How SPASE Can Facilitate Data Science?

Shing F. Fung

ITM Physics Laboratory

NASA Goddard Space Flight Center

&

ISWAT O2-02 Team

Data Science in Heliophysics & Space Weather Research: A Few Examples

- NASA ROSES 2019 (NNH19ZDA001N-HTMS) for Theory, Modeling and Simulati becomes more affordable and more available…to solve the fundamental prob (AI)…, machine learning (ML), techniques have become potentially effective m and...can lead the way to new understanding and drive science concepts for fu

NASA research program supporting AI/ML

- McGranaghan, R. M., Bhatt, A., Matsuo, T., Mannucci, A. J., frontier in geospace through data science, Journal of Geopl https://doi.org/10.1002/2017JA024835

Data science in Heliophysics advocacy

- McGranaghan, R. M., Mannucci, A. J., Wilson, B. D., Mattmann, C. A., & Chadwick, R prediction of high-latitude ionospheric scintillation: A novel approach with machine 1817–1846. https://doi.org/10.1029/2018SW002018
- Galvez et al. (2019), A Machine-learning Data Set Prepared from the NASA Solar Dy Astrophys. J. Supplement Series, 242:7 (11pp), https://doi.org/10.3847/1538-4365/

Machine learning applications in Heliophysics research

- NASA Frontier Development Lab (FDL) 2020

NASA public-private collab. research Initiative

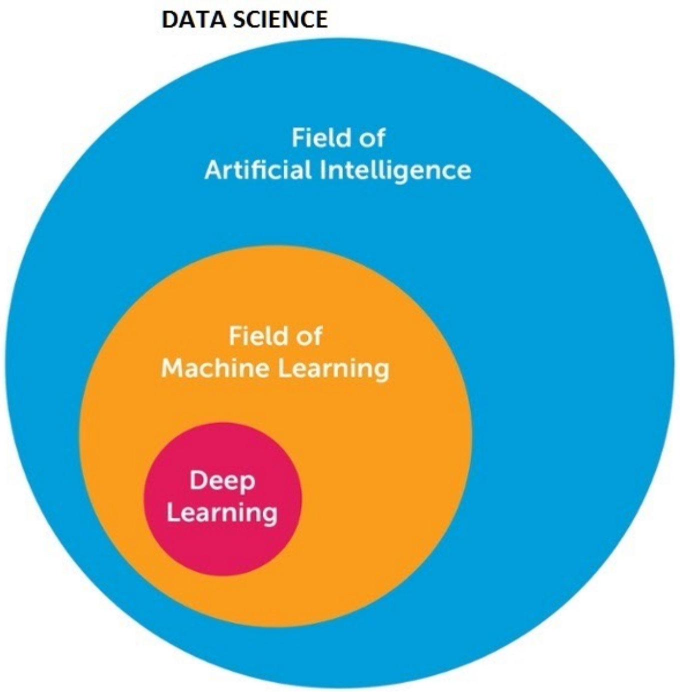
- Camporeale, E., & The Scientific Organizing Committee of ML-Helio (2020). ML-Helio: An emerging community at the intersection between heliophysics and machine learning. Journal of Geoph e2019JA027502. https://doi.org/10.1029/2019JA027502
- Machine Learning in Heliophyiscs, ML- Helio conference, 30 August - 3 September 2022, Boulder (CO), USA

Machine learning conferences in Heliophysics

- Frontiers in Astronomy and Space Science: Machine Learning in Helioph

Journal for ML in Heliophysics

Data Science, Artificial Intelligence, Machine Learning

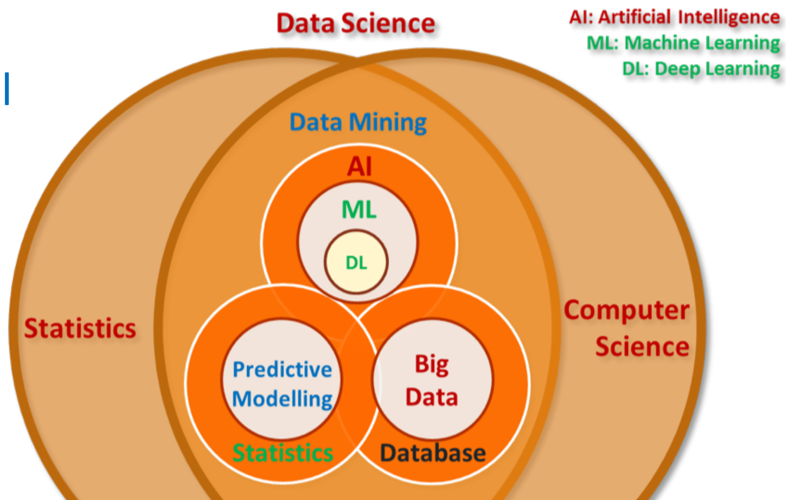


https://www.mygreatlearning.com/blog/difference-data-science-machine-learning-ai/

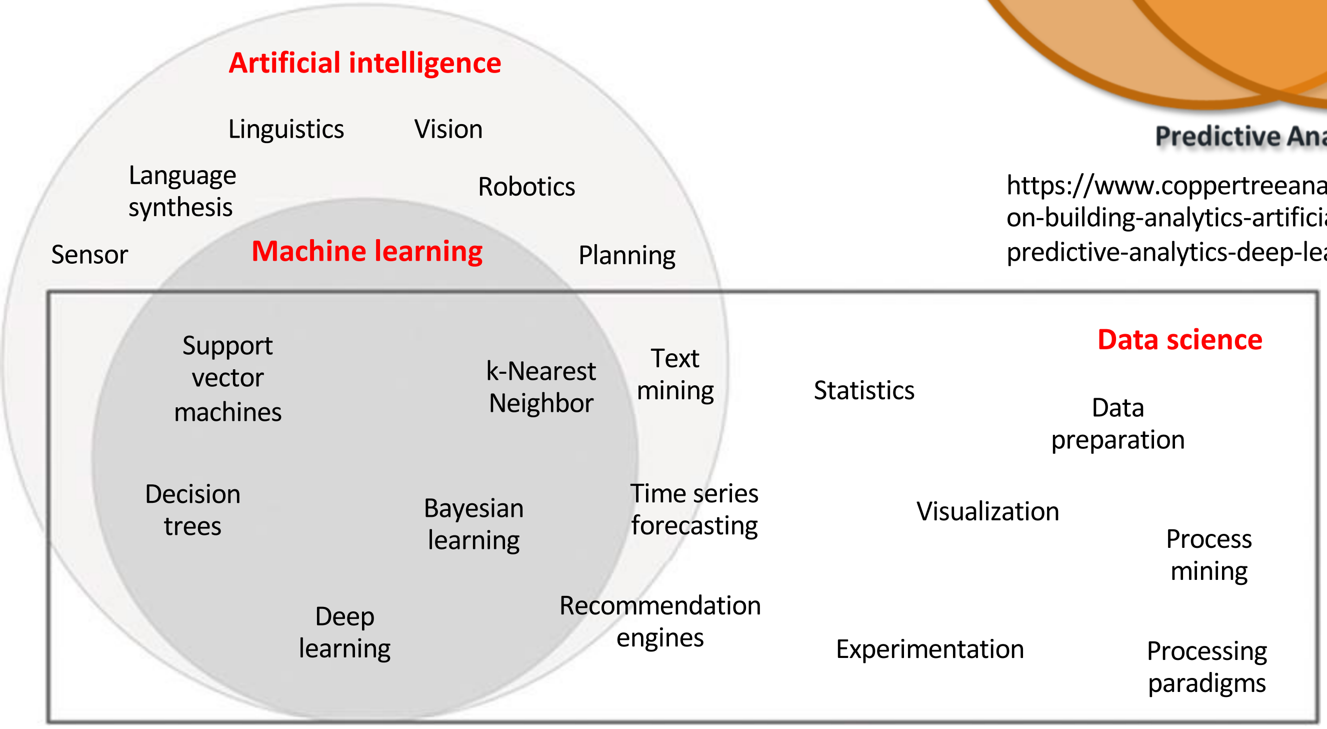
Data science – Using statistical and computer science techniques to extract nonobvious & useful patterns and relationships from data sets

Artificial intelligence – Giving machines the capability of mimicking human behavior, particularly cognitive functions (decision making)

Machine learning – Enabling machines to learn from experience



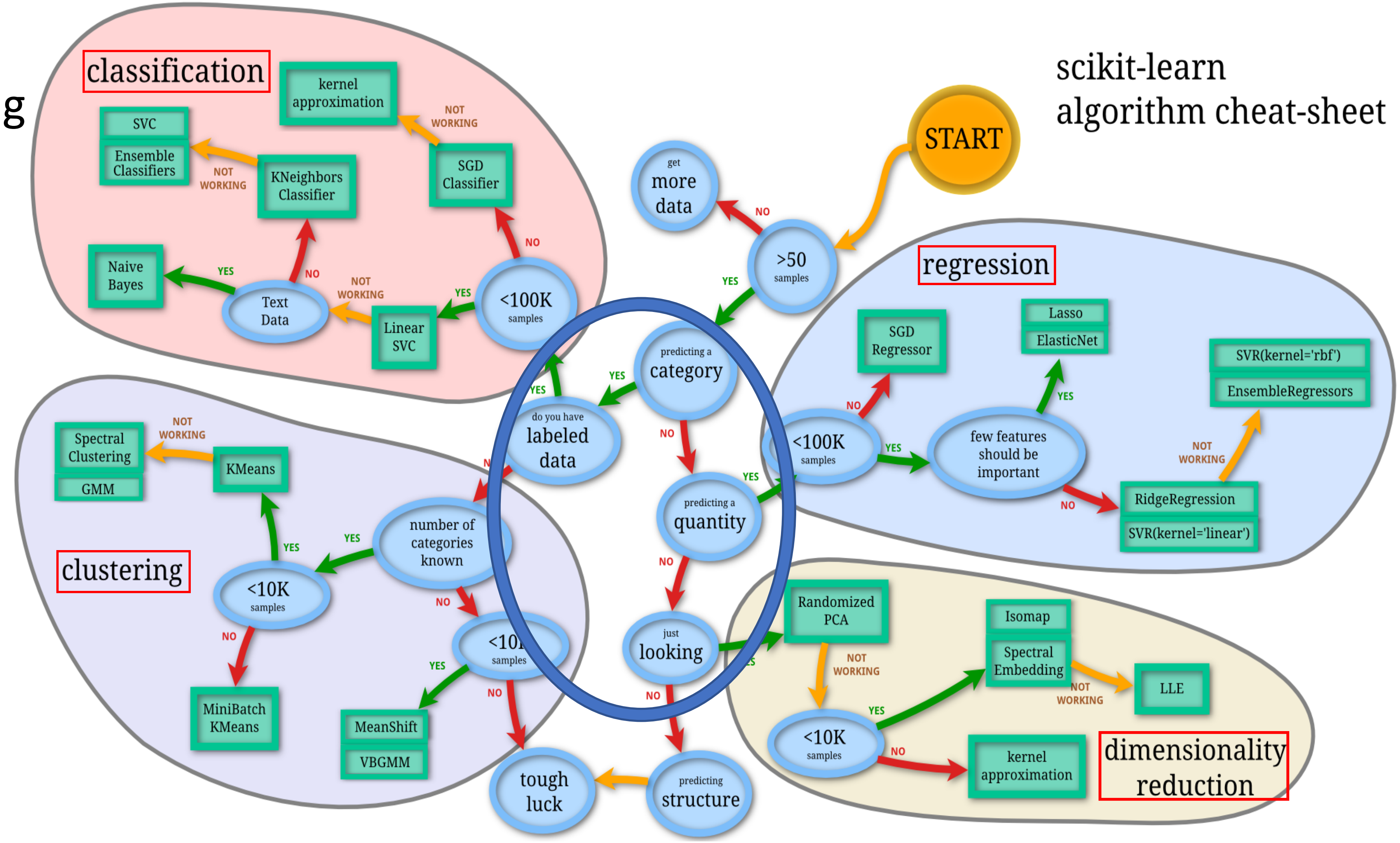
https://www.coppertreeanalytics.com/fundamental-series-on-building-analytics-artificial-intelligence-machine-learning-predictive-analytics-deep-learning-whats-the-difference/



[Adapted from Kotu & Deshpande, 2019]

Data Requirements Vary With ML Models

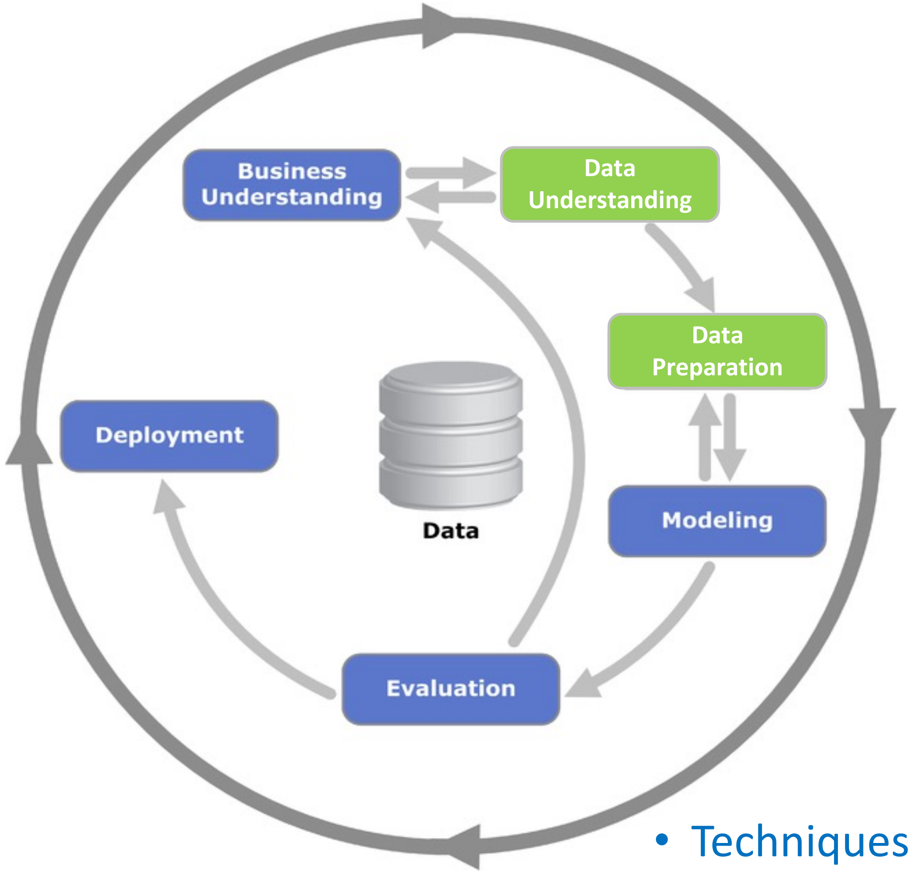
- A problem with machine learning is often in finding the right estimator/tool for the job.
- **But different estimators are better suited for different problems involving different types of data.**
- The flowchart gives a rough guide on how to match up problems with estimators based on available data.

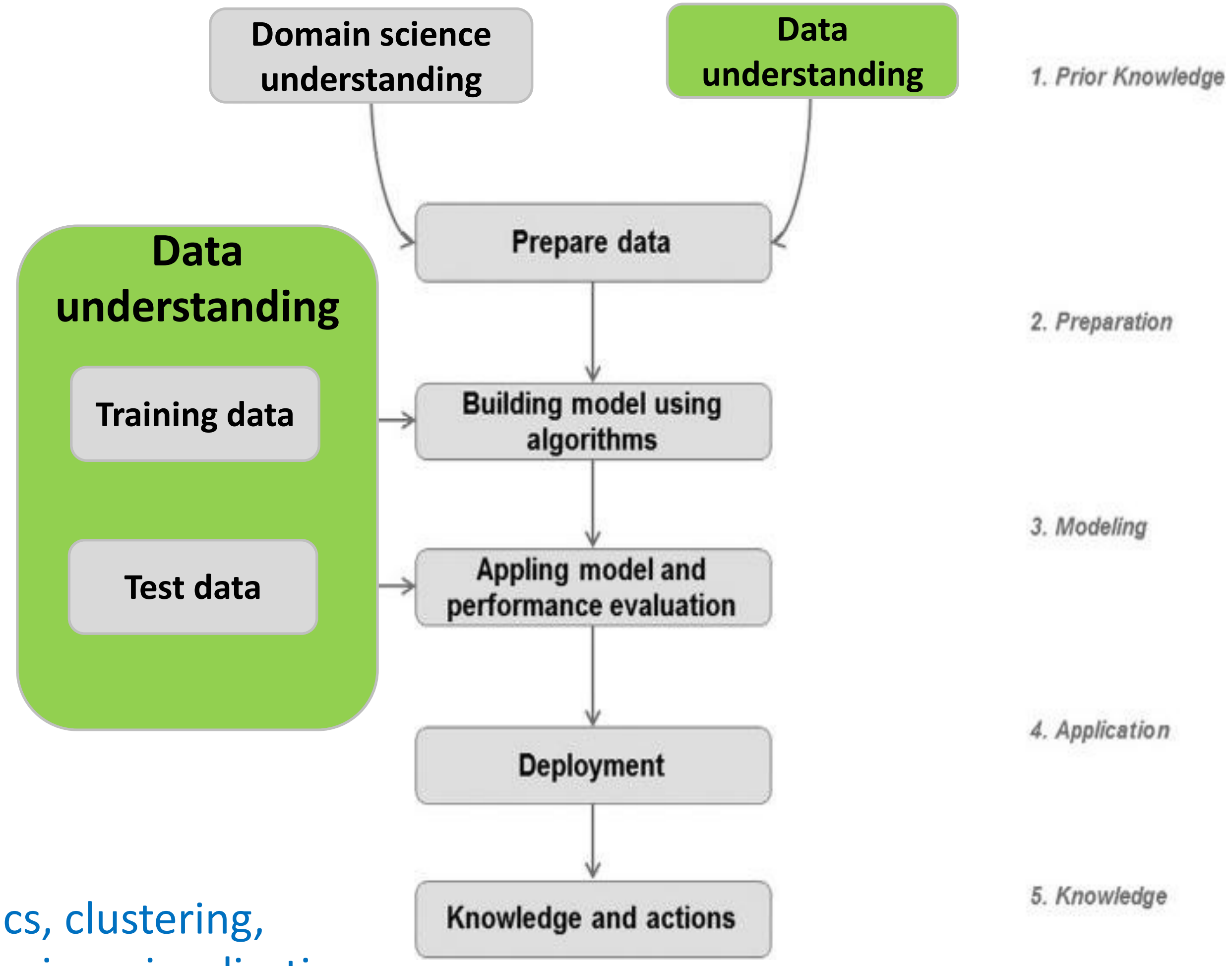


https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Data Science Process

Cross-Industry Standard Process for **Data Mining** (CRISP-DM) Model in Business World [Shearer, 2000]





[Adapted from Kotu and Deshpande, 2019]

- Techniques include statistics, clustering, classification, machine learning, visualization, logic, computer science,…etc.
- **Data understanding** and **data preparation** are two crucial steps in the data science process

Preparing/Wrangling Data for Data Science Analysis (1/2)

Preparing datasets to suit a particular data science (analysis) task is perhaps the most time-consuming part of the data science process.

- Data exploration (aka exploratory data analysis)
  - Obtain descriptive statistics to learn about data structure in terms of patterns and correlations, parameter value characteristics (means, standard deviations, extreme values) and distributions.
  - Visualize data to reveal possible inter-relationships within the dataset (preparing for supervised learning or clustering analysis).
- Data quality (Data cleaning I)
  - Errors in data will impact the representativeness of the model.
  - Techniques used to process and transform data and provide data alerts should be documented.
  - Documentation should include data cleansing practices for elimination of duplicate records, quarantining outlier records, standardization of attribute values, substitution of missing values, etc.
- Data gaps (Data cleaning II)
  - Missing data means that those missing data cannot take part in or cannot have an influence on the data science process. It is therefore important to assess the impact of data gaps in a particular data science task.
  - The first step of managing missing values is to understand the source of the missing data, which will often guide the mitigation methodology to use.
    - Substituting with artificial data, e.g., derived from the dataset, so that the issue can be managed with marginal impact on the later steps in the data science process. This method is useful if the missing values occur randomly and the frequency of occurrence is quite rare.
    - Ignoring all the data records with missing data values or records with poor data quality. This method can be applied to data records that are not inherently (e.g., temporally or spatially) related. Some data science algorithms are good at handling records with missing values, while others expect the data preparation step to handle it before the model is inferred. For example, k-nearest neighbor (k-NN) algorithm for classification tasks are often robust with missing values. Neural network models for classification tasks do not perform well with missing attributes, and thus, the data preparation step is essential for developing neural network models.
- Outliers (Validating data records)
  - Outliers may occur because of correct data capture or erroneous data capture. Regardless, the presence of outliers needs to be understood and will require special treatments. The purpose of using data science technique is to generalize a pattern or a relationship within a dataset in order to create a representative model. The presence of outliers may skew the representativeness of the inferred model. Detecting outliers may also be the primary purpose of some data science applications.

Preparing/Wrangling Data for Data Science Analysis (2/2)

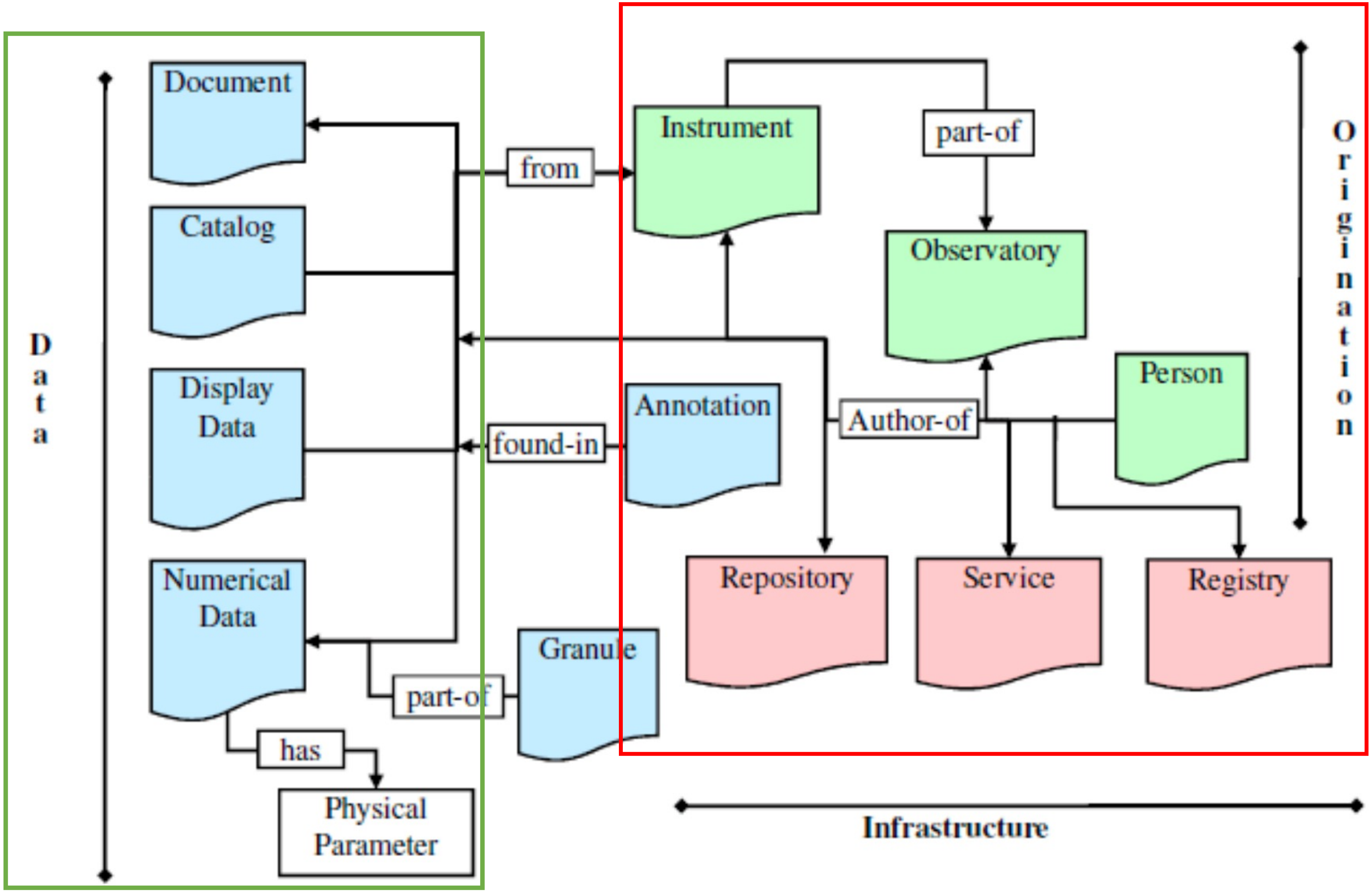
- Data transformation
  - Data science algorithms often require data to be in a form different from how the data are provided naturally, so some conversion will be necessary.
  - Some algorithms, like k-NN, expect input data attributes to be numeric and normalized, because the algorithm compares the values of different attributes and calculates the normalized distance between the data points.
  - Normalization prevents one attribute dominating the distance results because of large values. The distance calculation will always be dominated by slight variations in the attributes. One solution is to use a more uniform scale from 0 to 1 by normalization. This way, a consistent comparison can be made between the two different attributes with different units.
- Data types and conversion
  - The attributes in a dataset can be of different types, such as continuous numeric (e.g., the AE or Dst index) or categorical (like the Kp index).
  - Different data science algorithms can require different data attribute types. In case of linear regression models, the input attributes have to be numeric. If the available data are categorical, they must be converted to continuous numeric attribute. Similarly, numeric values can be converted to categorical data types by a technique called binning, where a range of values are specified for each category.
- Feature selection
  - Many data science problems involve a dataset with many attributes.
  - Large number of attributes increases the complexity and dimensionality of a model. Data samples can be sparse in high-dimensional space, resulting in low reliability of (some parts of) the models
    - If all the attributes are not equally useful in a model, their presence might be counterproductive.
    - Some attribute may be correlated with each other (like Kp and AE).
  - While more detailed information is desired in data science, especially in clustering and classification, reducing the number of attributes to just the “most important features” without significant loss in the performance of the model, is called feature selection, which may lead to a more simplified model and helps synthesize a more effective explanation of the model.
- Data sampling
  - Selecting a representative subset of the original dataset for analysis or modeling is called data sampling. It is typically used to create training and test datasets.
  - Data sampling is usually done by using simple sampling, or class label specific sampling, that does not have enough samples of outlier classes.
  - Stratified sampling is a sampling process in which the total population is first subdivided into smaller groups (called strata) based on different characteristics in the population data. Random samples are then taken from each of stratus such that the sample numbers from the strata maintain the same proportions of representation as in the original population. Compared to sampling the original population randomly, stratified sampling ensures that the samples used to create a model properly represent the classes in the original population having normal and outlier records. In classification applications, sampling is used to create multiple base models, each developed using a different set of sampled training datasets. These base models are used to build one meta model, called the ensemble model, where the error rate is improved when compared to that of the base models.

Kuto, V. and B. Deshpande, Data Science Concepts and Practice, Morgan Kaufman Publishers, an imprint of Elsevier, 2019.
https://medium.com/@ODSC/top-data-wrangling-skills-required-for-data-scientists-8a6b7dc604a7

Data Preparation/Wrangling for Data-Science & AI Studies

Data preparation for data science	**AI Training Data Set Requirements** (see https://github.com/rmcgranaghan/data_science_tools_and_resources/wiki/Curated-Reference%7CChallenge-Data-Sets)	**Metadata description requirements**
Data exploration	• Metadata are appended to be consistent with common formats, ensure complementarity between the features, and any other supplemental information is applied to make the data usable. • Data release standards are satisfied, with "all clear" from the preparers of the dataset for general distribution and use.	• Uniform metadata model with parameter descriptions • Data and version release date, instrument PI and/or POC • AccessURL & InformationURL
Data quality	• Data are [calibrated, promoted in level, standardized, etc.] so that values correspond well to the physical system being studied	• Calibration info
Data gap	• Spurious data and non-physical values are either corrected or identified	• Caveat descriptions
Outliers	• A description of the types of physical processes that are well-sampled and therefore can be addressed or approached with each given data set	• ???
Data type and conversion	• Data are interpolated, patched, etc. to provide even or consistent sampling. May be re-sampled in space or time in order to synchronize with other data that could be used in the feature set.	• Interpolation procedure documentation • Sampling resolution or cadence specification
Data transformation	• The data are then made available in a way that is easy to read into an AI algorithm (e.g. Keras, PyTorch); • Feature scaling optimized for learning (such as normalizing from 0 to 1 for scalars, -1 to 1 for vectors, logarithmic for multi-scale sensitivity) [Outside purview of metadata???]	• Implement standard access protocol server
Feature selection	• Data labels (such as features and event indicators), if relevant, are compatible with the feature set, i.e. Labels "Y" should have a format compatible for learning with feature set "X" as an input. [used for creating training output data for ML]	
Data sampling	• Usable code with commonly performed steps (query/sampling, data ingest, etc.). The target is to provide the steps that almost every user would have to develop on their own.	• InformationURL

SPASE Ontology



Data Characteristics Needed for Data Science Analysis

The key data characteristics are the (a) availability and (b) usability of (c) large amount of (d) similarly processed data that can be (e) readily ingested into data science analysis (e.g., ML) algorithms or tools:

a) Availability
   i. What: Measurement types
   ii. Where: Observatory, instrument, repositories
   iii. How: Access information (ways to access data, data formats)

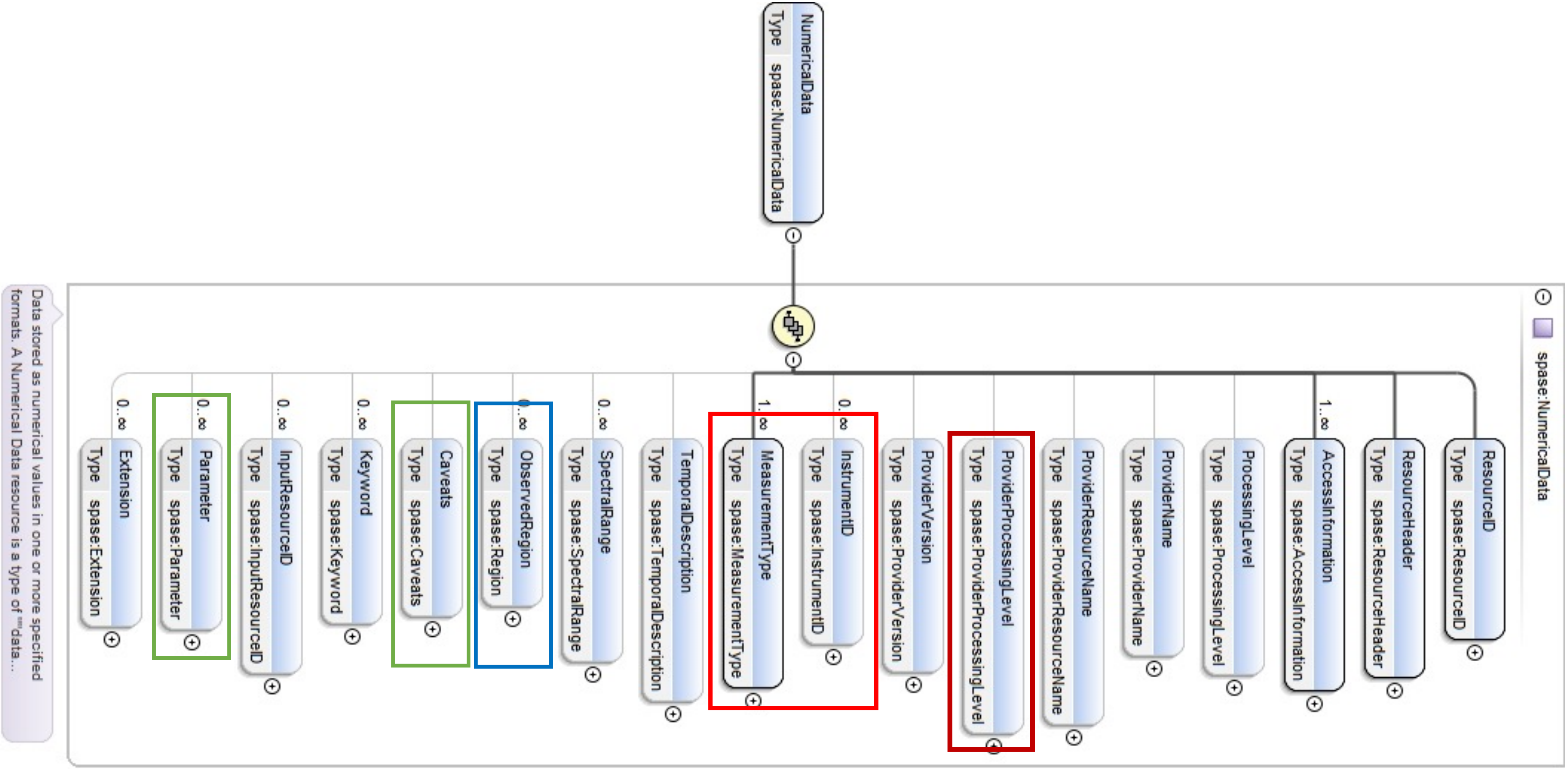
b) Usability
   i. Parameter descriptions [parameters (attributes & labels), coordinate system, units, cadence, min & max values (for scaling)]
   ii. Data gaps
   iii. Caveats

c) Data quantity – sample size
   i. Time coverage
   ii. Location coverage

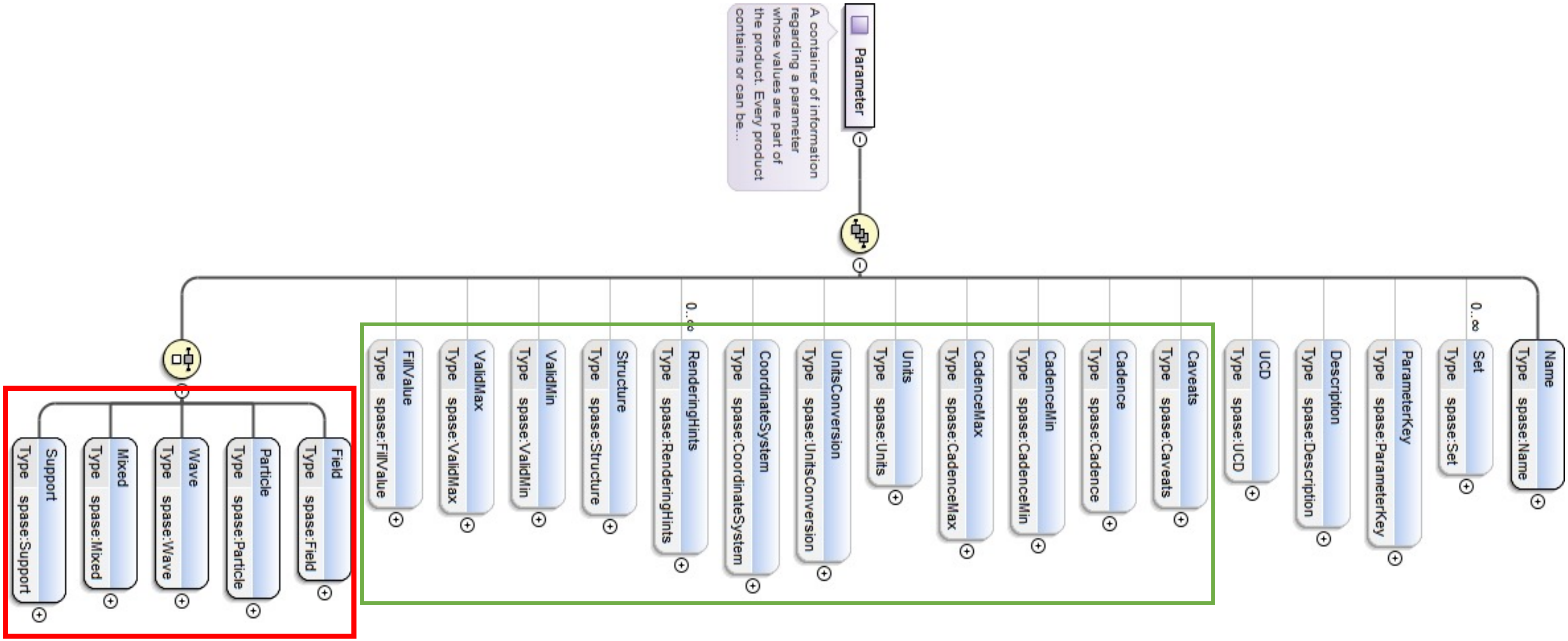
d) Processing – Data of the same type obtained from different sources need to be processed in the same way to ensure statistical consistency
   i. Processing levels
   ii. Calibration & cross-calibrations

e) Ingestion readiness
   i. Data formats
   ii. Parameter list
   iii. Temporal and spatial synchronization of datasets (compatibility in cadence and coverage)

Numerical Data



Parameter Description



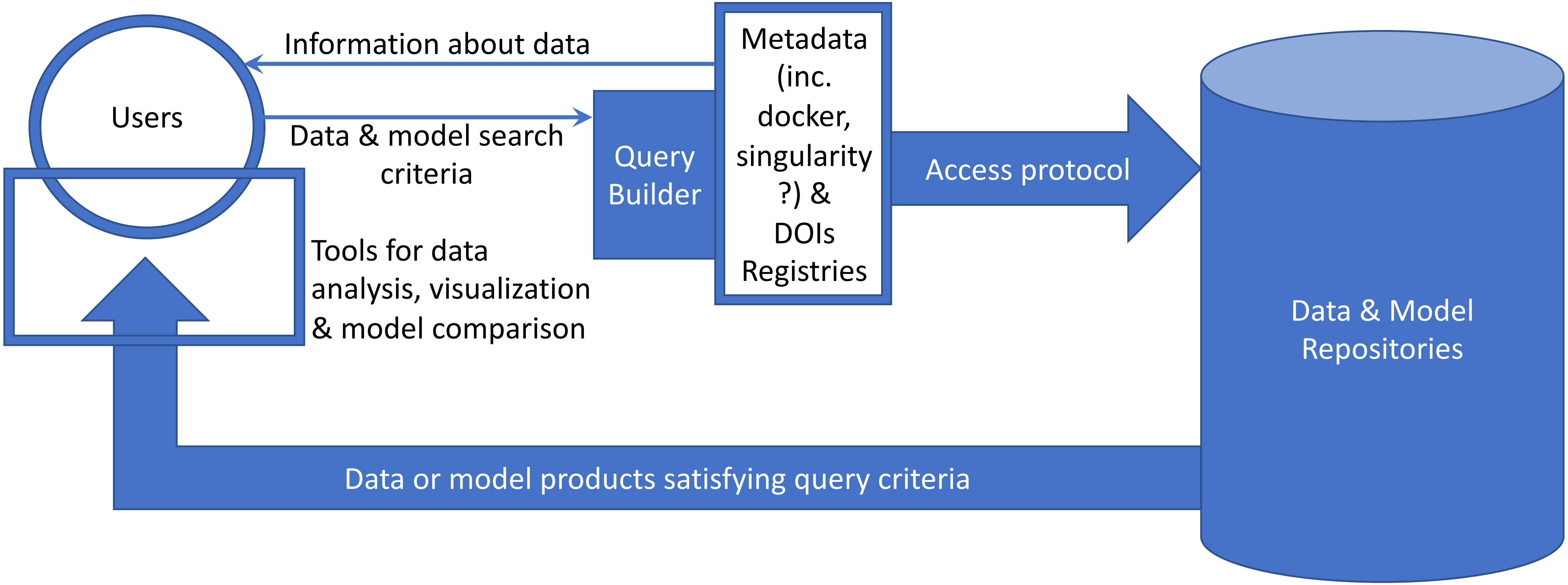
A container of information regarding a parameter whose values are part of the product. Every product contains or can be...

Measurement types

Infrastructure needs

- Archiving AI/ML-ready & training datasets @ SPDF
  - Archiving cleaning documentation and code and environment
  - Docker registry?
- Archiving and storing of AI/ML models @ CCMC
- Can SPASE be used effectively to
  - Facilitate the compilation of training datasets?
  - Identify and help search for a DS/ML datasets?
- Consistency of variable names.

Information Architecture


Users
Information about data
Data & model search criteria
Query Builder
Metadata (inc. docker, singularity ?) & DOIs Registries
Access protocol
Data & Model Repositories
Tools for data analysis, visualization & model comparison
Data or model products satisfying query criteria

Metadata Needs for Data Science

- Event identification (in model data vs. observations)
- Ontologies
- What do data science analysis algorithms need?
- For ML, metadata should include measurement error info (source)
- Operational history of instruments