Attain Capabilities Offering for Attain AI/ML CoE

# ToC

**(Red Highlights means it is not filled in yet)**

1. **Background Statement**
2. **Mission Statement**
3. **SMART Goals for CoE**
4. **Meeting Cadence/ Agenda**
   1. **Agile for DS**
5. **Means of Communication**
6. **CoE Engagement and Education Opportunities**
   1. **Training**
   2. **Open Discussion Forums**
      1. **Lunch and Learn Talks**
      2. **Project Pitches**
      3. **Book Club**
7. **AI/ ML Reference Material**
   1. **AI / ML Newsletter**
8. **Data** 
   1. **Data Science Lifecycle**
      1. **Institute Business Understanding**
      2. **Establish Ground Truth**
      3. **Model Development**
      4. **Visualization, Tools, and Reports**
      5. **Integration, Production, Monitoring**
         1. **DevSecOps**
         2. **QA Testing**
         3. **Peer Review Process**
            1. **Research Phase**
            2. **Development Phase**
   2. **Data Compliance**
9. **Technology** 
   1. **Planning & Development**
   2. **Infrastructure**
   3. **Tools and Services**
   4. **Insights**
10. **People** 
    1. **Talent Definition**
    2. **Talent Recruitment**
    3. **Talent Placement**
    4. **Talent Development**

Introduction

**What is the purpose of a Center of Excellence (CoE)?**

A CoE serves the following functions for an organization:

* They help connect people and provide shared context for people to communicate and share information.
* They act as a blueprint to jump start the next project.
* They provide dialogue for people with a shared interest in solving similar problems.
* They stimulate learning and offer resources for tools, coaching, and self-reflection.
* They capture and disperse existing knowledge and expertise in necessary subject matter.
* They generate new knowledge.

# 1 Background Statement

A background statement for a CoE sets the course for all other aspects of the CoE. Include in your background statement:

* Explanation of the environment that created the need for your CoE
* Description of the value your CoE will have to your organization and its individual members.
  + Why do we need a CoE? What is the value? (Capability Offering)
  + Who will your community serve? Who are your stakeholders?
  + What will members get out of your community?
  + Why would they want to contribute to the community?

# 2 Mission Statement

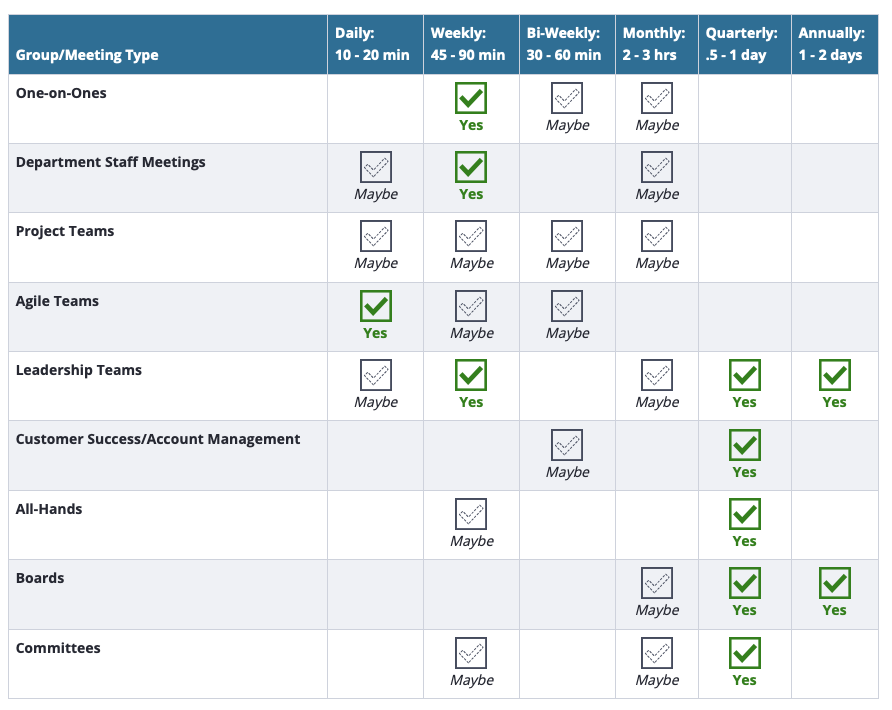
The mission statement builds on the background statement but focuses on the core purpose of your community by answering the following questions:

* What are the main objectives of your community?
* What are the responsibilities of the community to its members?
* What are the values of the community?

# 3. SMART Goals for CoE

The mission statement outlines big picture goals for the community. The purpose of the goal statements is to break these longer term goals into specific actions the community will take to fulfil its mission.

# 4 Meeting Cadence/ Agenda



*Image source:* [*link*](https://blog.lucidmeetings.com/blog/how-often-should-you-meet-selecting-the-right-meeting-cadence-for-your-team)

**Establish a consistent meeting cadence with crisp agendas to keep discussions focused**

* + Frequency of meetings will be dependent on what you want to accomplish and how many events/activities the CoE is able to manage. There may be some months where there will be an uptick in meetings due to planned activities.

## 4.1 Agile for Data Science

Incorporating the data science team into the agile process will be beneficial for many reasons. Aspects of agile framework to incorporate into data science include (but are not necessarily limited to):

* Sprints should be broken up into 2-3 week cycles
* Planning and prioritizing at the beginning of sprints
* Creating tickets with user stories, action items, deliverables, and timelines clearly defined
* Incorporating QA and PR practices into the Data Science team scrum board
* Participating in retrospectives and demos at the end of the sprint. Invite stakeholders to all relevant meetings to reduce uncertainty and gain visibility and a common understanding

# 5 Means of Communication

Determine the communication mode(s) that will be used

* Email: create an internal alias to easily communicate with members
* Intranet/SharePoint site
* Slack or other social business platforms
* Shared folder for documents

# 6 CoE Engagement and Education Opportunities

## 6.1 Training

### 6.1.1 Analyze the task and skill gaps within the project, and map skill needs to internal staff.

* Where are their project shortfalls?
* What project roles are they having trouble staffing?
* What specific technical and personal skills are needed?
* Are any project tasks getting delayed because there is no one who can do them?

Based on your findings, you can compile a list of task and skills gaps by project. From here, assess internal personnel to see who has the aptitude and background to step into these tasks and skills gaps, and then identify them as trainees. For each trainee, design a curriculum and find a project suitable for them.

### 6.1.2 Host a two-part informal training series on data and analytics activities occurring in the organization.

These trainings should be very high level. Don’t get into the weeds with the attendees; rather have your audience walk away with some ideas and better insight as to how analytics is being leveraged in the organization and create a space for ongoing conversation:

* 101: “How We Do It” training
* 201: “Ask Us Anything” session: once the organization understands what the analytics team does (101), host a session so they can gain a deeper understanding of the work taking place and its broader impact

### 6.1.3 Deep Dive Seminar/Boot Camp

* Educational and interactive day-long session: define the terminology (101 analytics), share use cases, conduct an interactive workshop, highlight the capabilities of the analytics team and some of the current projects in motion.

6.1.4 MOOC (Massive Open Online Courses)

* Create a curated list of the top 10 data and analytics courses to take online and promote across the organization. For example, Coursera ([2018’s Most Popular Data Science Courses](https://www.coursera.org/collections/popular-data-science-courses-2018)) or Khan Academy ([Computing](https://www.khanacademy.org/computing)); schedule follow up discussions with employees that have completed coursework and discuss how learnings apply to current work and activities.

### 6.1.5 Continuously revise the curriculum to keep up with real world project requirements.

## Some project needs will remain relatively constant while others will evolve as technology and business changes. It is essential, if you are developing training, to keep pace with these changes so your training always delivers the skills education that your projects need. You can ensure this cohesion by constantly evaluating projects, and then going back to your curriculum to ensure that the training is in sync with project needs.

## 6.2 Open Discussion Forums

### 6.2.1 Lunch and Learn Talks

A monthly lunch and learn should take place project wide, and should cover a broad range of interesting topics related to data science, analytics, and big data.

A list of scheduled lunch and learns should exist on a SharePoint page, including:

* The talk’s title and presenter
* An abstract of the talk
* A link to a presentation, reference material, or further readings(updated after the talk is given, if applicable)

Lunch and learns typically range from 30-60 minutes, depending on the audience size and goal of the discussion. At least 20% of the time should be dedicated to Q&A.

### 6.2.2 Project Pitches

### 6.2.3 Book Club

A quarterly Data Science book club should be available project / company wide. This will allow for data scientists and technical / non technical employees to come together and learn relevant information within the data sector. Book topics should not be extremely technical (I.e. it is best to not choose a book that goes into the mechanics of how a specific set of algorithms work), and instead should be books regarding general best practices in the world of Data Science/ Big Data / Data Analytics. They should be informative to a large audience of varied backgrounds.

Ideally, there should be a group in a communication channel (Slack, Microsoft Teams, etc...) dedicated to the book club members, where people can voice insights or ask questions.

At the end of the quarter, a literary review should be written and added to the SharePoint for literary reviews.

# 7 AI/ ML Reference Material

The following resources should all be available in a common directory of SharePoint and accessible to everyone on the project.

## 7.1 AI / ML Newsletter

Choose a cadence (preferably monthly or quarterly) to disseminate a newsletter containing the following relevant information:

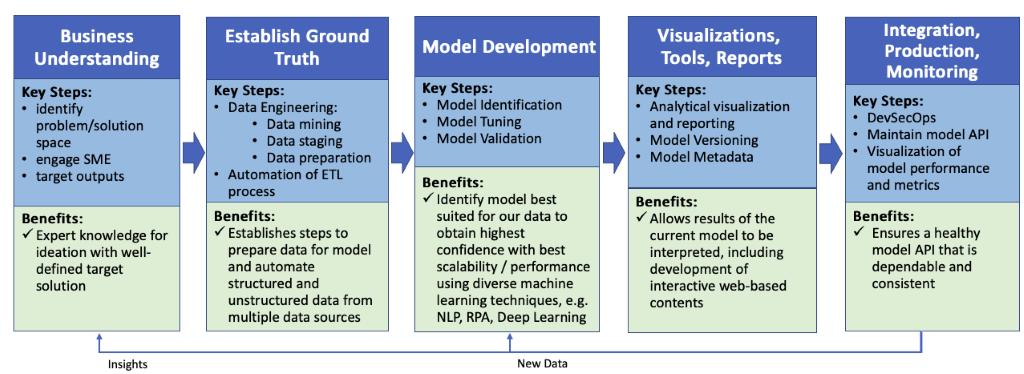
* Curated Data Science Industry news (specific to industry of project, not internal)- how is data science being used to tackle industry business problems?
* Tools / Techniques
* Data visualizations from internal data
* Data Science Tip
* Internal Blog Posts
* Project member spotlight / kudos
* Get Involved plug – data projects /collaborations internally (or externally) seeking involvement

## 7.2 Literature Review Process

Literature reviews are a vital step in any R&D project. The following template should serve as a template for data science literature reviews, which should be posted in a community SharePoint site:

|  |
| --- |
| **Article Title:** **Length of Article:**  **Topic:**  **Publishing Journal:**  **Search Terms used to Find Article:**  **Target Audience:**  **Links to previous reviews of paper:**  **References to Article:**  **Notes On Article** *(Use this space to document interesting findings, quotes, example code, etc, and if applicable cite where in the paper you are pulling it from):*  **Application for User** *(what does this article offer from a business perspective?)*:  **Next Steps** *(follow up tasks prompted by this article):* |

# 8 Data



## 8.1 Data Science Lifecycle

### 8.1.1 Business Understanding

The start to any data science project begins by asking the right business questions and identifying strategic processes to transform these questions into outcomes via models and analytics.

#### 8.1.1.1 Develop a Mission Statement

A mission statement created by Subject Matter Experts (SMEs) along with scientists and analysts will serve as a vision for the project as a whole, ensuring the entire team understands the goals of the project and what each person’s role is in achieving these outcomes.

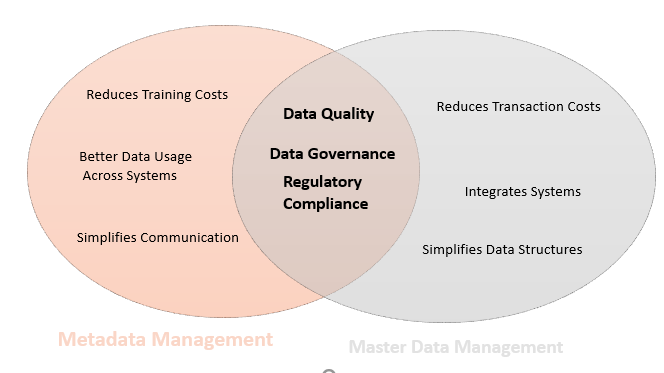
#### 8.1.1.2 Identify Business Problems, Solutions, and Target outputs

Once the mission statement is in place, the team can build upon it to extract key problems that need solving, and what form the metrics for the solutions should be in.

#### 8.1.1.3 Build a Business Glossary

Analysts and scientists work together with business experts to develop a business glossary. A business glossary defines terms within the business domain and provides authoritative sources for all business operations. Many experts concur that a business glossary is the most important component for the data governance team, as it allows the business to own its terms and meaning and emphasizes how vocabulary may differ across business functions. It is common that different business stakeholders will have different definitions for words, e.g. the word “invoice” from the perspective of a consumer refers to a document requesting for payment, while from an IT engineer’s point of view, invoice signifies the process of creating a bill.

#### 8.1.1.4 Establish Metadata Management and Master Data Management (MDM)



*Image source:* [*link*](https://www.dataversity.net/metadata-management-vs-master-data-management/#)

8.1.2 Establish Ground Truth

#### *8.1.2.1 Data Engineering Pipeline*

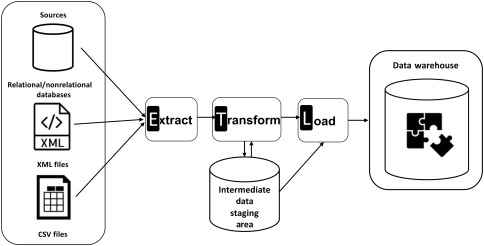
If the data engineering pipeline is already in place, the data scientists should have initial meetings with the engineers / owners / maintainers of the pipeline to establish the following points:

* Get an understanding of the data flow from pipeline owners
* Communicate data needs from pipeline owners.
  + Most likely, this will be in terms of formatting and properties of data. The data engineers and data scientists need to agree upon these data properties and check back with business experts to ensure that the data being captures is that data needed to solve the business problems.

If the data pipeline is not already set up, these are the steps that need to be taken to do so. These steps are usually performed by data engineers, but data scientists may be a part of the process, or at the very least it is good practice for data scientists to understand and have visibility around the process:

Before we can begin to manipulate and explore the data, we need to have a reliable data flow and infrastructure for our data ETL (extract, transform, load) process. The first phase in this process is to have data engineers design, build and maintain a data warehouse(s). The following steps outline the best practices for setting up and maintaining a **cloud based data flow**:

* **Have a data model**. This process was laid out in section 9.1.1.
* **Have a data flow diagram (DFD)**. This allows visibilities into where all of the business’ data repositories exist, as well as how the data traverses throughout the company’s infrastructure. The DFD should also provide information about the outputs and inputs of each entity and process.
* **Build a source agnostic integration layer.**
* **Adopt a data warehouse architecture standard (star schema, data vault)**
* **Data Staging and Processing.** Whether you pull data from production systems or from a staging area, the data has to be processed before it lands in the data warehouse.

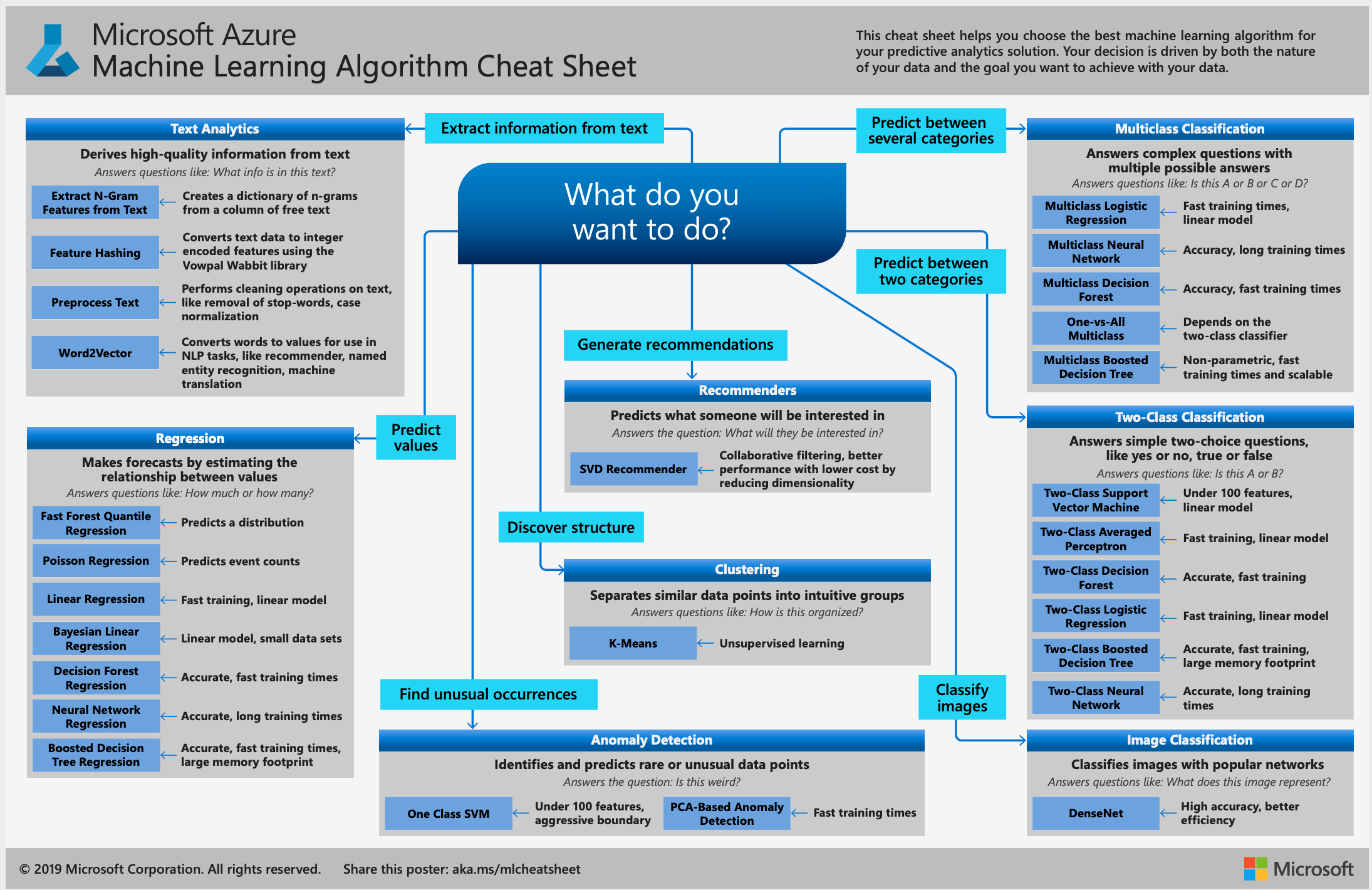


* **ETL / ELT:**
  + **Extract** the data from operational/ external data sources
    - Data is consolidated from numerous, disparate sources that are potentially stored in different formats, converting data into a suitable format for transformation process.
  + **Transform** the data into an appropriate format
    - validation /rejection of records
    - Conversion
    - Clearing duplicates
    - Standardizing, filtering, sorting
    - Translating and looking up or verifying if data sources are inconsistent
  + **Load** the data into a data warehouse repository.
    - Import extracted and transformed data into a target data warehouse.
* **Automate data warehouse.**

8.1.3 Model Development

*8.1.3.1 Model Identification*

* + **Define Success**. This step defines how the model will be used and its accuracy requirements, and provides guidance for the technologies which should be leveraged to build the model. The key metrics to consider when defining success:
    - **What is the problem you are solving/ What are the obtainable objectives?** This should be defined in section 9.1.1.
    - **Building A Better Business**. When evaluating a data science project, there should exist a list of items that, if successful, describe a completed and successful project. E.g. a document management system will contain factors such as breadth of documents, intensity of NLP training, how quickly a reinforcement learning bot can master a task.
  + **Explore Data**. This step allows data scientist to better understand the problem before they tackle it, as well as allows them to create reports and visualizations for business stakeholders so the entire team can understand the raw data. Key factors to look at when exploring data include looking at summary statistics of variables, ranges and dispersion, examining scatter matrix plots of variables and targeted responses to try to understand pairwise behaviors, consider missing variables within records, apply visualization techniques when possible.
  + **Condition Data**. It is necessary to perform operations to the data (usually which can be automated) in order to make it cohesive, thorough, and suitable for a machine learning model. The following steps are taken to achieve this:
    - (when applicable) re-scale or normalize features, so that all values are within a similar scale that is compatible for input of a model. Also, to sometimes it is necessary to artificially create limit points, ensuring the proper limit behaviors if not present in the training data.
    - Fill or remove missing values, as ML models cannot take a null value as input. It is not wise to arbitrarily fill a missing value with zero or another naïve value, so having an understanding of the statistical meaning of the value’s field or the business meaning of the field can help accomplish this.
    - Map ordinal (categorical/ qualitative) data to a numerical representation, e.g. the feature *performance* {poor, fair, good, great} would be mapped to {1,2,3,4}.
    - **Select Variables**. Ensures that we do not include spurious or correlated variables into our model. If we do, this can both slow down model performance by using unnecessary data in the model, as well as can cause errors in predictions.
      * Use **Symbolic regression** for nonlinear approach
      * Use **Principal Component Analysis (PCA)** for linear approach.
  + **Balance Data**. Say, for example, we are predicting whether a customer will default on a loan, and our historical data shows that only 10% of customers default on loans. This is considered to be unbalanced data, and many machine learning models perform much better when the target categories are equal in size.
  + **Choose the Algorithm**



*Image Source:* [*link*](https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-cheat-sheet)

*There are a lot of questions that are involved in choosing the algorithm we want to use in our model. Many of them can be illustrated in the flow chart provided by Microsoft Azure depicted above, but there are some additional questions to keep in mind.*

* + *All algorithms require at least 50 samples (and many require even more than that)*
  + *If your goal is to predict* ***category membership****, then you have two choices:*
    - *If your* ***data is labeled,*** *then you would use a classification model*
      * *If you have* ***more than 100k sample****s, then choose SGD classifier (and apply kernel approximation if necessary)*
      * *If you have* ***less than 100k sample****s, then choose Linear SVC.* 
        + *If linear SVC is not working and you have text data, choose Naive Bayes, or if you do not have text data choose K-neighbors classifier or an ensemble classifier*
    - *If your* ***data is unlabeled****, you would use a clustering method*
      * *If the* ***number of categories is unknown*** *and you have at least 10k samples, choose Mean Shift or VBGMM (if you have less than 10k sample then it might be hard to choose a great algorithm for this data set).*
      * *If the* ***number of categories is known*** *and you have more than 10k samples, then choose K Means, or if that does not work Spectral clustering or GMM. If you have less than 10k samples, choose Minibatch KMeans.*
  + *If you are predictions are* ***regression or dimensionality reduction,*** *then there are a few things to keep in mind as well:*
    - *For regression, if you have* ***less than 100k samples,*** *choose a SGD Regressor. If you have* ***more than 100k samples*** *and you only have a few important features, choose Lasso or ElasticNet, otherwise choose Ridge Regression or SVR with a linear or rbf kernel.*

*8.1.3.2 Model Tuning*

Tuning is the process of maximizing a model’s performance without overfitting or introducing too much variance into a model’s performance. The main way to tune a data science model is to adjust its hyperparameters. Using automated approaches to hyperparameter selection is most commonly the best approach, and the three most common methods are outlined below:

* **Grid Search, aka parameter sweeping.** 
  + One of the most basic and traditional methods of hyperparametric optimization. Manually defines subsets of the hyperparameter space and exhausts all combinations of hyperparameters. Based on model performance using cross validation, the best performing hyperparameters are chosen. This can be computationally expensive, I.e. if a model takes 2 minutes to run and there are 100 different possibilities for hyperparameters, than this process takes 200 minutes. If this is not a viable option, the below two options may be more ideal.
* **Random Search**
  + Similar to grid search, but less computationally expensive as they do not exhaust the possible parameter combinations (using random selection, as the name suggests). The downfall to random search is that since it is completely random, you may completely miss the most optimal selection. This is where Bayesian optimization prevails:
* **Bayesian Optimization**
  + This process uses a probabilistic model for a given function to analyze and determine where to evaluate the function. Bayesian optimization contains 2 main components:
    - A **prior function** which takes into account the behavior of the unknown objective function and an observation model that describes the data generation mechanism
    - A **loss function** (usually [regret](https://en.wikipedia.org/wiki/Loss_function#Regret)) that describes how optimal a sequence of queries are.

*8.1.3.3 Model Validation*

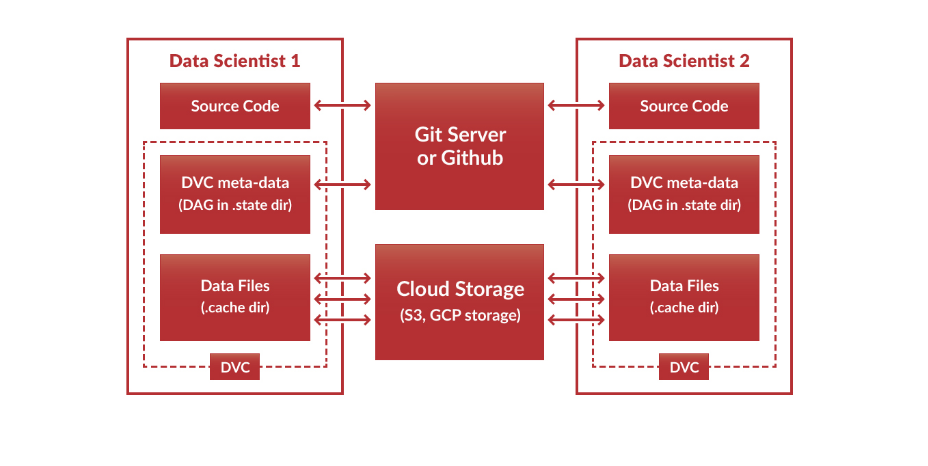
*Model validation takes on two components:*

* *validating a model from a data science perspective is the process of evaluating a trained model with a testing data set, separate from the data set that the model was trained with. The purpose of model validation is to assure the model is not overfit and can generalize to new data that will be fed to it overtime.*
* *Validating a model from a data governance perspective is, according to the American Academy of Actuaries,* “the practice of performing an independent challenge and thorough assessment of the reasonableness and adequacy of a model based on peer review and testing across multiple dimensions.”1

*At Attain, we follow the eight core validation principles set forth by the North American CRO Council\*:*

1. **Model design and build need to be consistent with the model’s intended purpose.** This involves having a clear understanding of the problem statement that the model is intended to answer, understanding the model’s limitations, and reviewing the model to ensure connection with its logic and flow.
2. **Ensure that model validation is an independent process.** The model validation step is separate from building the model, and typically an independent person or group should be responsible for validating a model.
3. **Establish an owner of model validation.** In alignment with point (2), a single individual should be held accountable whom is empowered to escalate related concerns and issues with model validation, as well as is responsible for attesting, resolving and reporting on model validation aspects.
4. **Ensure appropriateness of established model governance.** This ensures access rights are updated and takes internal audit into consideration of the model.
5. **Make model validation efforts proportional to evidenced areas of materiality and complexity.** Here, we define materiality and complexity and model and align validation with the time and resources needed. This includes incorporating a balance between model precision and timelines of results.
6. **Validate the model components.** This includes validation of model input components, calculation components, and output components.
7. **Address limitations of model validation.** The validation limitations should be recognized and made clear, and priorities should be set for future enhancements of both the model design and validation.
8. **Document the model validation.** This documentation should encompass aspects of preceding model validation principles and be aligned with its usefulness to the business. Keep in mind that validation is primarily about accountability, not documentation.

8.1.4 Model Visualizations, Tools and Reports



*8.1.4.1 Analytical Visualization and Reporting*

This step brings quite possibly the most quantifiable benefits to the stakeholders, as this is the stage in the data science process where results and findings from the data are interpreted and passed along to the business.

A **business stakeholder POC dashboard** should be established early on in the data science project. The steps involved in the process are:

* Involve the business stakeholders in a design thinking workshop at the inception of the project where analysts and stakeholders may brainstorm around what meaningful analytics would look like to them.
* The dashboard should be usable even by people with limited visualization skills, and it should include visuals on Business ROI and before vs. After scenario analysis.

**Add more here, but what?**

*8.1.4.2 Model Versioning*

The machine learning process is extremely iterative, and versioning our models is thus particularly important. As our data, infrastructure, and business needs adapt and grow over time, our model needs to do the same to keep up with our needs. Models are typically rolled out gradually, in order to ensure failure tolerance and incorporate QA processes. Versioning allows the developer to deploy the correct version of the model at the appropriate time.

The most widely accepted tool used in model version control is **Git** (Github, Gitlab). Data Version Control (DVC) is also an essential component of version controlling, and is offered as a git extension that works directly with cloud storage.

Another useful tool in versioning is using a software like [**Metaflow**](https://docs.metaflow.org/introduction/what-is-metaflow), a python library developed to help data scientists build and manage data science projects by managing libraries, loading and storing data, debugging, versioning, and architecting jobs.

8.1.5 Integration, Production, Monitoring

*8.1.5.1 DevSecOps*

A traditional DevOps workflow consists of:

* establishing a workflow
* ensuring existence of a constant feedback loop
* maintaining a culture of experimentation.

With DevSecOps, security is always a top priority and therefore identifying the most immediate security challenges will always be step 1, and a workflow will be built around it. Having an emphasis on security allows a boost in speed and innovation as it reduces the potential of having to fix vulnerabilities farther along in the development process. It also benefits AI developers, as continuous testing leads to more reliable results from their models.

*8.1.5.2 QA Testing*

In data science, performing tests on data and properties of data is typically more valuable than testing on just the code.

* **Unit Testing**. Unit testing is a required step in any data science project. A good unit test is **Readable, Simple, and Unique-** each unit test should test exactly 1 edge case, and if it is not human readable then it is too hard to test if your coverage is sufficient.
* When testing SQL Queries, these are common components to test on:
  + Joins
  + Text formatting
  + Anything that could result in a duplicate (or be affected by a duplicate)
  + Fields that will appear in conditional statements further along in the model building pipeline
  + Everything in a WHERE clause
  + Everything in an ORDER BY clause

#### 8.1.5.3 Peer Review Process

*This process was borrowed from* [*this article*](https://towardsdatascience.com/peer-reviewing-data-science-projects-7bfbc2919724)*, and more details can be found by following the link provided.*

##### 8.1.5.3.1 Data Science Research Phase Review

The motivation behind peer reviewing the research phase of the project is to prevent choosing the wrong approach too early in the project.

The goal is to catch costly errors early in the project, as well as improving the ability of the data scientist to explain and defend her decisions in future review processes.

**Types of Checks:**

* **Technical Validity Check.** Done in tandem with data and software engineers supporting the project.
* **Research Review.**
* **Scope and KPIs validity check**.

**Structure:**

* The reviewed data scientist prepares a presentation of the research process he went through.
* The reviewed data scientist sets up a long meeting with reviewer(s), to discuss:
  + Project scope and product needs
  + Initial guidelines and KPIs
  + Assumptions made
  + Data Used, and how it was explored
  + List of possible approaches / solutions arrived at
  + Approach going forward
  + Possible failures will lead to alternative choices
* The reviewer goes over checklist before meeting
* A suggested structure for the review meeting includes:
  + DS under review presents research phase presentation
  + Review gives general feedback
  + Reviewer goes over the checklist
  + Reviewer approves or rejects
  + Together, resulting action points are made

**Checklist:**

This checklist contains questions that should be addressed during research phase.

* **Data Properties**
  + Regarding the initial dataset: How was it generated / sampled/ updated? Did this introduce noise, sampling bias, or missing data? Can you modify sampling/ generation to reduce/ eliminate noise?
  + How was the dataset labeled and did this introduce label bias and can this be measured?
  + How similar is the initial data set to the input data expected in production, structure and schema-wise?
* **Approach Assumptions**
  + What assumptions does each approach make on the data / data generation process / phenomenon under study? Can these assumptions be validated independently?
* **Past Experience**
  + What experience do you have applying this approach? Did you find any published success/failure stories of applying this approach to similar problems?
* **Objective Alignment**
  + What loss functions can be applied and how do they relate to the project KPIs? What measure does the method optimize?
* **Implementation**
  + Is the implementation up to date and consistently supported?
* **Scaling**
  + How does computation/ training time scale with the number of data points or features?
* **Composability / Breakability**
  + Can models for different clients be composed in a sensible way and can their results be integrated in a meaningful way?
  + Can a general model be broken down into per-domain models?
* **Information Requirements**
  + To what degree does each approach rely on the amount of available information?
* **Cold Start & Domain Adaptation**
  + How fast will new clients expect to reach the minimal information required for an approach?
* **Noise/ Bias/ Missing Data Resilience**
  + Does the approach handle noise in data and how?

##### 8.1.5.3.2 Data Science Model Development Phase Review

This is the stage in the project where most data processing has been performed, models have been applied to actual data, and models have been benchmarked and validated.

This process is much the same as the research phase, with the following differences:

During the presentation, the Reviewed Data Science should present the following points:

* Project scope and product needs
* Selected approach after research phase
* Metric and soft/ hard constraints for model selection
* Data used, preprocessing and feature engineering
* Models examined, training regime, hyperparameter optimization
* Selected model, alternatives and pros/cons of each
* Next steps: automation, productization, performance optimization, monitoring, etc.

Checklist for reviewing the model development phase.

* Data Assumptions
* Preprocessing
* Leakage
* Causality
* Loss/Evaluation Metric
* Overfitting
* Runtime
* Stupid Bugs
* Trivial Questions

*8.1.5.4 Maintain Model API*

Once a model is ready for production, it needs to be turned into an API with a choice of web service framework (Flask is most common for Python applications- [here's](https://www.datacamp.com/community/tutorials/machine-learning-models-api-python) a great tutorial).

The data scientist will work with the client to put together documentation regarding the API input and output (S[wagger](https://swagger.io/blog/api-documentation/what-is-api-documentation-and-why-it-matters/) is a useful software to build API documentation).

*8.1.5.5 Visualization of Model Performance and Metrics*

*The model performance metrics chosen will be highly based on both the type of algorithm chosen as well as the type of data and business question being solved by the model.*

**Classification metrics**

* **Accuracy**- used to determined the percent of correct predictions, out of all classes.
* **Precision**- a.k.a. positive predictive value, the number of accurate positives compared to the number of positives it predicts. (how many selected items are relevant?)
* **Recall (sensitivity)**– a.k.a. true positive rate, the number of true positive predictions our model makes compared to actual positives. (how many relevant items are selected?). If correctly identifying positives is important, then choose a model with higher recall.
* **F-Score** - the harmonic mean of precision and recall.

*Note: precision and recall should be used when classes are not evenly distributed, e.g. if you are modeling for cancer detection, the percentage of people who have it to people don’t is extremely varied. Since F-Score takes both precision and recall into account, this is best to use when you want both scores weighted in the metric of the model.*

* **ROC** (Receiver operating characteristic)- plot of the true positive rate against false positive rate, helps in deciding the best threshold value in a logistic model.
* **AUC** (Area under the curve)- gives the rate of successful classifications by a logistic model and allows us to compare the ROC curve of one model to another.

**Regression metrics**

* **Mean Absolute Error (MAE)-** measures the average magnitude of errors, and is the most interpretable of the metrics.
* **Root Mean Squared Error (RMSE)-** quadratic scoring rule that measures the average magnitude of error, and gives a higher weight to larger errors than MAE, so use when higher errors are less desirable.
* **Mean Squared Error (MSE)**- average of the squared differences between predicted and actual output, commonly used since it is indifferent to whether the prediction was too high or low but just cares how incorrect it was. Since it is squared, it more heavily penalizes larger errors.
* **R2 Score (Coefficient of Determination)** proportion of variance in the outcome that our model is capable of predicting (between 0 and 1). R2 is great for summarizing the strength of a relationship, however it should not necessarily be used to validate a models performance or to compare models.

## 8.2 Data Compliance

### 8.2.1 Regulation-Compliant AI

*8.2.1.1 Data De-identification*

One requirement of the GDPR is data de-identification, which removes information from real people and desensitizes it for third part access. There are three main ways we approach this challenge:

* **Data Anonymization.** Erases or encrypts identifying components of data. Just anonymizing data is typically not sufficient in itself, though, as there are de-anonymization methods that can retrace the data to the original source.
* **Data Pseudonymization.** Replaces private identifiers with fake identifier, i.e. renaming. This method allows the preservation of statistical accuracy and data integrety.
* **Data Generalization.** Removes parts of data to make it less identifiable.

*8.2.1.2 Data Encryption*

Data encryption can be employed when training neural networks and other AI models. Methods such as symmetric encryption, asymmetric encryption, or hybrids of the two, help ensure that the initial state of the data cannot be regenerated by foreign agents. In order to compute the encrypted data, scientists use homomorphic encryption which produces an encrypted output. This process is a priority for data scientists as it brings up many complications, but is a necessary component in data security.

*8.2.1.3 Synthetic Data Generation*

Synthetic data is common in AI models. The generation method algorithmically manufactures new data with no connection to real sources. Not only is this useful for model performance, but it also increases privacy and security.

9 Technology

## 10 People

## 10.1 Talent Definition

Defining talent in the data science space is an incredibly challenging, and vital, task for an organization. It is important to understand all of the roles encompassed within data science. When setting up a data science team, there are generally four main types roles that a company wants to consider:

* **Data theorist/ generalist**, potentially the highest value-add in the set, these are the people who have extensive familiarity with big data tools and experience working with real life data sets and the technologies necessary to wrangle such data.
* **Data engineer,** these people help to set up data infrastructure and work with the data modelers and analysts to ensure that these foundational steps are set up properly so the end result can be obtained as painlessly as possible.
* **Data modeler / machine learning engineer,** the most typical application of a “data science” job description, these are the people who take existing models and apply them to the companies data to solve business problems. Consumer-facing companies with tons of data as well as companies offering data-based services would serve best from hiring for these roles.
* **Data analyst,** the primary description of a data analyst is to search for and manipulate data for further advanced analysis.

## 10.2 Talent Recruitment

Recruiters should work closely with hiring managers to generate accurate job discriptions based on the roles outlined in section 10.1. When creating and hiring for the role, it is important to think long term to understand how to leverage the organizations plans and product roadmaps.

## 10.3 Talent Placement

Data scientists thrive at the focal point of business, development, and analysis. It is hard for a data scientist to thrive in isolation. As discussed in earlier sections, data scientists need to converse with business experts and stakeholders within the company in order to build a clear picture of the business needs from the data. Their should be a culture of free-flowing communication amongst the data scientists and product managers, business developers, engineers, and customer-facing members.

## 10.4 Talent Development

In terms of continued learning and development, much of this is already laid out in section 6.1.

Attain already has standard for development, should we just add that here?

*Resources:*

* [*https://towardsdatascience.com/peer-reviewing-data-science-projects-7bfbc2919724*](https://towardsdatascience.com/peer-reviewing-data-science-projects-7bfbc2919724)
* [*https://www.iianalytics.com/blog/2019/8/23/creating-an-analytics-community-of-practice*](https://www.iianalytics.com/blog/2019/8/23/creating-an-analytics-community-of-practice)
* [*https://www.techrepublic.com/article/how-to-develop-an-internal-data-science-training-program/*](https://www.techrepublic.com/article/how-to-develop-an-internal-data-science-training-program/)
* [*https://www.dataversity.net/business-glossary-basics/*](https://www.dataversity.net/business-glossary-basics/)
* [*https://www.snowflake.com/blog/top-9-best-practices-for-data-warehouse-development/*](https://www.snowflake.com/blog/top-9-best-practices-for-data-warehouse-development/)
* [*http://bi-insider.com/data-warehousing/three-steps-in-etl-processing/*](http://bi-insider.com/data-warehousing/three-steps-in-etl-processing/)
* [*https://riskspan.com/news-insight-blog/tuning-machine-learning-models/*](https://riskspan.com/news-insight-blog/tuning-machine-learning-models/)
* [*http://proceedings.mlr.press/v22/lacoste12/lacoste12.pdf*](http://proceedings.mlr.press/v22/lacoste12/lacoste12.pdf)
* *1* [*https://www.actuary.org/sites/default/files/files/publications/ModelGovernancePracticeNote\_FinalDraft\_10.30.2016.pdf*](https://www.actuary.org/sites/default/files/files/publications/ModelGovernancePracticeNote_FinalDraft_10.30.2016.pdf)
* *\** [*http://www.crocouncil.org/images/CRO\_Council\_-\_Model\_Validation\_Principles.pdf*](http://www.crocouncil.org/images/CRO_Council_-_Model_Validation_Principles.pdf)
* [*https://algorithmia.com/blog/how-to-version-control-your-production-machine-learning-models*](https://algorithmia.com/blog/how-to-version-control-your-production-machine-learning-models)
* [*https://technology.cloverhealth.com/a-data-scientists-guide-to-writing-unit-tests-ff1dd716aabc*](https://technology.cloverhealth.com/a-data-scientists-guide-to-writing-unit-tests-ff1dd716aabc)
* [*https://towardsdatascience.com/peer-reviewing-data-science-projects-7bfbc2919724*](https://towardsdatascience.com/peer-reviewing-data-science-projects-7bfbc2919724)
* [*https://medium.com/ml-cheat-sheet/machine-learning-evaluation-metrics-b89b8832e275*](https://medium.com/ml-cheat-sheet/machine-learning-evaluation-metrics-b89b8832e275)