Ministry of Higher Education and Scientific Research
Djilali BOUNAAMA University - Khemis Miliana(UDBKM)
Faculty of Science and Technology
Department of Mathematics and Computer Science

Chapter 2

Data Science Pipeline

AIBD-M1-UEM112 : Introduction to Data Science

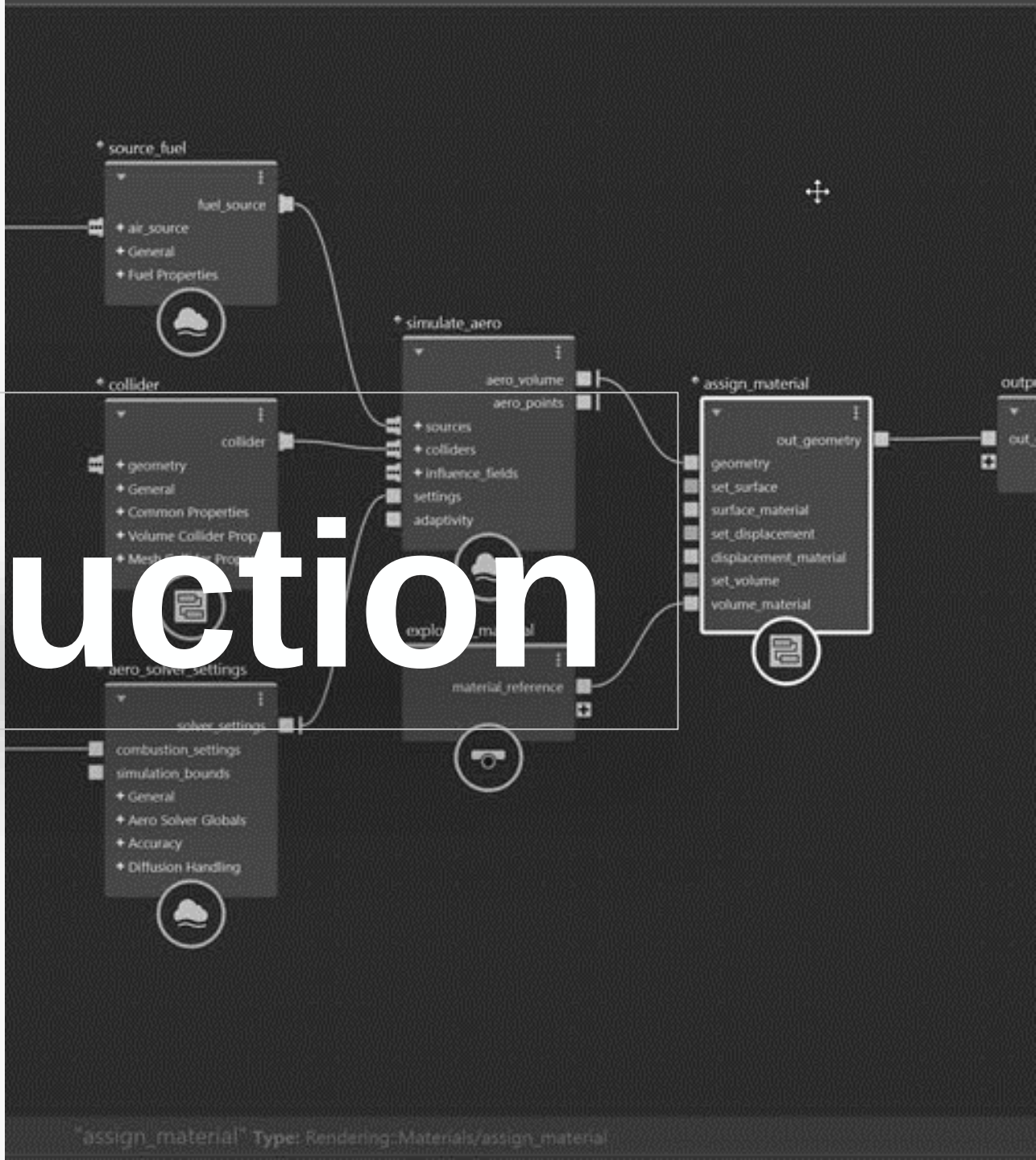
Noureddine AZZOUZA

n.azzouza@univ-dbkm.dz

Course Topics

1. Introduction
2. Data Science Pipelines
3. Types
4. Stages
3. References

Introduction

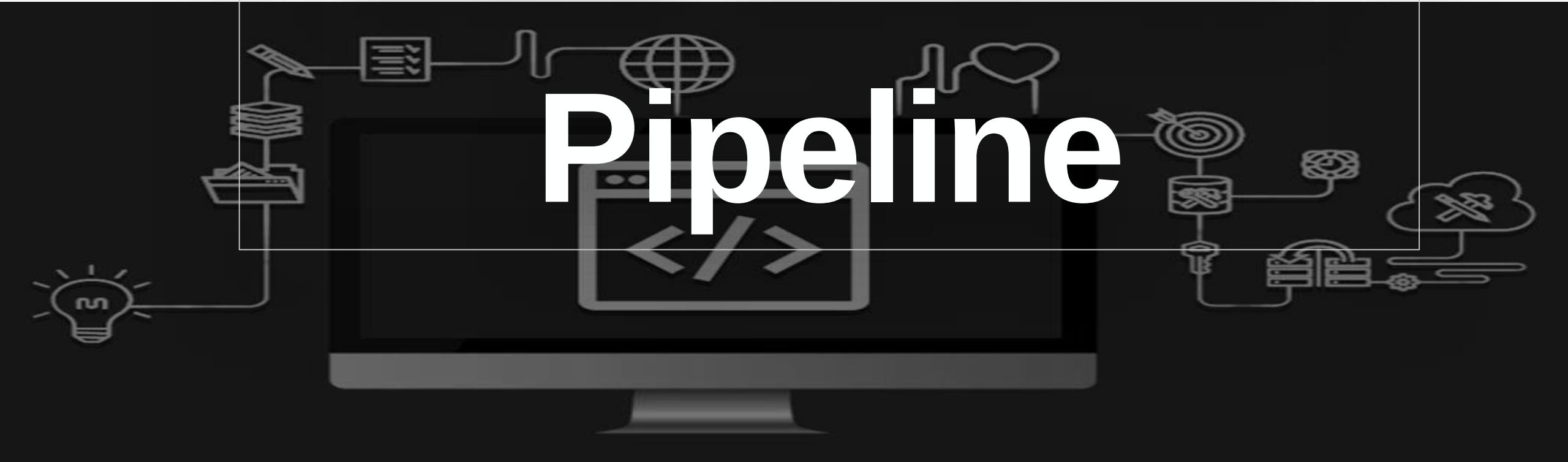


Introduction

- ✓ Data science is quickly evolving to be one of the hottest fields in the technology industry.
- ✓ With rapid advancements in computational performance, we can uncover patterns and insights about user behavior and world trends to an unprecedented extent.
- ✓ With the influx of buzzwords in the field of data science and relevant fields, a common question is “Data science sounds pretty cool - how to get started?”
- ✓ Here is a brief overview of steps that make up a data science lifecycle / Pipeline. For each step, we provide some useful resources.

Data Science

Pipeline



Definition

- ✓ ***Data science processes***, also called data science stages as in stages of a pipeline, for descriptive, predictive, and prescriptive analytics are becoming integral components of many software systems today.
- ✓ The data science stages are organized into a ***data science pipeline***, where data might flow from one stage in the pipeline to the next
- ✓ These data science stages generally perform different tasks such as ***data acquisition, data preparation, storage, feature engineering, modeling, training, evaluation of the machine learning model***, etc.

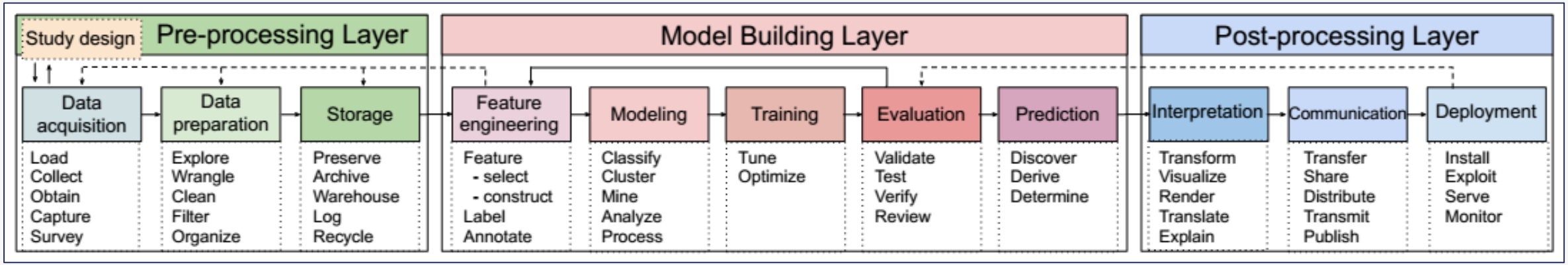
Terminology

- ✓ Terms like :
  - ➢ “data science pipeline”,
  - ➢ “machine learning pipeline”,
  - ➢ “big data lifecycle”,
  - ➢ “deep learning workflow”,
- ✓ With permutation of these keywords, usually refers to the “***data science pipeline”*** concept (with some minor differences).

Definition

- ✓ The term pipeline was introduced by Garlan with box-and-line diagrams and explanatory prose that assist software developers to design and describe complex systems.
- ✓ By data science pipeline (DS pipeline), we are referring to a series of processing stages that interact with data.
- ✓ The stages are defined to perform particular tasks and connected to other stage(s) with input-output relations

Stages of Data Science Pipeline

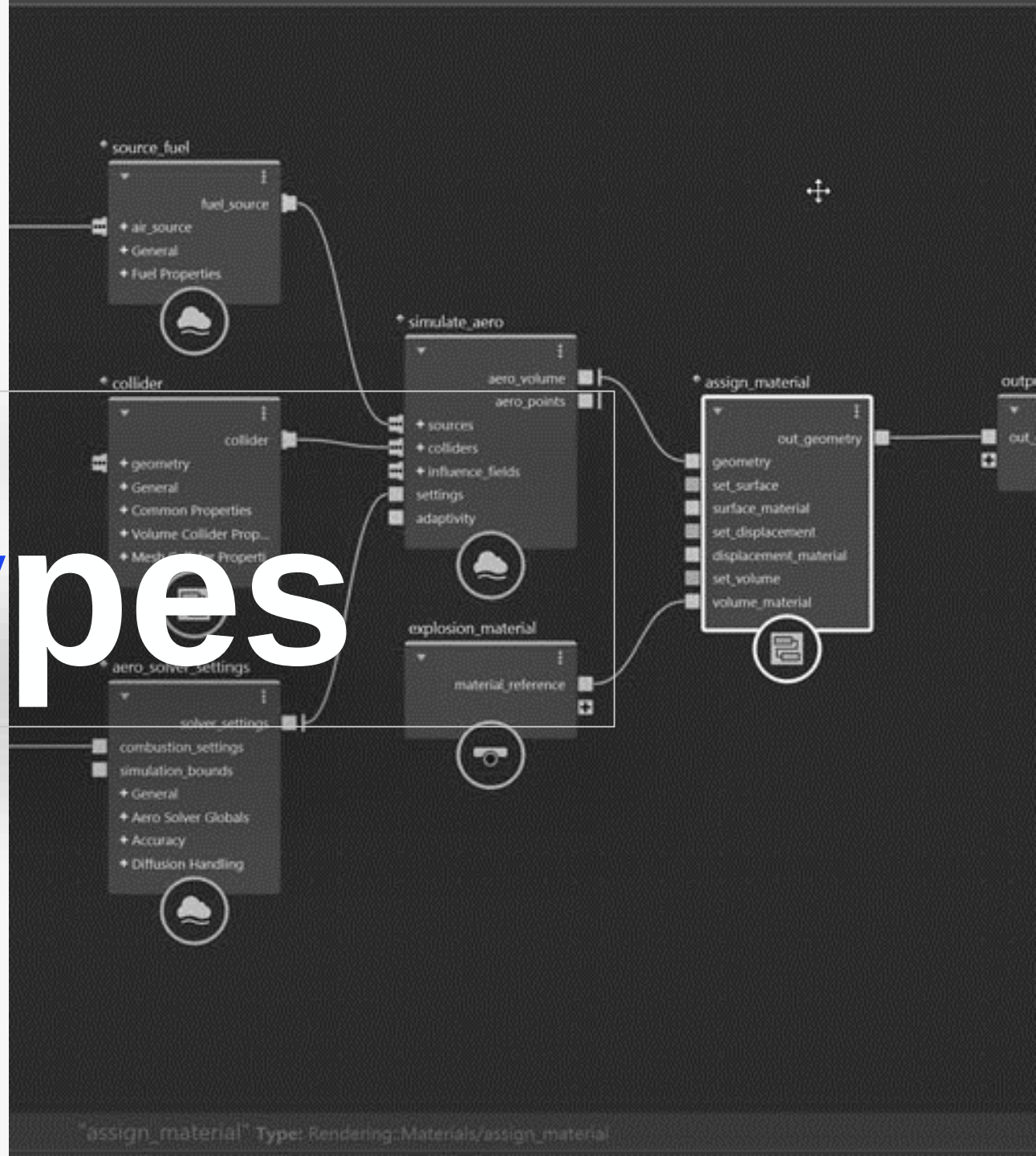


- ✓ The stages are connected with feedback loops denoted with arrows.
- ✓ The sub-tasks are listed below each stage.
- ✓ Solid arrows are always present in the lifecycle, while the dashed arrows are optional.
- ✓ Distant feedback loops are also possible through intermediate stage(s)

Stages of Data Science Pipeline

- ✓ This DS pipeline is represented with 3 layers, 11 stages
- ✓ the preprocessing layer, the stages are data acquisition, preparation, and storage (appeared in team process pipelines)
- ✓ The algorithmic steps and data processing are done in the model building layer. Modeling does not necessarily imply the existence of an ML component, since DS can involve custom data processing or statistical modeling.
- ✓ Post-processing layer includes the tasks that take place after the results have been generated.

Types


source_fuel
fuel_source
air_source
General
Fuel Properties
collider
collider
geometry
General
Common Properties
Volume Collider Prop...
simulate_aero
aero_volume
aero_points
sources
colliders
influence_fields
settings
adaptivity
explosion_material
material_reference
assign_material
out_geometry
geometry
set_surface
surface_material
set_displacement
displacement_material
set_volume
volume_material
aero_solver_settings
solver_settings
combustion_settings
simulation_bounds
General
Aero Solver Globals
Accuracy
Diffusion Handling
"assign_material" Type: Rendering::Materials/assign_material

Types of Data Science Pipeline

- ✓ there is no standard methodology to develop comparable and inter-operable DS pipelines.
- ✓ Pipeline in the literature can be classified into three types of DS pipelines :
  1. Machine Learning process,
  2. Big data management process,
  3. team process.

Machine Learning process

- ✓ Most of the pipelines in the literature are describing machine learning processes (about 46% of all). The recent advent of AI and DL has led to more DS systems that involve ML components.
- ✓ The pipelines in this category emphasize the algorithmic process, learning patterns, and building predictive models.
- ✓ However, the post-processing stages are rare in these type of pipelines.
- ✓ incorporating the post-processing stages would be desired to ensure safe real-world deployment of such pipelines.

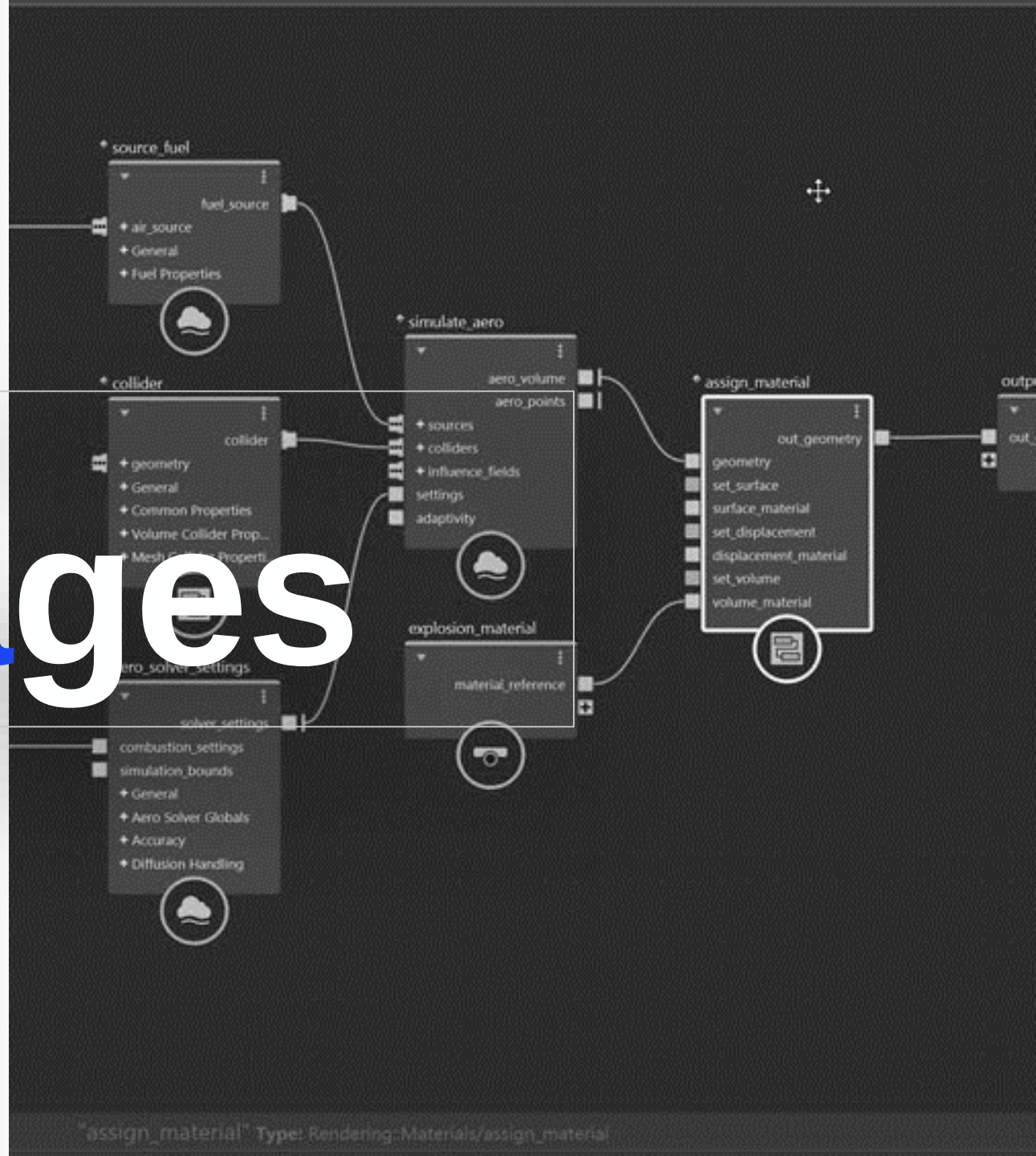
Big Data Management

- ✓ The references in this category present DS pipelines that
  - ➢ manage a large amount of data or
  - ➢ describes a framework (software-hardware) for data processing
- ✓ Do not contain machine learning components in the pipeline.
- ✓ Processing large amount data often requires specific algorithms and engineering methods for efficiency and further processing.
- ✓ About 18% of existing pipelines fall in this category.

Team process

- ✓ These pipelines describe workflow of human activities that needs to be followed in a DS pipeline.
- ✓ These pipelines present a high-level view for building DS component in a team environment.
- ✓ The data science teams require specific expertise and management to build successful DS pipelines
- ✓ Do not describe a DS software architecture.

Stages


source_fuel
fuel_source
air_source
General
Fuel Properties
collider
collider
geometry
General
Common Properties
Volume Collider Prop...
simulate_aero
aero_volume
aero_points
sources
colliders
influence_fields
settings
adaptivity
explosion_material
material_reference
assign_material
out_geometry
geometry
set_surface
surface_material
set_displacement
displacement_material
set_volume
volume_material
combustion_settings
simulation_bounds
General
Aero Solver Globals
Accuracy
Diffusion Handling
"assign_material" Type: Rendering/Materials/assign_material

1. Data Acquisition (ACQ)

- ✓ In the beginning of DS pipeline
- ✓ Data are collected from appropriate sources.
- ✓ Data can be acquired manually or automatically.
- ✓ Data acquisition also involves
  - ➢ understanding the nature of the data,
  - ➢ collecting relevant data, and
  - ➢ integrating available datasets

2. Data Preparation (PRP)

- ✓ Data are generally acquired in a raw format that needs certain preprocessing steps.
- ✓ This involves exploration and filtering, which helps identify the correct data for further processing.
- ✓ Well prepared data reduces the time required for data analysis and contributes to the success of the DS pipeline.

3. Storage (STR)

- ✓ It is important to find an appropriate hardware-software combination to preserve data so that it can be processed efficiently.
  - ➢ For example, Miao et al. used graph database system Neo4j to build a collaborative analysis pipeline since Neo4J supports querying graph data properties.

4. Feature Engineering (FTR)

- ✓ The entire dataset might not contribute equally to decision making.
- ✓ In this stage, appropriate features that are useful to build the model are identified or constructed.
- ✓ Features that are not readily available in the dataset, require engineering to create them from raw data

5. Modeling (MDL)

- ✓ When data are preprocessed and features are extracted, a model is built to analyze the data.
- ✓ Model building includes :
  - ➢ model planning,
  - ➢ model selection,
  - ➢ mining and deriving important properties of data.
- ✓ Appropriate data processing strategies and algorithms are selected to create a good model.

6. Training (TRN)

- ✓ For a specific model, we need to train the model with available labeled data.
- ✓ By each training iteration, we optimize the model and try to make it better.
- ✓ The quality of the training dataset contributes to the training accuracy of the model.

7. Evaluation (EVL)

- ✓ After training the model, it is tested with a new dataset which has not been used as training data.
- ✓ Also, the model can be evaluated in real-life scenarios and compared with other competing models.
- ✓ Existing metrics are used or new metrics are created to evaluate the model.

8. Prediction (PRD)

- ✓ The success of the model depends on how good a model can predict in an unknown setup.
- ✓ After a satisfactory evaluation, we employ the model to solve the problem and see how it works.
- ✓ There are many prediction metrics such as classification accuracy, log-loss, F1-score, to measure the success of the model.

9. Interpretation (INT)

- ✓ The prediction result might not be enough to make a decision.
- ✓ We often need a transformation of the prediction result and post-processing to translate predictions into knowledge.
- ✓ For example, only numerical results do not help much but a good visualization can help to make a decision.

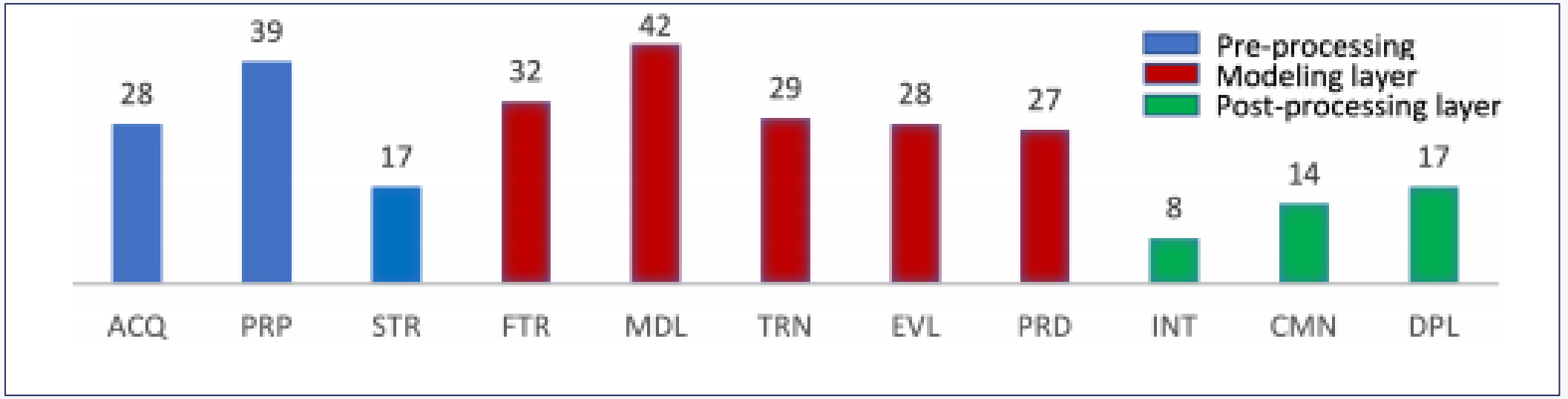
10. Communication (CMN)

- ✓ Different components of the DS system might reside in a distributed environment.
- ✓ So, we might need to communicate with the involved parties (e.g., devices, persons, systems) to share and accumulate information.
- ✓ Communication might take place in different geographical locations or the same.

11. Deployment (DPL)

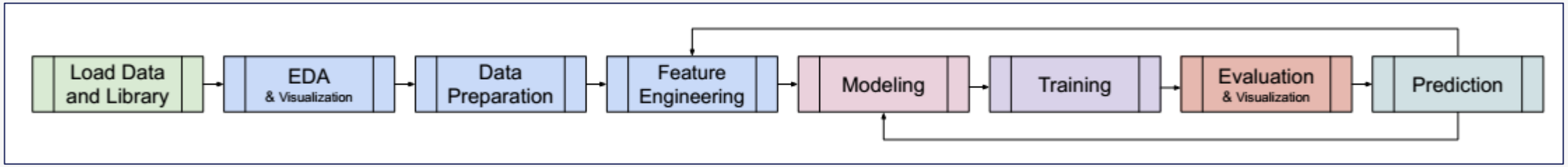
- ✓ The built DS solution is installed in its problem domain to serve the application.
- ✓ Over time, the performance of the model is monitored so that the model can be improved to handle new situations.
- ✓ Deployment also includes model maintenance and sending feedback to the model building layer.

Frequency of pipeline stages in theory



- ✓ Post-processing layers are included infrequently (52%) compared to pre-processing (96%) and model building (96%) layers of pipelines in theory.

Data Science Pipeline in small projects



- ✓ Among the 11 pipeline stages described before, we found only 6 –or 8- stages in the DS pipeline for small projects.
- ✓ Other stages (e.g., storage, interpretation, communication, deployment) are not found in these programs because these stages occur while building a production-scale large DS system and often not present in the DS notebooks.

Data Science Pipeline in small projects

- ✓ Data preparation can occur before or after all other stages.
- ✓ Apart from that, data acquisition is followed by data preparation most of the time, which in turn is followed by modeling.
- ✓ Modeling is followed mostly by training, which in turn is followed by prediction.
- ✓ Evaluation is mostly surrounded by prediction and data preparation

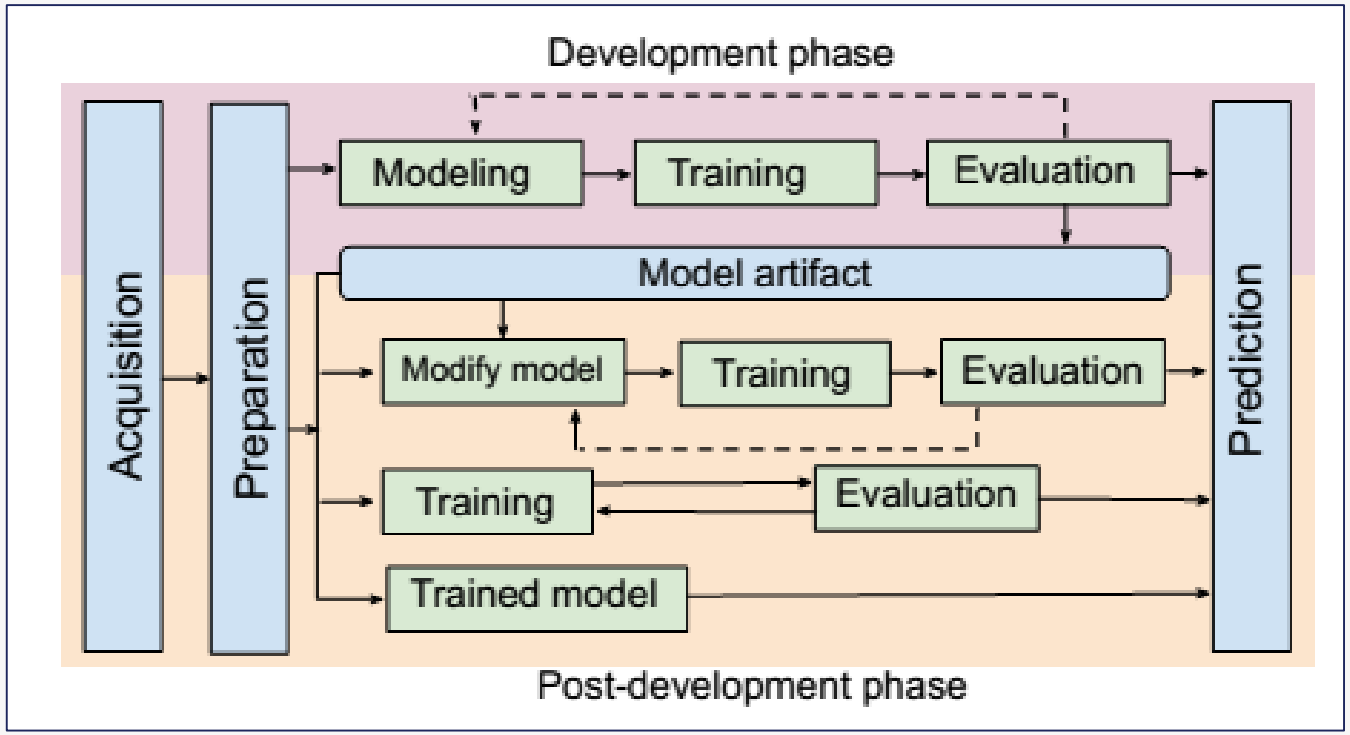
Data Science Pipeline in small projects

- ✓ Data preparation tasks (e.g., formatting, reshaping, sorting) are not limited to just before the modeling stage, rather it is done on a whenever-needed basis.
  - ➢ For example, (code snippet from a Kaggle competition), while creating model-layers, data preprocessing API has been called in line 2

```
1 x = Conv2D(mid, (4, 1), activation='relu', padding='valid')(x)
2 x = Reshape((branch_model.output_shape[1], mid, 1))(x)
3 x = Conv2D(1, (1, mid), activation='linear', padding='valid')(x)
4 x = Flatten(name='flatten')(x)
5 head_model = Model([xa_inp, xb_inp], x, name='head'
```

Data Science Pipeline in large projects

- ✓ Development phase (top) runs during model building
- ✓ post-development phase (bottom) runs for making prediction

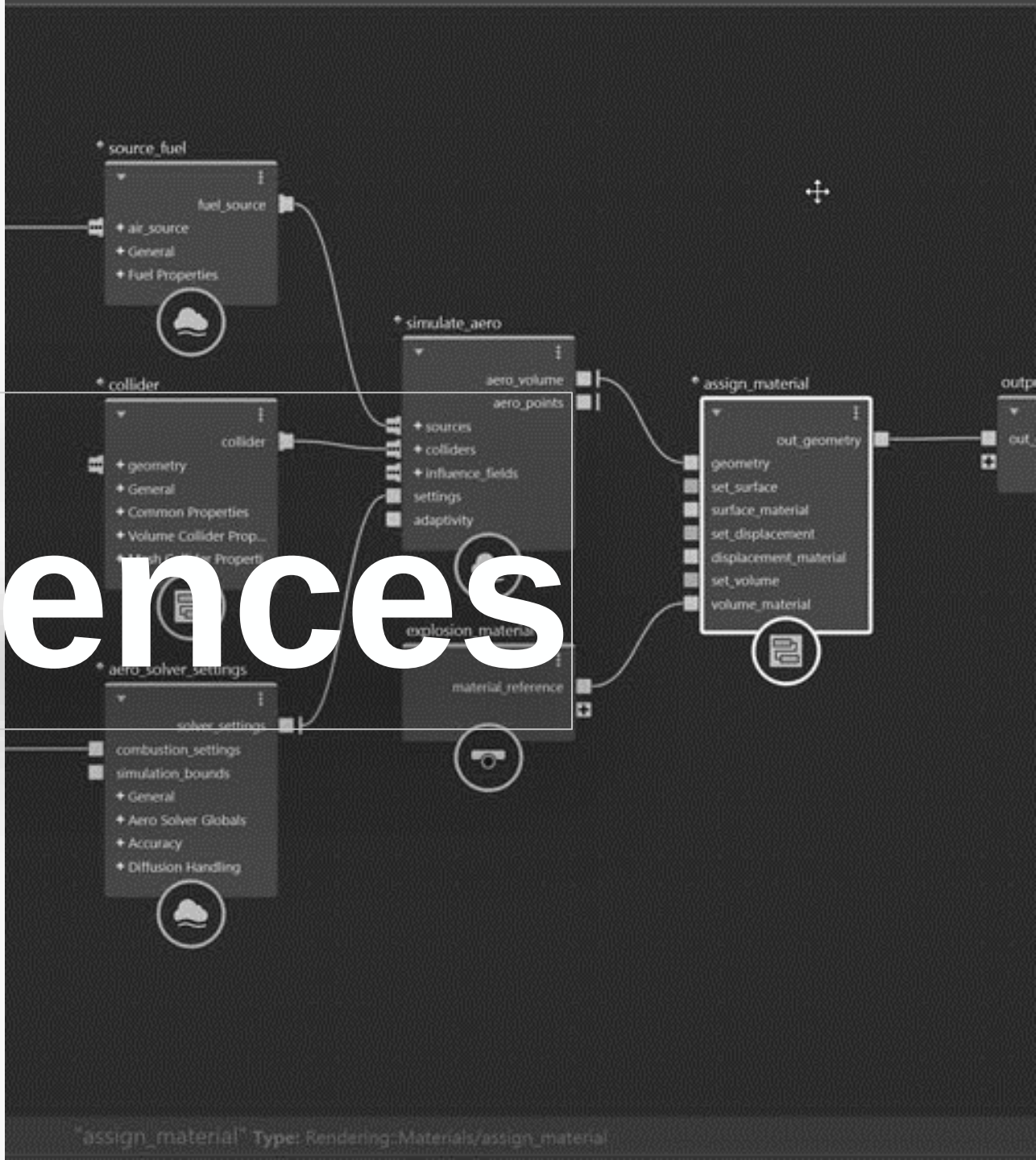


Data Science Pipeline in large projects

- ✓ Each of the projects contains 6 stages
- ✓ since the projects are not coupled to a specific dataset and they solve a more general problem, the projects are not limited to one single pipeline.

1. ***development phase*** : the main goal is to build a model that solves the problem in general. After completing, the final model is created and saved as an artifact. the user modify and exploit the model in the post-development phase.
2. ***post-development phase***: he users access the pre-built model and use that for prediction. the users can download the pre-trained model and at the end, the prediction result is obtained

References


source_fuel
fuel_source
air_source
General
Fuel Properties
collider
collider
geometry
General
Common Properties
Volume Collider Prop...
simulate_aero
aero_volume
aero_points
sources
colliders
influence_fields
settings
adaptivity
explosion_materia
material_reference
assign_material
out_geometry
geometry
set_surface
surface_material
set_displacement
displacement_material
set_volume
volume_material
aero_solver_settings
solver_settings
combustion_settings
simulation_bounds
General
Aero Solver Globals
Accuracy
Diffusion Handling
"assign_material" Type: Rendering::Materials/assign_material

Ministry of Higher Education and Scientific Research
Djilali BOUNAAMA University - Khemis Miliana(UDBKM)
Faculty of Science and Technology
Department of Mathematics and Computer Science

Chapter 2

Data Science Pipeline

AIBD-M1-UEM112 : Introduction to Data Science

Noureddine AZZOUZA

n.azzouza@univ-dbkm.dz