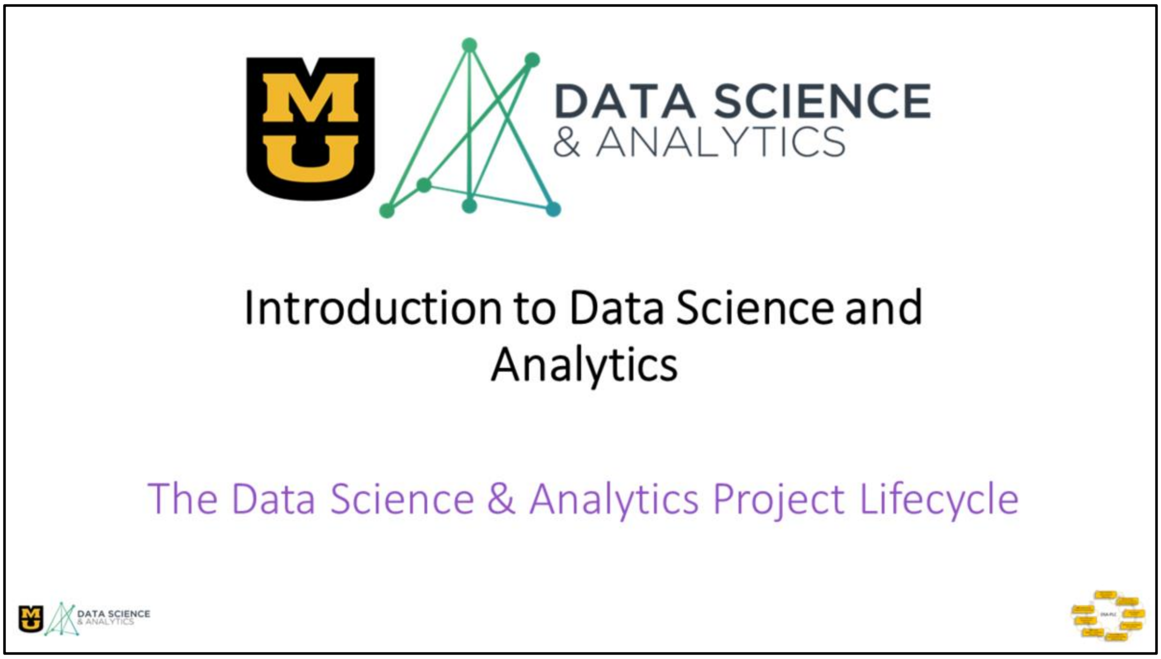


DATA SCIENCE
& ANALYTICS

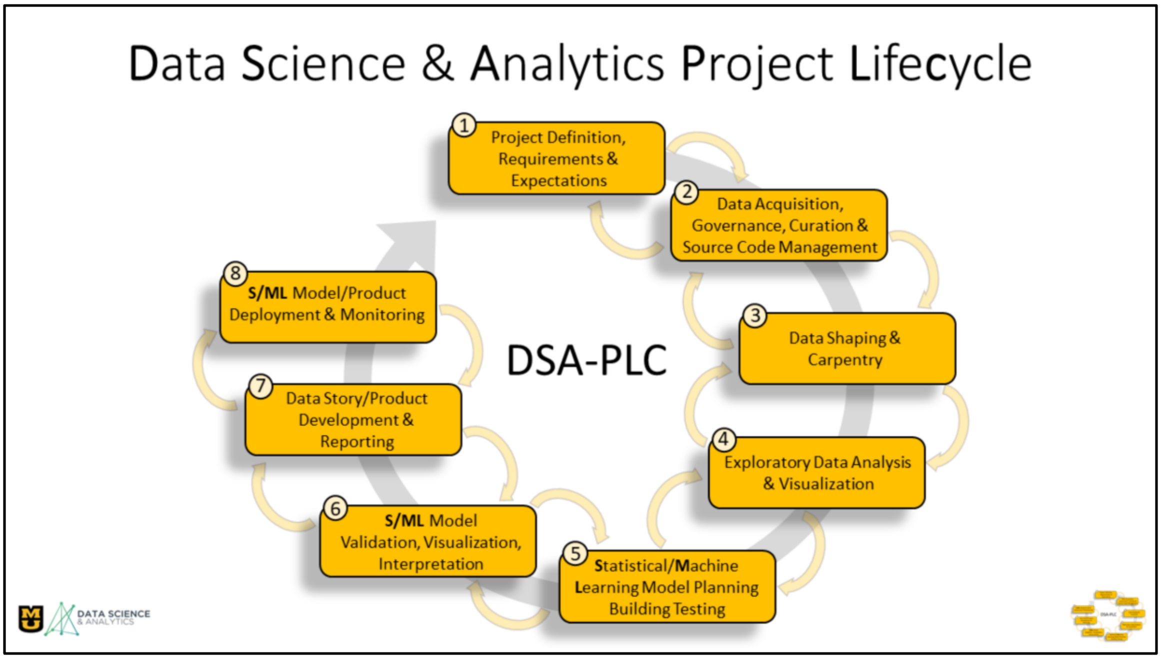
Introduction to Data Science and Analytics

The Data Science & Analytics Project Lifecycle

Many analytics problems or ideas seem huge and daunting at first, but using a well-defined process enables you to break down complex problems into smaller steps that you can more easily address.

Utilizing an established process for performing analytics is critical for ensuring that you produce a comprehensive, documented, and repeatable method for conducting analysis.

This presentation provides an overview of the Data Science and Analytics Project Lifecycle. The DSA-PLC is a reproducible, iterative process for designing and conducting data science and analytics projects.



Data Science & Analytics Project Lifecycle

DSA-PLC

1. Project Definition, Requirements & Expectations
2. Data Acquisition, Governance, Curation & Source Code Management
3. Data Shaping & Carpentry
4. Exploratory Data Analysis & Visualization
5. **S**tatistical/**M**achine **L**earning Model Planning Building Testing
6. **S/ML** Model Validation, Visualization, Interpretation
7. Data Story/Product Development & Reporting
8. **S/ML** Model/Product Deployment & Monitoring

DATA SCIENCE & ANALYTICS

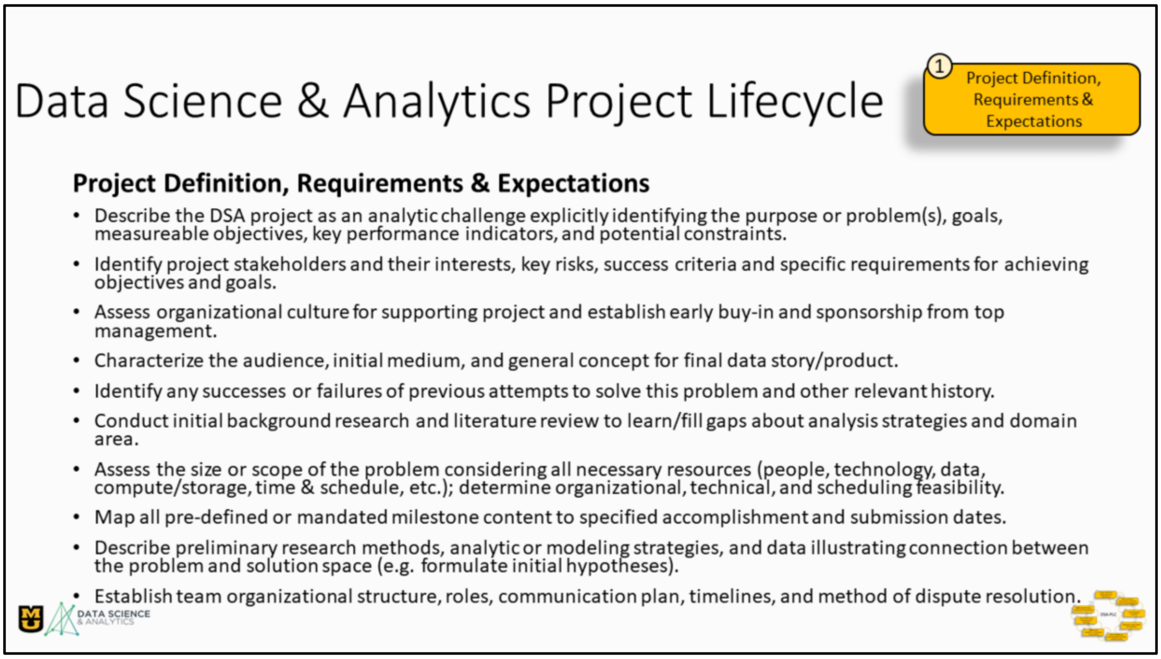
There are many example life-cycle models used to portray the relationships between the stages and phases of various processes involved in projects; some lifecycle models have more or less stages but this particular model characterizes much of the work performed in Data Science and Analytics and particularly how our academic program addresses DSA projects.

There are 8 stages in this DSA project Lifecycle. Work on a project will normally begin in stage 1 but once enough planning is complete, additional work can be done in several stages at once. For example, while data are being acquired, shaped, and explored, drafting the outline and narrative content of the data story can also occur. Movement from any stage to another and back again to previous stage occurs throughout the project Lifecycle. A stage is not a single fixed that you advance to and never return like with the waterfall model; the stages in the DSA lifecycle is dynamic and highly iterative.

Once you have enough information and have made enough progress in any given stage, you can move to the next stage of the process. As you can see in the graphic, you can often learn something new in a stage to cause you to go back and refine work done in prior stage given new insights and information that you've uncovered.
For this reason, the graphic is shown as a cycle and the circular arrows are intended to convey that you can move iteratively between stages until you have sufficient information to continue moving forward. In fact, within each stage you will likely be working iteratively to accomplish major tasks.

Data Science projects tend to be less well-structured and lack definite scope or size at the start. For this reason, it is common to revise the action plan derived in in the planning stage one or more times, and thus work through each stage again to address the revisions and prescribed changes.

Going forward, I will provide an overview of each of the stages and discuss their most salient features and content.



Data Science & Analytics Project Lifecycle

1
Project Definition, Requirements & Expectations

Project Definition, Requirements & Expectations

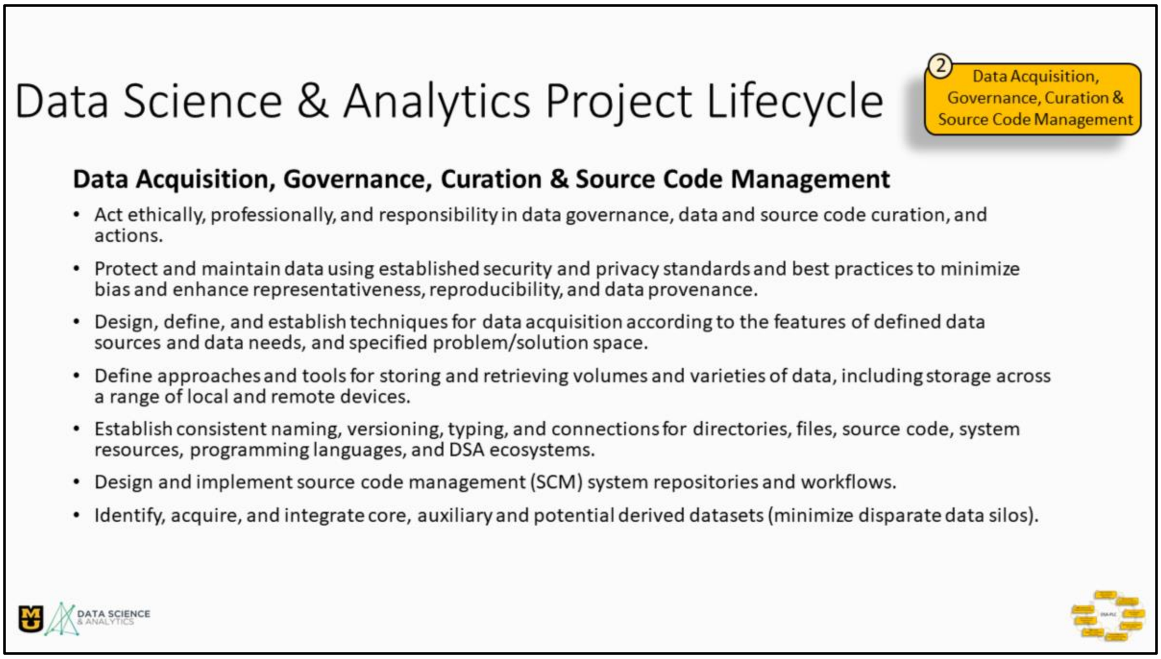
- Describe the DSA project as an analytic challenge explicitly identifying the purpose or problem(s), goals, measureable objectives, key performance indicators, and potential constraints.
- Identify project stakeholders and their interests, key risks, success criteria and specific requirements for achieving objectives and goals.
- Assess organizational culture for supporting project and establish early buy-in and sponsorship from top management.
- Characterize the audience, initial medium, and general concept for final data story/product.
- Identify any successes or failures of previous attempts to solve this problem and other relevant history.
- Conduct initial background research and literature review to learn/fill gaps about analysis strategies and domain area.
- Assess the size or scope of the problem considering all necessary resources (people, technology, data, compute/storage, time & schedule, etc.); determine organizational, technical, and scheduling feasibility.
- Map all pre-defined or mandated milestone content to specified accomplishment and submission dates.
- Describe preliminary research methods, analytic or modeling strategies, and data illustrating connection between the problem and solution space (e.g. formulate initial hypotheses).
- Establish team organizational structure, roles, communication plan, timelines, and method of dispute resolution.

Stage 1 is the Project Definition, Requirements, & Expectations stage where the project plan originates.
It is crucial to define the business or problem domain, and frame the business problem as an analytic challenge that can be addressed in subsequent stages. Formulate initial hypotheses to test and begin learning the data. This is also the time to identify and document background information including relevant history of the problem and seek material describing whether your organization, business unit or professional literature has attempted similar projects in the past. You should learn as much as you can and fill in the content and technical knowledge gaps early in this stage.

It is also vital to assess the resources you will need to support the project, in terms of people, technology, time, and data. Identify the intended audience for which your story is relevant. Address stakeholder expectations and management buy in and support.

Set an initial timeline for milestone due dates. Describe research methods and analytic strategies that align with the problem space and assess the overall feasibility of project success. It is also the time to establish the team structure and roles and methods of dispute resolution since the will be times of disagreement.

The outcome of this stage is a relatively detailed written action plan that may need management review and approval before moving on to other stages.



Data Science & Analytics Project Lifecycle

2 Data Acquisition, Governance, Curation & Source Code Management

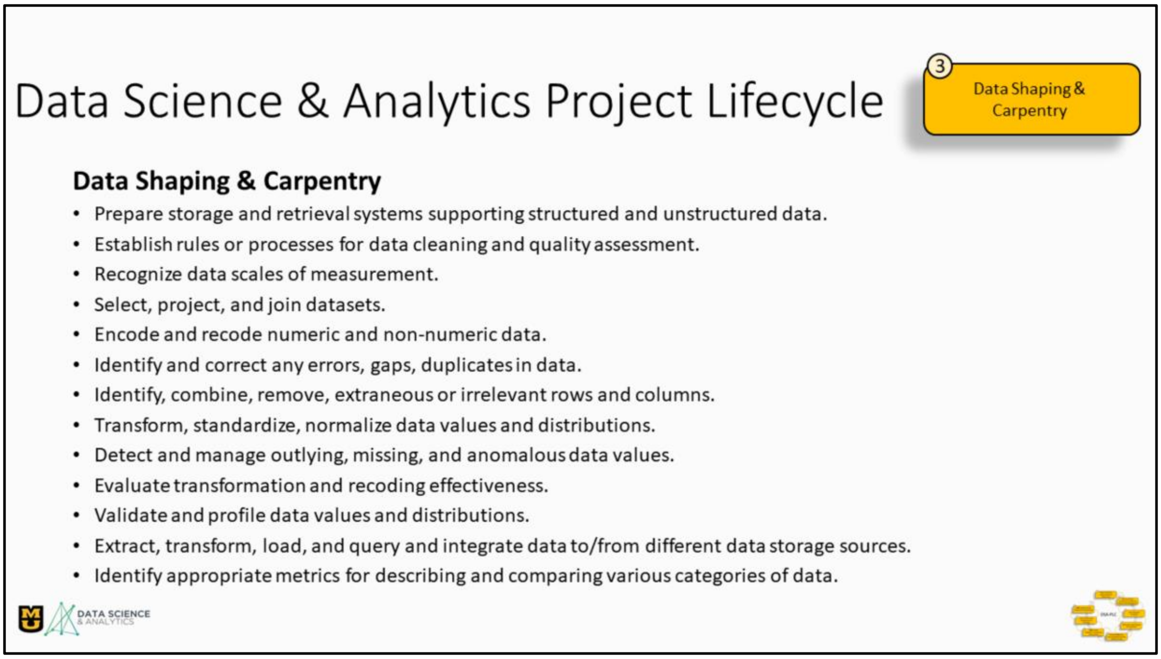
Data Acquisition, Governance, Curation & Source Code Management

- Act ethically, professionally, and responsibility in data governance, data and source code curation, and actions.
- Protect and maintain data using established security and privacy standards and best practices to minimize bias and enhance representativeness, reproducibility, and data provenance.
- Design, define, and establish techniques for data acquisition according to the features of defined data sources and data needs, and specified problem/solution space.
- Define approaches and tools for storing and retrieving volumes and varieties of data, including storage across a range of local and remote devices.
- Establish consistent naming, versioning, typing, and connections for directories, files, source code, system resources, programming languages, and DSA ecosystems.
- Design and implement source code management (SCM) system repositories and workflows.
- Identify, acquire, and integrate core, auxiliary and potential derived datasets (minimize disparate data silos).

Stage 2 deals particularly with how you will acquire your data, govern access and security/control, and manage data and source code. Here, it is important to address any laws, regulations, and codes of conduct applicable to our project. Consider the extent to which you ethically and legally can use the data for your intended use case. It is necessary to also address issues of personal privacy, anonymity, and other legal rights for individuals represented in the data that may impede or otherwise stall the project.

Describing the explicit source of any data, whether acquired internally or externally, and documenting its availability and validity for your intended use will help achieve ethical accountability the project. In addition, defining and implementing your data curation/management and source code management workflow and repository is important for maintaining project currency and fallback versioning if needed.

Overall, it is important to create and document necessary internal protocols including policies, procedures, checklists, and reviews to enforce ethical and proper source code and data usage.



Data Science & Analytics Project Lifecycle

3 Data Shaping & Carpentry

Data Shaping & Carpentry

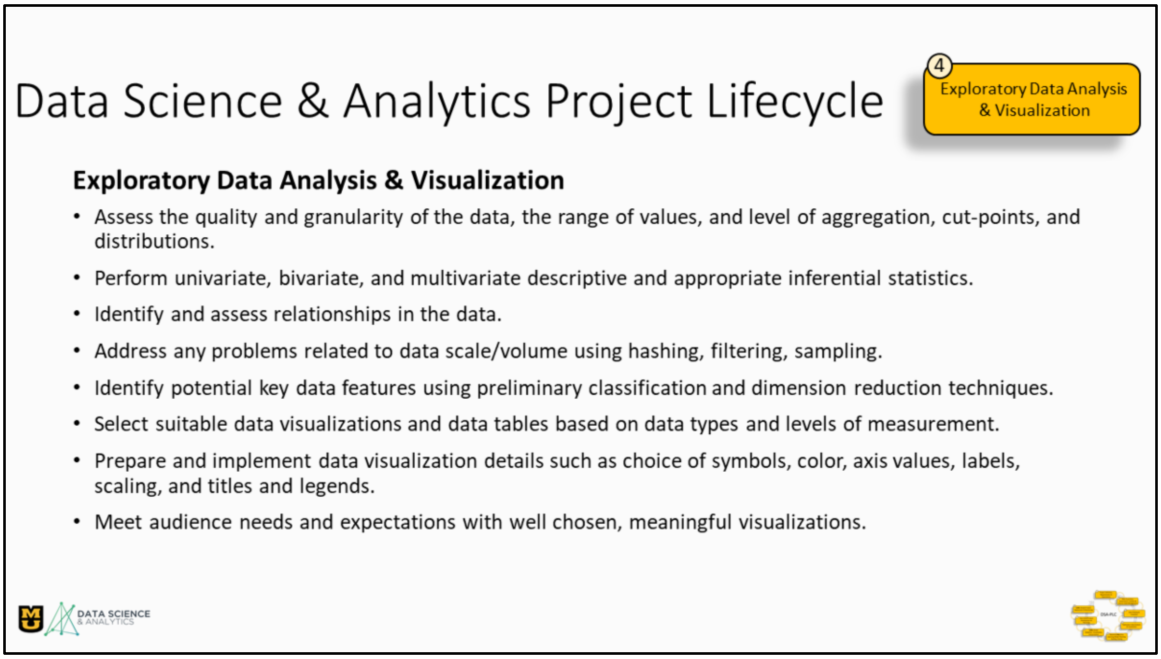
- Prepare storage and retrieval systems supporting structured and unstructured data.
- Establish rules or processes for data cleaning and quality assessment.
- Recognize data scales of measurement.
- Select, project, and join datasets.
- Encode and recode numeric and non-numeric data.
- Identify and correct any errors, gaps, duplicates in data.
- Identify, combine, remove, extraneous or irrelevant rows and columns.
- Transform, standardize, normalize data values and distributions.
- Detect and manage outlying, missing, and anomalous data values.
- Evaluate transformation and recoding effectiveness.
- Validate and profile data values and distributions.
- Extract, transform, load, and query and integrate data to/from different data storage sources.
- Identify appropriate metrics for describing and comparing various categories of data.

Once there is a relatively stable action plan from stage 1 and an overall shared understanding established from stage 2, it is now time to begin fashioning the data for use in other stages. Often, the tasks involved with the data shaping and carpentry are highly iterative and consume a large proportion of the overall project time and effort.

Never assume data is clean. Assume it is dirty unless proven otherwise. Assess the data from all possible perspectives and determine if it is clean and if not, can you clean the data and document how this is performed. Much of the methods used to successfully clean and shape the data are dependent on which scale of measurement underlies the data values and distributions. Remember there is significant value in preserving the raw data prior to any transforms, shaping and carpentry, so version wisely.

Familiarize yourself with the data thoroughly and take steps to condition the data especially when accessing high volumes and varieties of data for a Big Data project. Determine to what degree you have missing or inconsistent values, and if you have values deviating from normal or other known benchmarks; also look for any evidence of systematic error. This is also the time to join or merge different data sets, and subset your data based on the types of questions to be answered.

Without sufficient quality and quantity, or you cannot get access to good data, you will not be able to perform the subsequent steps in the lifecycle process. Getting deeply knowledgeable about the data will be critical when it comes time to plan and run your models later in the process.



Data Science & Analytics Project Lifecycle

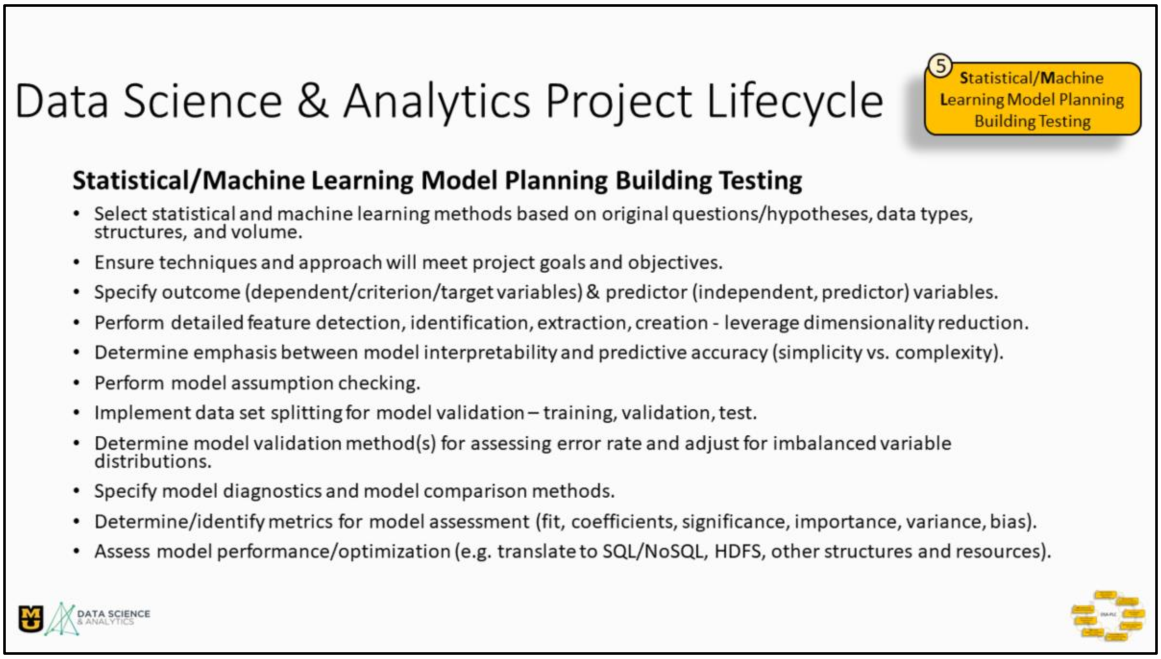
4
Exploratory Data Analysis & Visualization

Exploratory Data Analysis & Visualization

- Assess the quality and granularity of the data, the range of values, and level of aggregation, cut-points, and distributions.
- Perform univariate, bivariate, and multivariate descriptive and appropriate inferential statistics.
- Identify and assess relationships in the data.
- Address any problems related to data scale/volume using hashing, filtering, sampling.
- Identify potential key data features using preliminary classification and dimension reduction techniques.
- Select suitable data visualizations and data tables based on data types and levels of measurement.
- Prepare and implement data visualization details such as choice of symbols, color, axis values, labels, scaling, and titles and legends.
- Meet audience needs and expectations with well chosen, meaningful visualizations.

Once you have made progress with the necessary data shaping and carpentry to address any issues, performing exploratory data analysis can proceed, as can creating visualizations of your data in appropriate ways. Normally this stage is where basic descriptive and inferential statistical analysis is applied to the data in robust ways. This stage is also where, based on the data volume, high dimensionality, and available compute resources, sampling methods and dimension reduction techniques may need to be designed and applied.

In this stage, be sure to explore the data in ways to understand the relationships among the variables to inform selection of the variables to better relate the data back to the problem domain. It is also important to select suitable data visualizations – including plot types, themes, symbols, and color, and data table displays based on the data's inherent level of measurement and audience expectations.



Data Science & Analytics Project Lifecycle

5 Statistical/Machine Learning Model Planning Building Testing

Statistical/Machine Learning Model Planning Building Testing

- Select statistical and machine learning methods based on original questions/hypotheses, data types, structures, and volume.
- Ensure techniques and approach will meet project goals and objectives.
- Specify outcome (dependent/criterion/target variables) & predictor (independent, predictor) variables.
- Perform detailed feature detection, identification, extraction, creation - leverage dimensionality reduction.
- Determine emphasis between model interpretability and predictive accuracy (simplicity vs. complexity).
- Perform model assumption checking.
- Implement data set splitting for model validation – training, validation, test.
- Determine model validation method(s) for assessing error rate and adjust for imbalanced variable distributions.
- Specify model diagnostics and model comparison methods.
- Determine/identify metrics for model assessment (fit, coefficients, significance, importance, variance, bias).
- Assess model performance/optimization (e.g. translate to SQL/NoSQL, HDFS, other structures and resources).

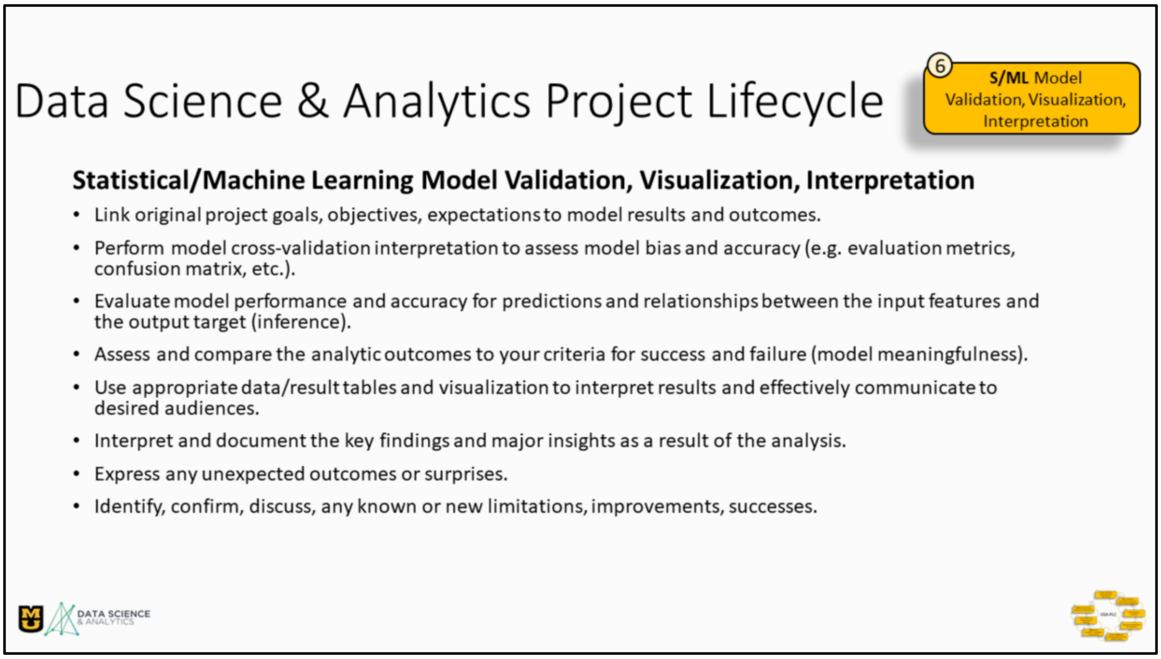
DATA SCIENCE & ANALYTICS

In stage 5, aim for capturing the most essential predictor variables, rather than considering every possible variable that you think may influence or be related to the outcome or target variable.

Like stage 3 – data shaping and carpentry, this stage will be very iterative using plan, build, and test cycles in order to identify the most essential variables for the analyses and models you select. Developing a general methodology, solid understanding of the variables and techniques to use, and preparing a description and diagram of the analytic workflow will facilitate documenting the history and outcomes of this stage. It is also good practice to document fully how you have split your data for model validation as well as for interpretability and accuracy.

Stage 5 represents the last step of preparations before executing the analytical models and requires you to be thorough in planning the analytical work and experiments in the next stage.

This is the time to refer back to the outcomes from stages 1-3 particularly the hypotheses or questions developed in Stage 1 based on the business problems or domain area. These hypotheses will help you frame the analytics you'll execute in stage 6, and choose the right methods to achieve your objectives. Also be sure to consider alternative approaches during this stage.   Many times you can get additional ideas from analogous problems people have solved in different industry verticals – which may require additional literature searches. Using sound statistical and machine learning model planning will help ensure that the chosen analytical techniques will enable you to meet the business objectives and prove or disprove the working hypotheses.



Data Science & Analytics Project Lifecycle

6
S/ML Model Validation, Visualization, Interpretation

Statistical/Machine Learning Model Validation, Visualization, Interpretation

- Link original project goals, objectives, expectations to model results and outcomes.
- Perform model cross-validation interpretation to assess model bias and accuracy (e.g. evaluation metrics, confusion matrix, etc.).
- Evaluate model performance and accuracy for predictions and relationships between the input features and the output target (inference).
- Assess and compare the analytic outcomes to your criteria for success and failure (model meaningfulness).
- Use appropriate data/result tables and visualization to interpret results and effectively communicate to desired audiences.
- Interpret and document the key findings and major insights as a result of the analysis.
- Express any unexpected outcomes or surprises.
- Identify, confirm, discuss, any known or new limitations, improvements, successes.

In the statistical and machine learning model validation, visualization and interpretation stage, the designed model is fit on the training data and evaluated or scored against the test data. Generally, this work takes place in a controlled computation environment and not in the live production environment.

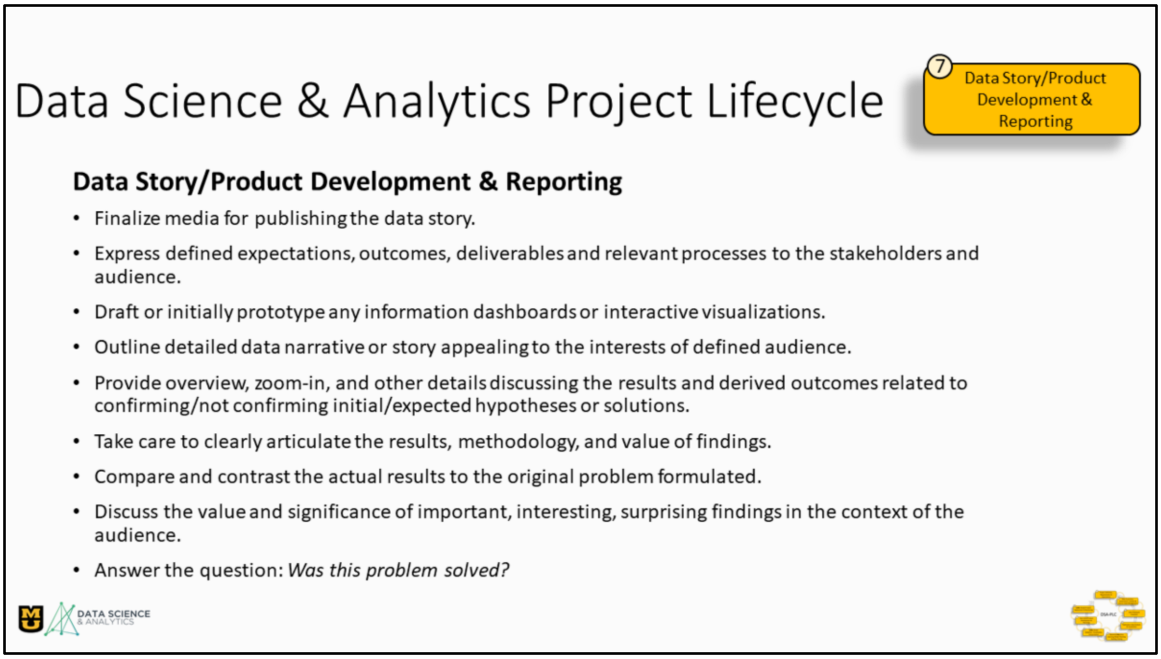
Here you apply the desired model to data sets for model training, testing, and validation to eventually prepare for reporting and production purposes. Depending on the selected models, the computational environment planned back in stage 1 for executing models and workflows can enhance performance and aid in minimizing wait times for tasks to run and execute.

Although the modeling techniques and logic required in this step can be highly complex and very iterative, the actual duration of this stage can be quite short compared to the other stages.

We can think of this approach as iterative in nature by executing the models defined in stage 5 and assessing the validity of the model results. Based on the results, we then fine-tune the model parameters to further optimize outcome performance. We also want to determine if the parameter values of the fitted model make sense in the context of the problem domain and if more data or more inputs are needed, or if inputs should be transformed or eliminated.

When assessing the model results we want to consider whether the model looks valid and accurate when executed against the test data and that output and model behavior makes sense to the domain experts and other stakeholders. That is, does it look like the model is giving “the right answers”, or answers that make sense in the defined problem domain and are the model results accurate enough to meet the goals and objectives of this project?

At the end of this stage, we need to know if this model achieves our desired outcomes as defined or if a different form of model is needed. If a new model is needed, you’ll return to the Model Planning stage 5 and revise the modeling approach.



Data Science & Analytics Project Lifecycle

7 Data Story/Product Development & Reporting

Data Story/Product Development & Reporting

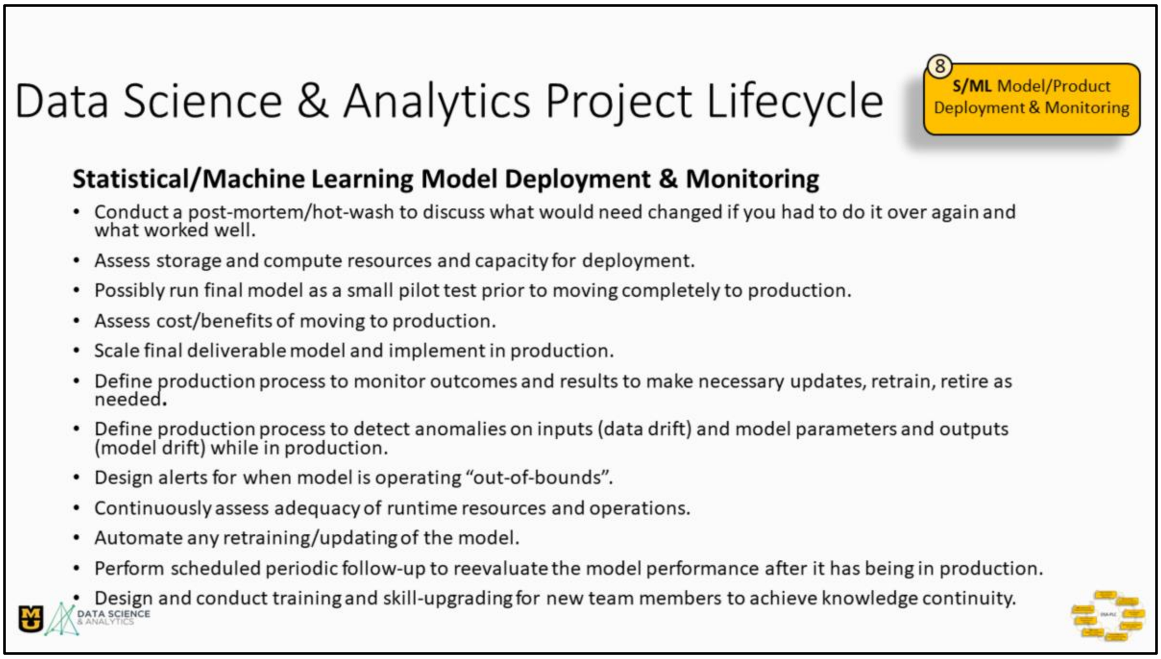
- Finalize media for publishing the data story.
- Express defined expectations, outcomes, deliverables and relevant processes to the stakeholders and audience.
- Draft or initially prototype any information dashboards or interactive visualizations.
- Outline detailed data narrative or story appealing to the interests of defined audience.
- Provide overview, zoom-in, and other details discussing the results and derived outcomes related to confirming/not confirming initial/expected hypotheses or solutions.
- Take care to clearly articulate the results, methodology, and value of findings.
- Compare and contrast the actual results to the original problem formulated.
- Discuss the value and significance of important, interesting, surprising findings in the context of the audience.
- Answer the question: *Was this problem solved?*

DATA SCIENCE & ANALYTICS

Fundamentally, stage 7 – Data Story/Product Development and Reporting is the communications stage.

The outcome from this stage can take many forms such as a presentation, document brief, infographic, traditional research paper, and even an interactive visualization. The deliverable will be a highly visible component of the project made available to the internal and external organizational stakeholders and sponsors. Take care to clearly articulate the results, methodology, business value of your findings, and moreover, how you met the criteria for success and answered the important questions defined in stage 1. Consider how best to articulate the findings and outcome to the various team members and stakeholders.

Determine the level of success or failure, identify your key findings, quantify the business value and develop a narrative to summarize your methods and findings and convey to stakeholders. Choose your media and frame the narrative in a way that is appropriate for the audience and demonstrates clear value of how the project goals and objectives have been accomplished. Even the most technically accurate analysis that cannot be translated into a language that speaks to the audience may easily miss the point and not clearly present the real value of the project. If this happens, much of your time and effort will have been wasted.



Data Science & Analytics Project Lifecycle

8 **S/ML** Model/Product Deployment & Monitoring

Statistical/Machine Learning Model Deployment & Monitoring

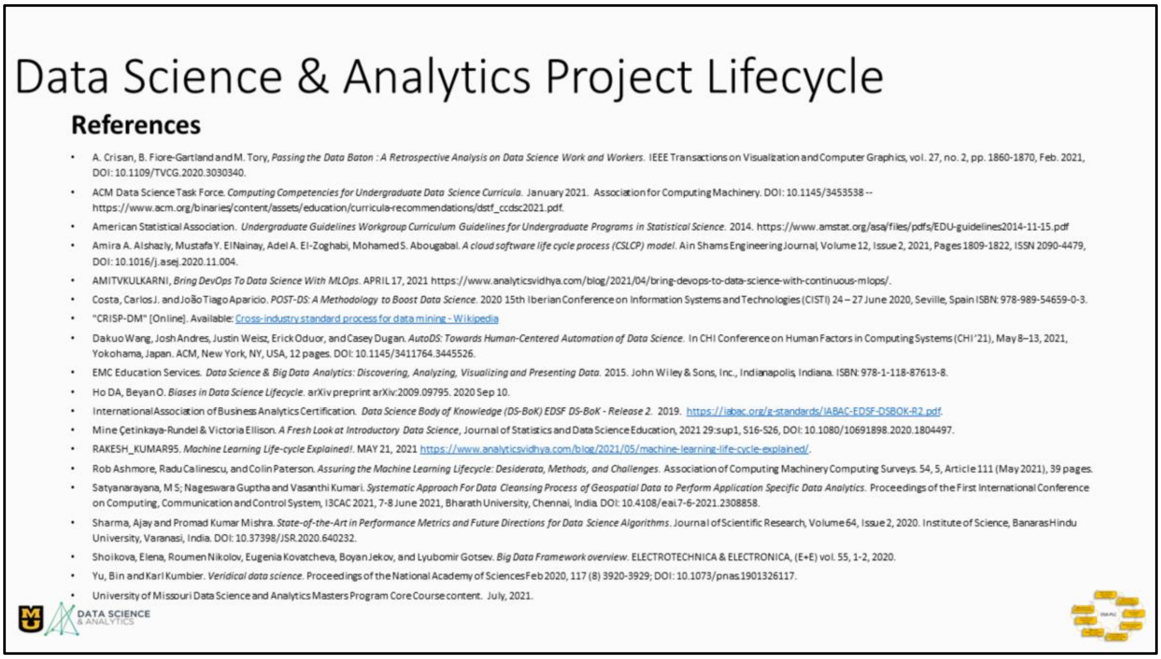
- Conduct a post-mortem/hot-wash to discuss what would need changed if you had to do it over again and what worked well.
- Assess storage and compute resources and capacity for deployment.
- Possibly run final model as a small pilot test prior to moving completely to production.
- Assess cost/benefits of moving to production.
- Scale final deliverable model and implement in production.
- Define production process to monitor outcomes and results to make necessary updates, retrain, retire as needed.
- Define production process to detect anomalies on inputs (data drift) and model parameters and outputs (model drift) while in production.
- Design alerts for when model is operating "out-of-bounds".
- Continuously assess adequacy of runtime resources and operations.
- Automate any retraining/updating of the model.
- Perform scheduled periodic follow-up to reevaluate the model performance after it has being in production.
- Design and conduct training and skill-upgrading for new team members to achieve knowledge continuity.

When you have reached the Deployment and Monitoring stage, your project has performed well in terms of process and outcomes. However, the project is not finished, it fact, a completely new set of iterative tasks await. By following a systematic process, as presented here, you increase the likelihood a project can be reach this stage and put into production.

To begin this stage, perform a post-mortem or review meeting with team and stakeholders to discuss all the details of the project identifying what worked well, not so well, and what needs to be changed

Rather than try to perform a full cutover of the current project to production, consider how to implement incrementally your model or product in a series of pilot projects in the production environment. This will help to maintain scope, minimize risk, and gain insight into infrastructure and performance needs. Doing incremental deployment will allow detection of any input or parameter anomalies and gauge any resource deficiencies and infrastructure dependencies. Design and implement any alerts for values considered out-of-bounds and automate as much operations as possible.

After deployment, conduct technical and result follow up to reevaluate the model. Assess whether the model is continuing to meet goals, objectives and expectations. If these outcomes are not occurring, determine if this is due to a technical issues, model inaccuracy, or if its predictions are not being acted on appropriately.
This is also the time to decide on and document what constitutes project obsolescence and how to retire or sunset this production level project.



Data Science & Analytics Project Lifecycle

References

- A. Crisan, B. Fiore-Gartland and M. Tory, *Passing the Data Baton : A Retrospective Analysis on Data Science Work and Workers.* IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 2, pp. 1860-1870, Feb. 2021, DOI: 10.1109/TVCG.2020.3030340.
- ACM Data Science Task Force. *Computing Competencies for Undergraduate Data Science Curricula.* January 2021. Association for Computing Machinery. DOI: 10.1145/3453538 -- https://www.acm.org/binaries/content/assets/education/curricula-recommendations/dstf_ccdsc2021.pdf.
- American Statistical Association. *Undergraduate Guidelines Workgroup Curriculum Guidelines for Undergraduate Programs in Statistical Science.* 2014. https://www.amstat.org/asa/files/pdfs/EDU-guidelines2014-11-15.pdf
- Amira A. Alshazly, Mustafa Y. ElNainay, Adel A. El-Zoghabi, Mohamed S. Abougabal. *A cloud software life cycle process (CSLCP) model.* Ain Shams Engineering Journal, Volume 12, Issue 2, 2021, Pages 1809-1822, ISSN 2090-4479, DOI: 10.1016/j.asej.2020.11.004.
- AMITVKULKARNI, *Bring DevOps To Data Science With MLOps.* APRIL 17, 2021 https://www.analyticsvidhya.com/blog/2021/04/bring-devops-to-data-science-with-continuous-mlops/.
- Costa, Carlos J. and João Tiago Aparicio. *POST-DS: A Methodology to Boost Data Science.* 2020 15th Iberian Conference on Information Systems and Technologies (CISTI) 24 – 27 June 2020, Seville, Spain ISBN: 978-989-54659-0-3.
- "CRISP-DM" [Online]. Available: Cross-industry standard process for data mining - Wikipedia
- Dakuo Wang, Josh Andres, Justin Weisz, Erick Oduor, and Casey Dugan. *AutoDS: Towards Human-Centered Automation of Data Science.* In CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 12 pages. DOI: 10.1145/3411764.3445526.
- EMC Education Services. *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data.* 2015. John Wiley & Sons, Inc., Indianapolis, Indiana. ISBN: 978-1-118-87613-8.
- Ho DA, Beyan O. *Biases in Data Science Lifecycle.* arXiv preprint arXiv:2009.09795. 2020 Sep 10.
- International Association of Business Analytics Certification. *Data Science Body of Knowledge (DS-BoK) EDSF DS-BoK - Release 2.* 2019. https://iabac.org/g-standards/IABAC-EDSF-DSBOK-R2.pdf.
- Mine Çetinkaya-Rundel & Victoria Ellison. *A Fresh Look at Introductory Data Science*, Journal of Statistics and Data Science Education, 2021 29:sup1, S16-S26, DOI: 10.1080/10691898.2020.1804497.
- RAKESH_KUMAR95. *Machine Learning Life-cycle Explained!.* MAY 21, 2021 https://www.analyticsvidhya.com/blog/2021/05/machine-learning-life-cycle-explained/.
- Rob Ashmore, Radu Calinescu, and Colin Paterson. *Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges.* Association of Computing Machinery Computing Surveys. 54, 5, Article 111 (May 2021), 39 pages.
- Satyanarayana, M S; Nageswara Guptha and Vasanthi Kumari. *Systematic Approach For Data Cleansing Process of Geospatial Data to Perform Application Specific Data Analytics.* Proceedings of the First International Conference on Computing, Communication and Control System, I3CAC 2021, 7-8 June 2021, Bharath University, Chennai, India. DOI: 10.4108/eai.7-6-2021.2308858.
- Sharma, Ajay and Promad Kumar Mishra. *State-of-the-Art in Performance Metrics and Future Directions for Data Science Algorithms.* Journal of Scientific Research, Volume 64, Issue 2, 2020. Institute of Science, Banaras Hindu University, Varanasi, India. DOI: 10.37398/JSR.2020.640232.
- Shoikova, Elena, Roumen Nikolov, Eugenia Kovatcheva, Boyan Jekov, and Lyubomir Gotsev. *Big Data Framework overview.* ELECTROTECHNICA & ELECTRONICA, (E+E) vol. 55, 1-2, 2020.
- Yu, Bin and Karl Kumbier. *Veridical data science.* Proceedings of the National Academy of Sciences Feb 2020, 117 (8) 3920-3929; DOI: 10.1073/pnas.1901326117.
- University of Missouri Data Science and Analytics Masters Program Core Course content. July, 2021.