIX), an Integrated Development Environment (IDE) (e.g., Jupyter Notebook, Sublime, etc.) and an issue tracking product (e.g., JIRA).

- Although it depends on the team, communication skills and business acumen aren't usually strong requirements.
- Companies may refer to this role as: data engineer, software engineer, software development engineer, software engineer - AI Infrastructure, software engineer-data.

This report is a work in progress and is being provided to the public for information purposes. Because it is a work in progress, there are parts that are either missing or will be progressively revised as our team learns more about the supply of and the demand for AI talent.

We welcome your comments and feedback. Please send any comments and/or questions to Kian Katanforoosh (kian@workera.ai).

- The Workera team

WORKERA

a deeplearning.ai company