Machine Learning Workflow

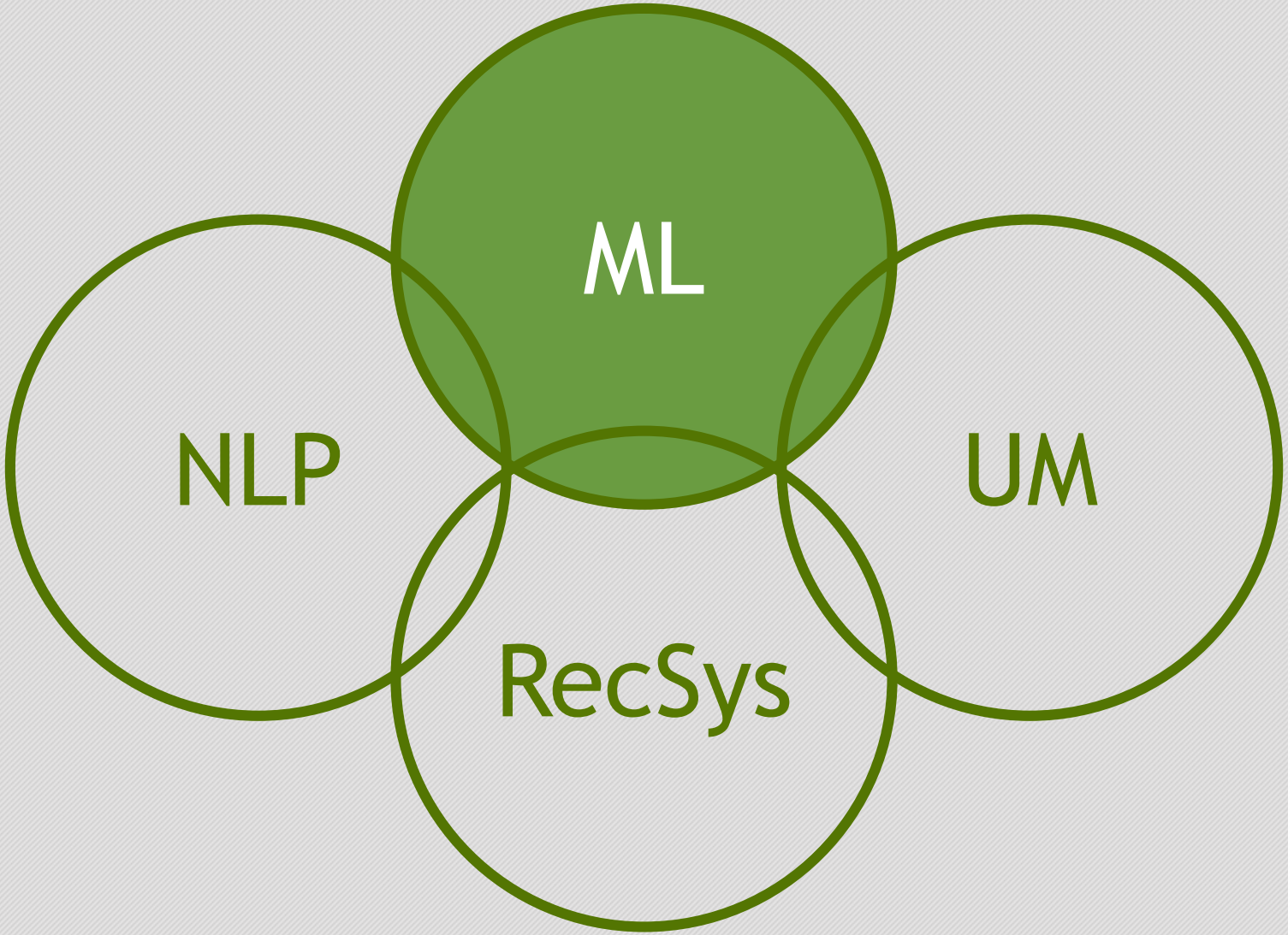
datalys

Ivan Srba

5th October 2017

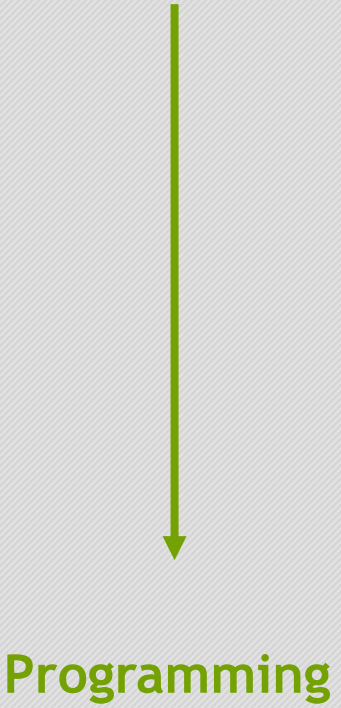
PeWe@FIIT
personalized web group

Topics Overview



Setting Context and Terminology

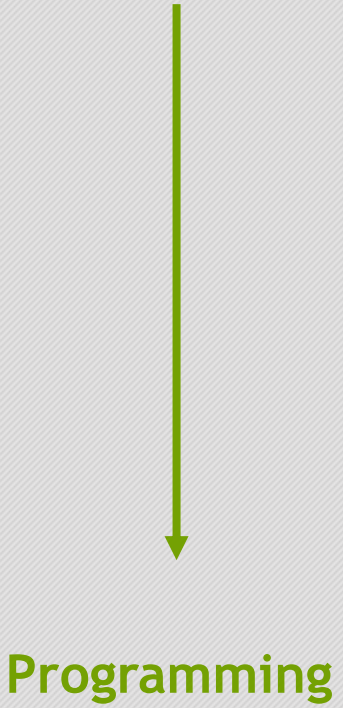
- **Statistics**
- **Data Mining**
- **Machine Learning**
- **Artificial Intelligence**



https://stats.stackexchange.com/questions/5026/what-is-the-difference-between-data-mining-statistics-machine-learning-and-ai

Setting Context and Terminology

- **Statistics** is just about the numbers, and quantifying the data.
- **Data Mining**
- **Machine Learning**
- **Artificial Intelligence**



https://stats.stackexchange.com/questions/5026/what-is-the-difference-between-data-mining-statistics-machine-learning-and-ai

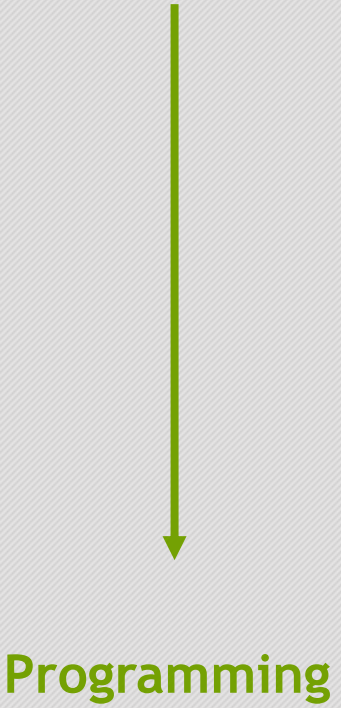
Setting Context and Terminology

- **Statistics** is just about the numbers, and quantifying the data.
- **Data Mining** is about using **Statistics** as well as other programming methods to find patterns hidden in the data so that you can *explain* some phenomenon.
- **Machine Learning**
- **Artificial Intelligence**

https://stats.stackexchange.com/questions/5026/what-is-the-difference-between-data-mining-statistics-machine-learning-and-ai

Setting Context and Terminology

- **Statistics** is just about the numbers, and quantifying the data.
- **Data Mining** is about using **Statistics** as well as other programming methods to find patterns hidden in the data so that you can *explain* some phenomenon.
- **Machine Learning** uses **Data Mining** techniques and other learning algorithms to build models of what is happening behind some data so that it can *predict* future outcomes.
- **Artificial Intelligence**



https://stats.stackexchange.com/questions/5026/what-is-the-difference-between-data-mining-statistics-machine-learning-and-ai

Setting Context and Terminology

- **Statistics** is just about the numbers, and quantifying the data.
- **Data Mining** is about using **Statistics** as well as other programming methods to find patterns hidden in the data so that you can *explain* some phenomenon.
- **Machine Learning** uses **Data Mining** techniques and other learning algorithms to build models of what is happening behind some data so that it can *predict* future outcomes.
- **Artificial Intelligence** uses models built by **Machine Learning** and other ways to *reason* about the world and give rise to intelligent *behavior* whether this is playing a game or driving a robot/car.

Mathematics

Programming

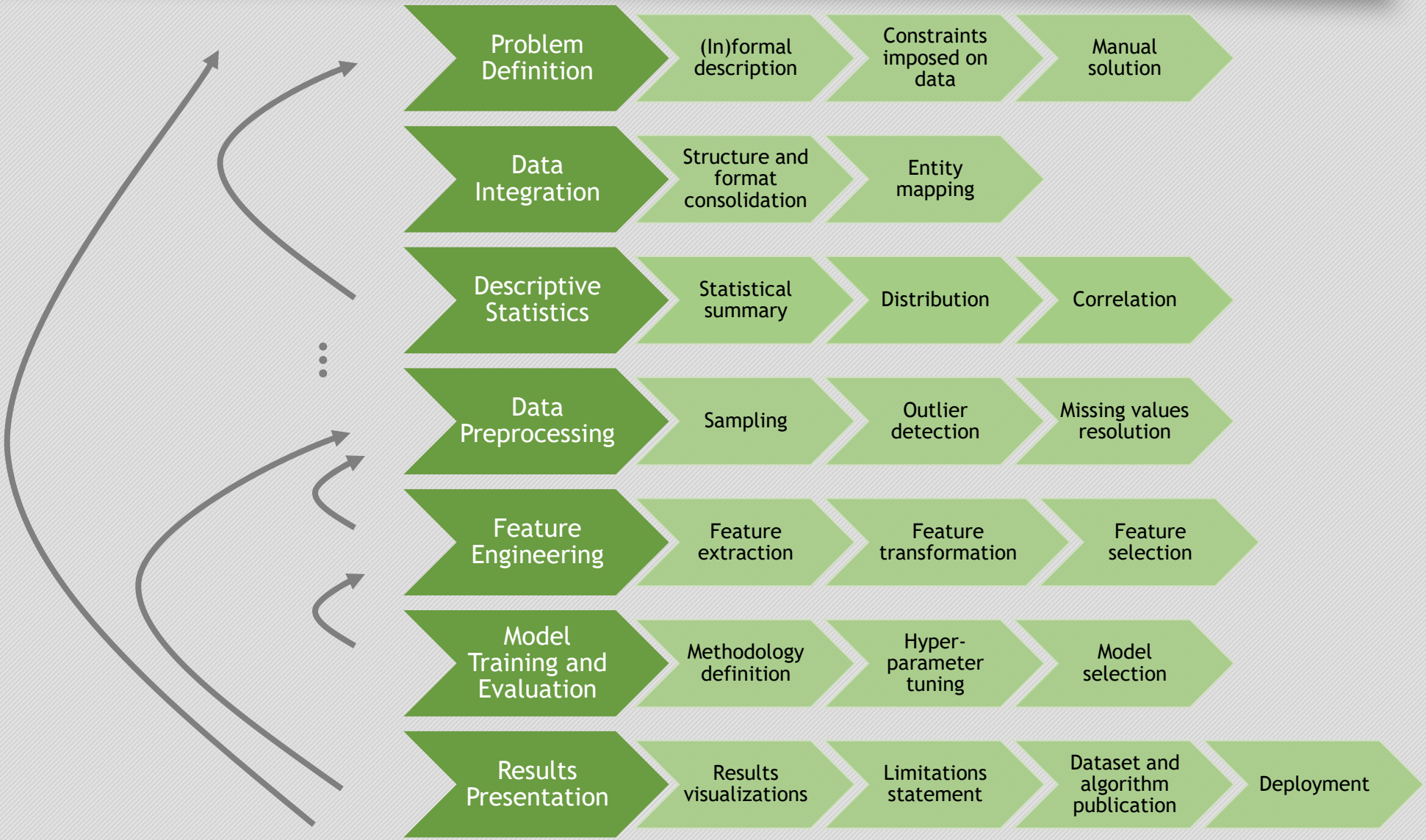
https://stats.stackexchange.com/questions/5026/what-is-the-difference-between-data-mining-statistics-machine-learning-and-ai

Slovak-English Domain Term Dictionary

https://docs.google.com/spreadsheets/d/1rs2EXUhhxflgJmxRR-a1Bjj6y9XVorTHjjCVTSxpaOQ/edit?usp=sharing

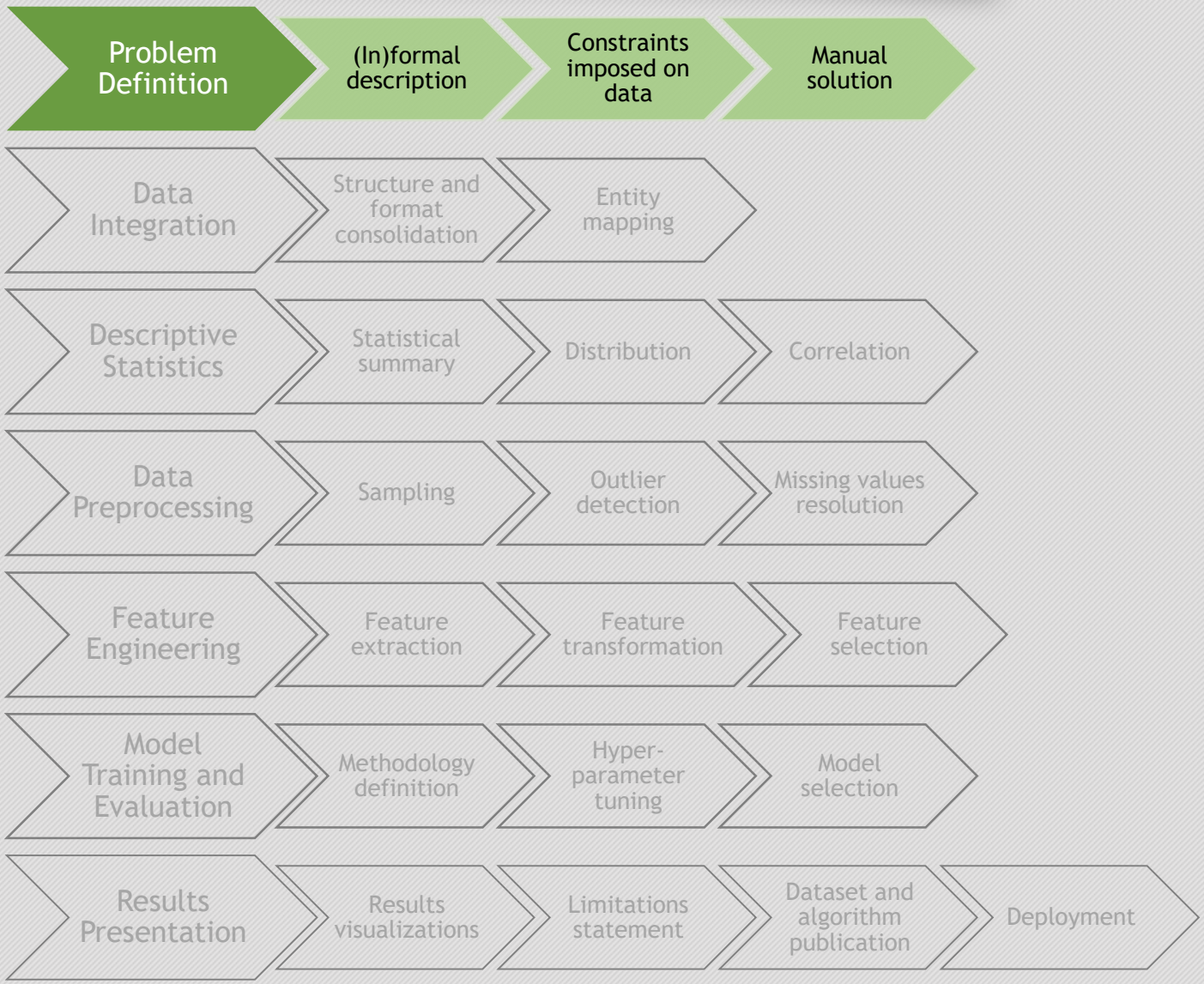
Link also available at pewe.sk > Datalys > Resources

ML Workflow: Overview



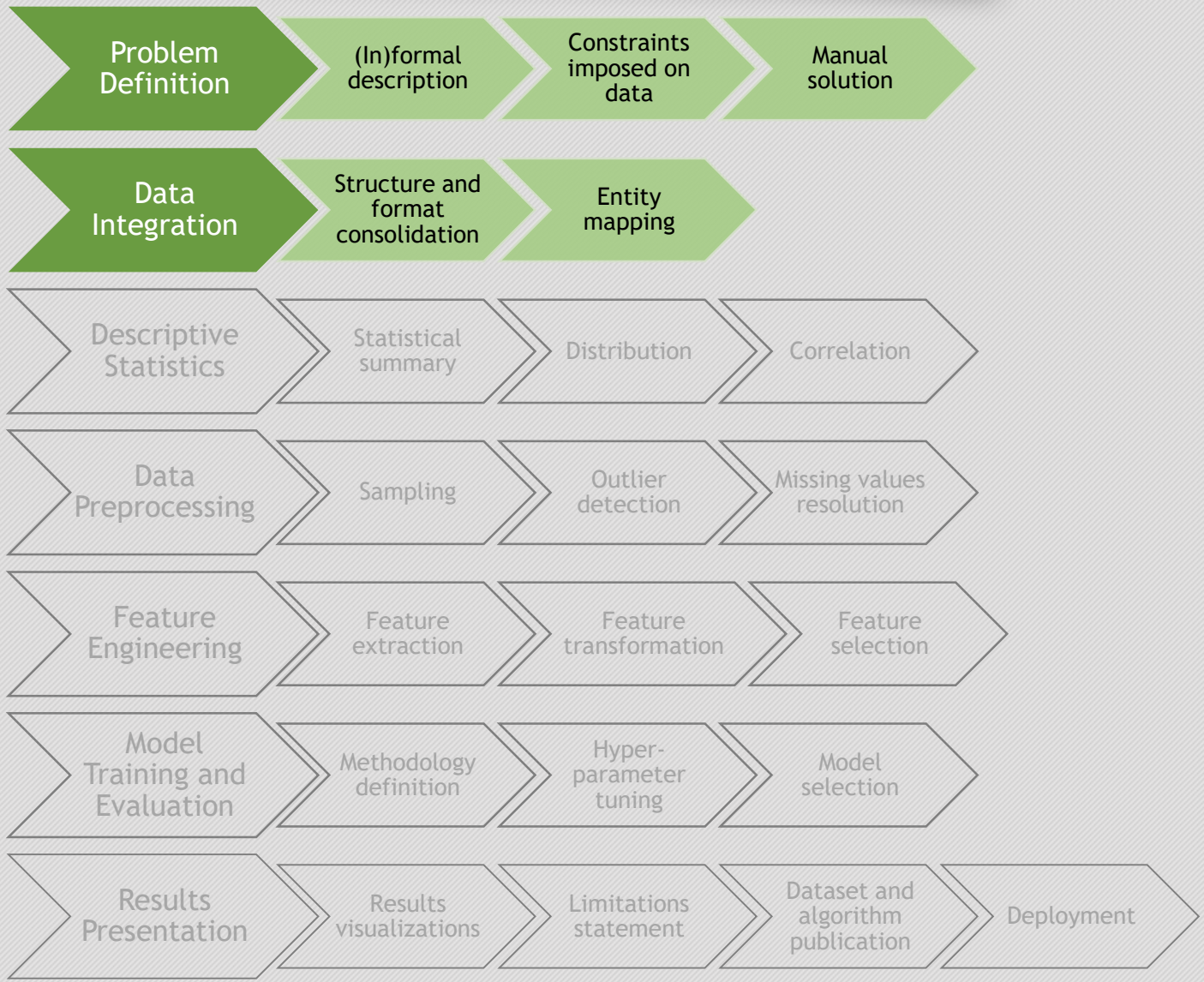
ML Workflow: Problem Definition

- Clearly identify your problem you are trying to solve
  - Informal description
    - As you would explain it to your friends
    - State your motivation to solve the problem
  - Formal description
    - As machine learning task
    - Hypothesis in your thesis
- Identify constraints imposed on required dataset
- Explore possible manual solutions
  - If they do not exist, it is not a problem any more (in many cases)



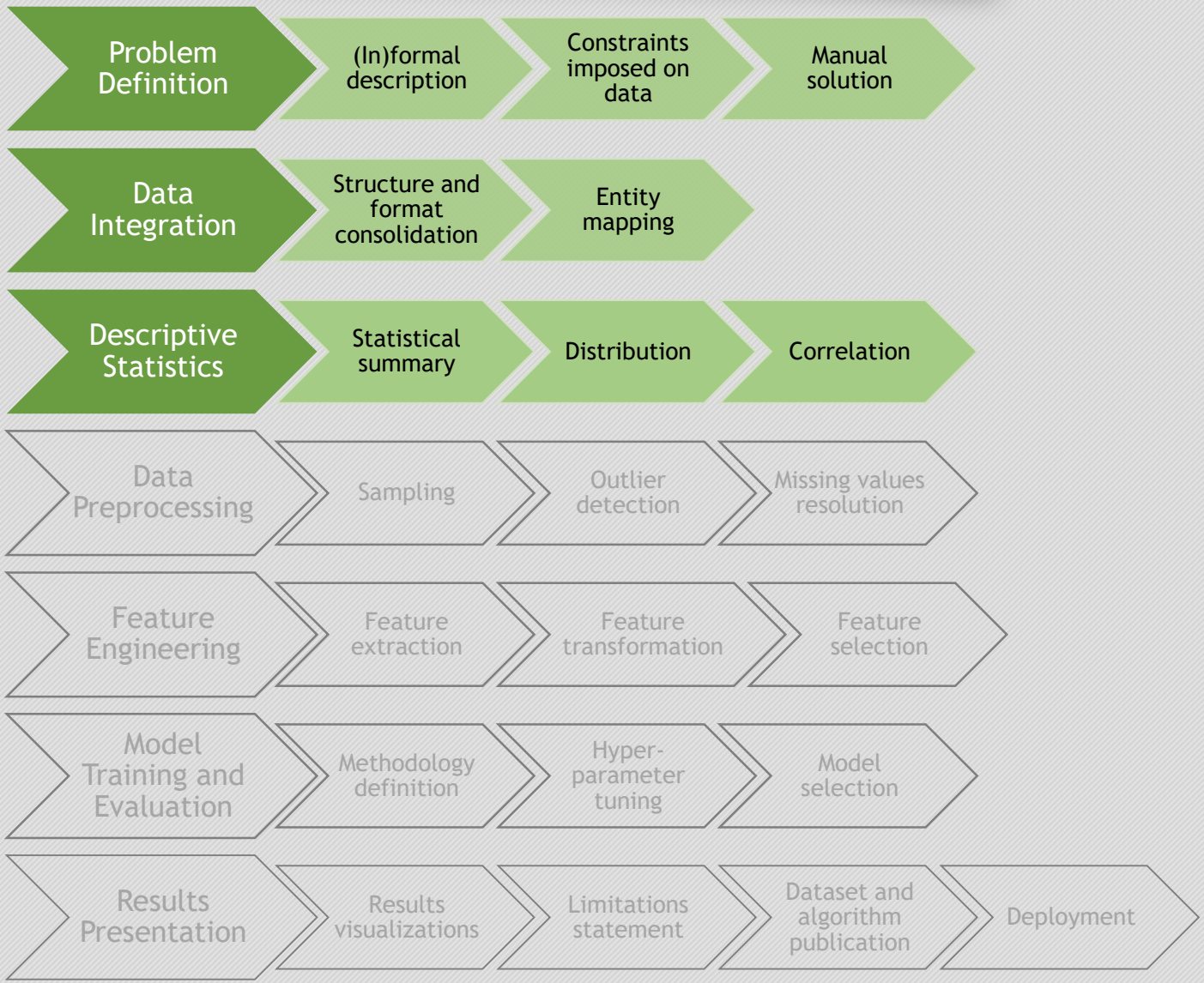
ML Workflow: Data Integration

- Some ML task can be solved only when you integrate data from several sources
- Different sources = different structure and format
  - Data consolidation is required
- Entity mapping
  - User IDs (email, DB ID, cookie, username, …)
  - Item IDs (code, name, …)



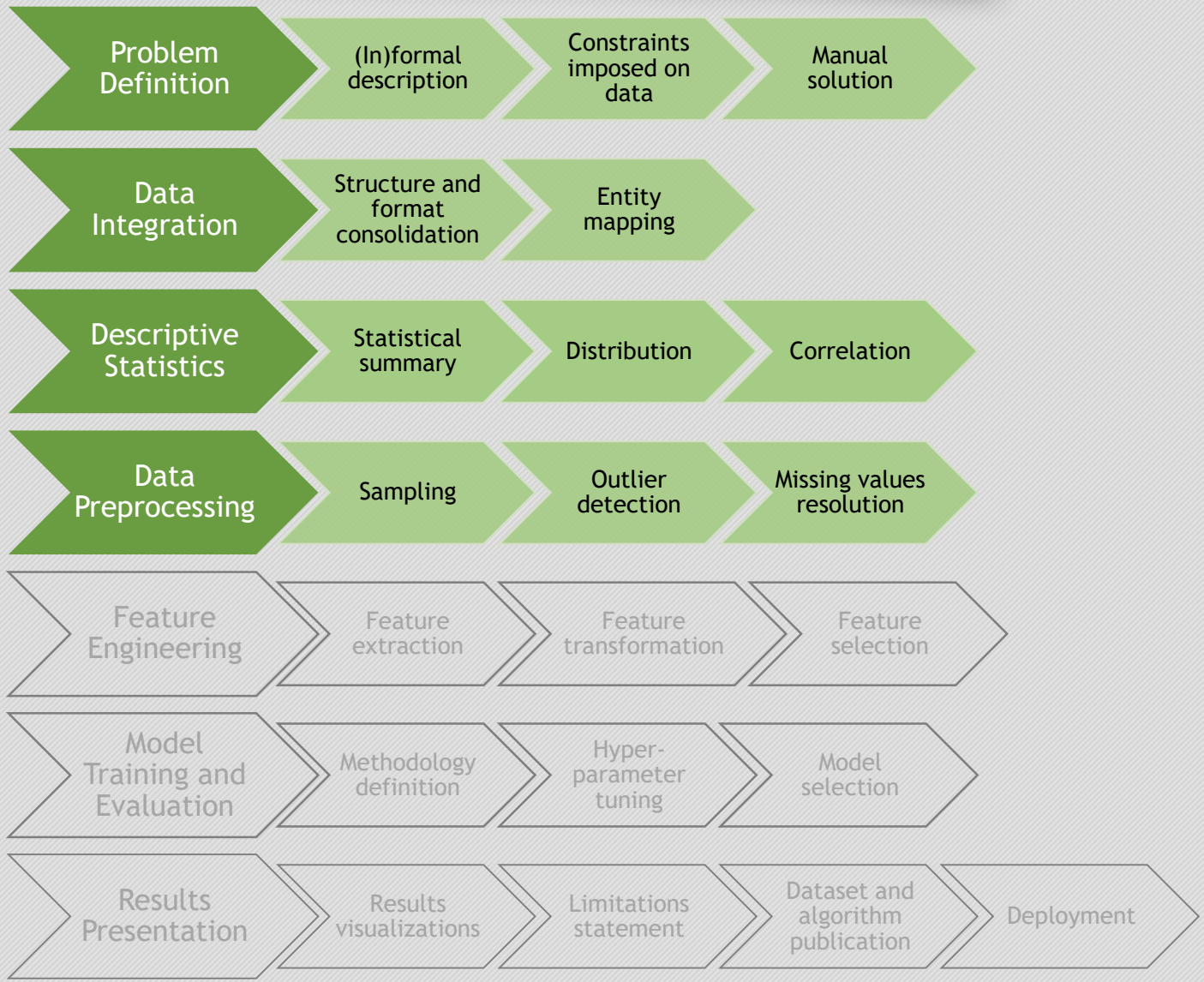
ML Workflow: Descriptive Statistics

- Known your data!
  - Otherwise, you are just guessing…
- Summarize data
  - Volume of data (attributes, instances)
  - Data types
  - Distribution of data
  - Relations in data
- Visualize data
  - Histograms, boxplots, scatterplots
- Result of descriptive statistics is an **important** input to all consequent steps



ML Workflow: Data Preprocessing

- Have large data? Do sampling!
  - Less data result in shorter training times
  - You can still finally run the model on larger portion of data
- Machine learning requires well-prepared data
  - Detect outliers
  - Replace missing values



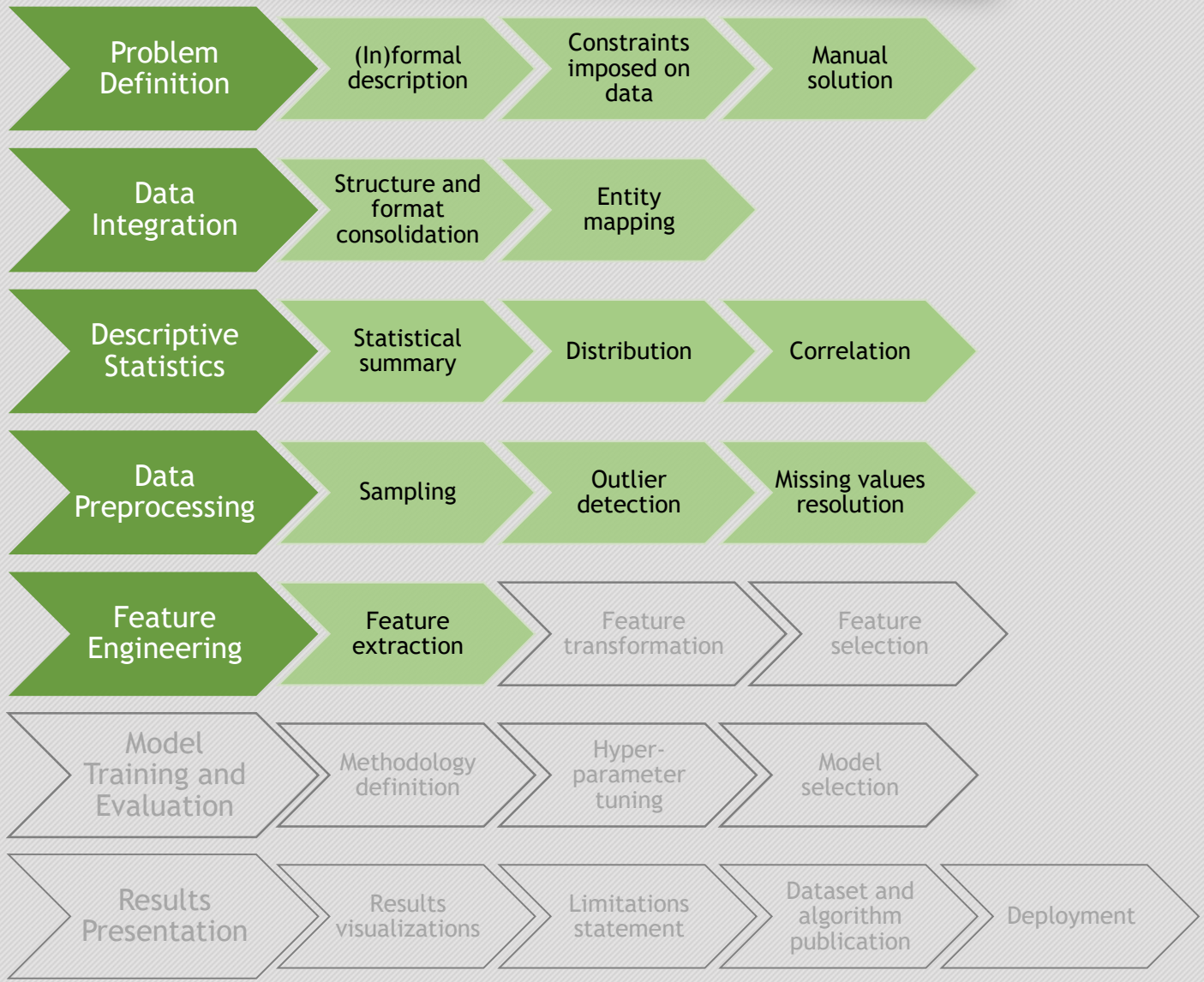
“Coming up with features is difficult, time-consuming, requires expert knowledge. *Applied machine learning* is basically feature engineering.”

Andrew Ng

Feature Engineering

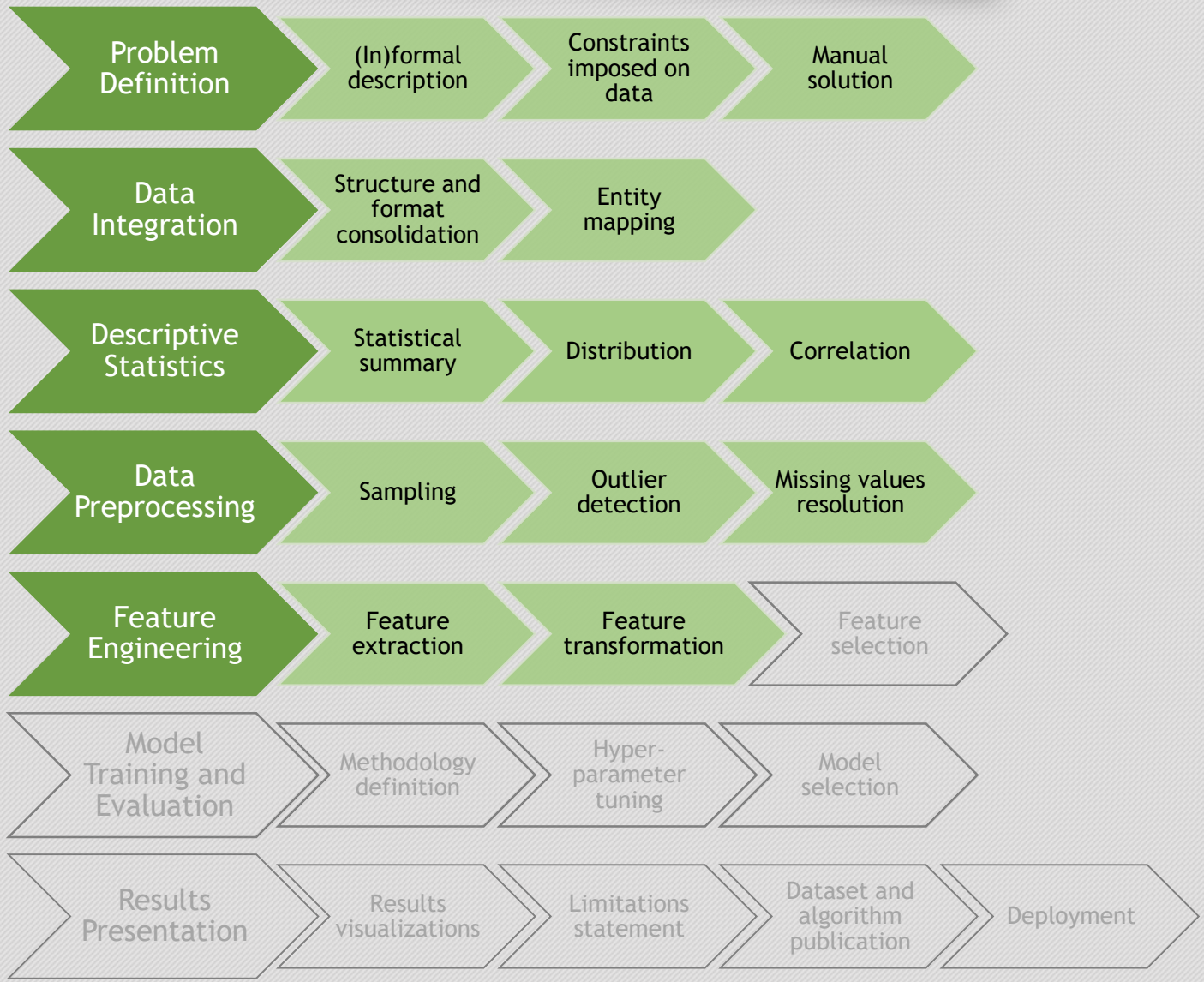
ML Workflow: Feature extraction

- Raw data need to be converted to features
  - Images, text, logs, …
- High-dimensional data need to be reduced
- Techniques
  - Expert-based (UM, NLP)
  - Automatized
  - Dimensionality-reduction (PCA)
  - …



ML Workflow: Feature transformation

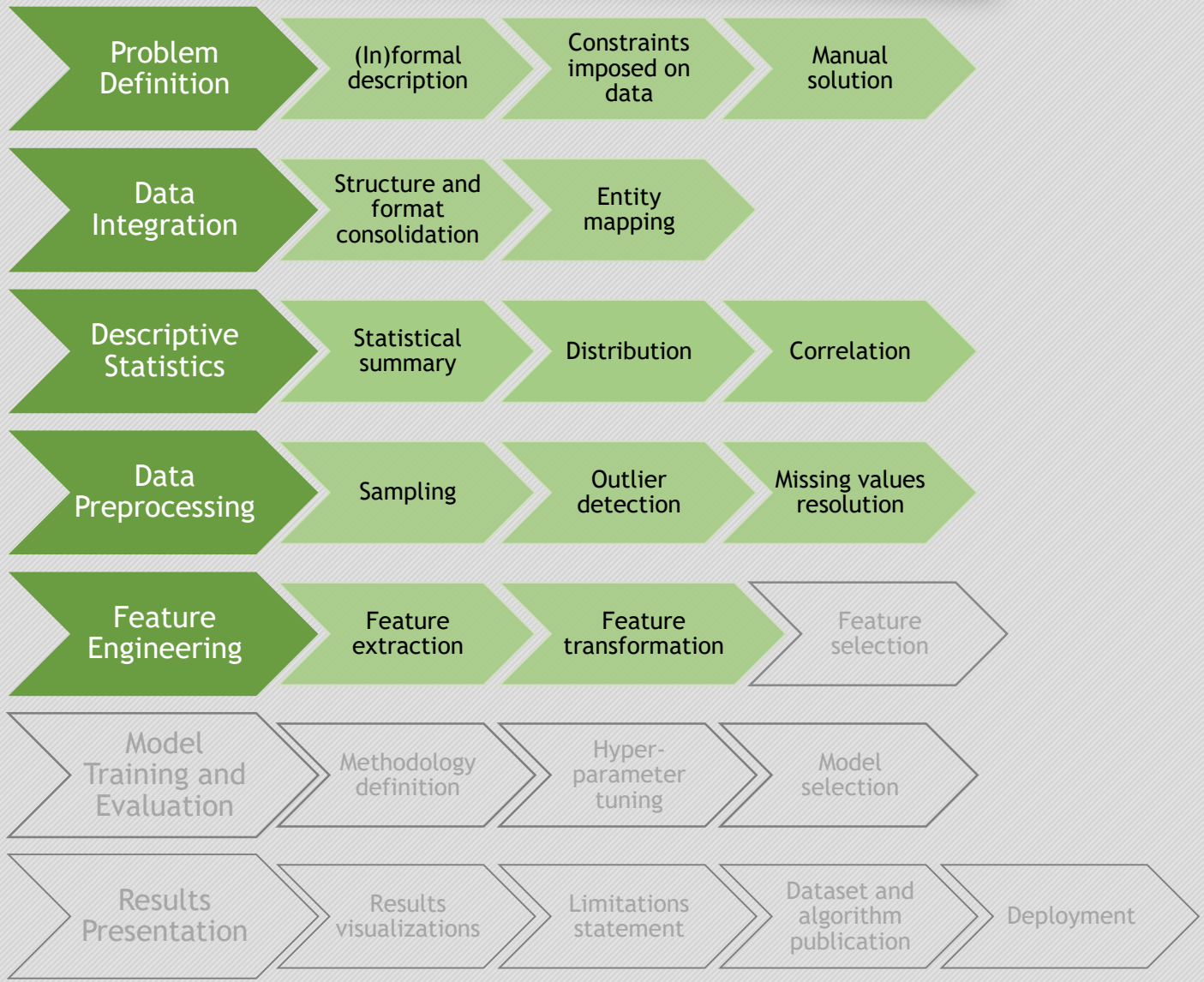
- Some ML algorithms work well when features have specific
  - Distribution
  - Range
  - Data type
- Techniques
  - Scaling
  - Normalization
  - Encoding categorical features
  - Binarization of features



ML Workflow: Feature transformation

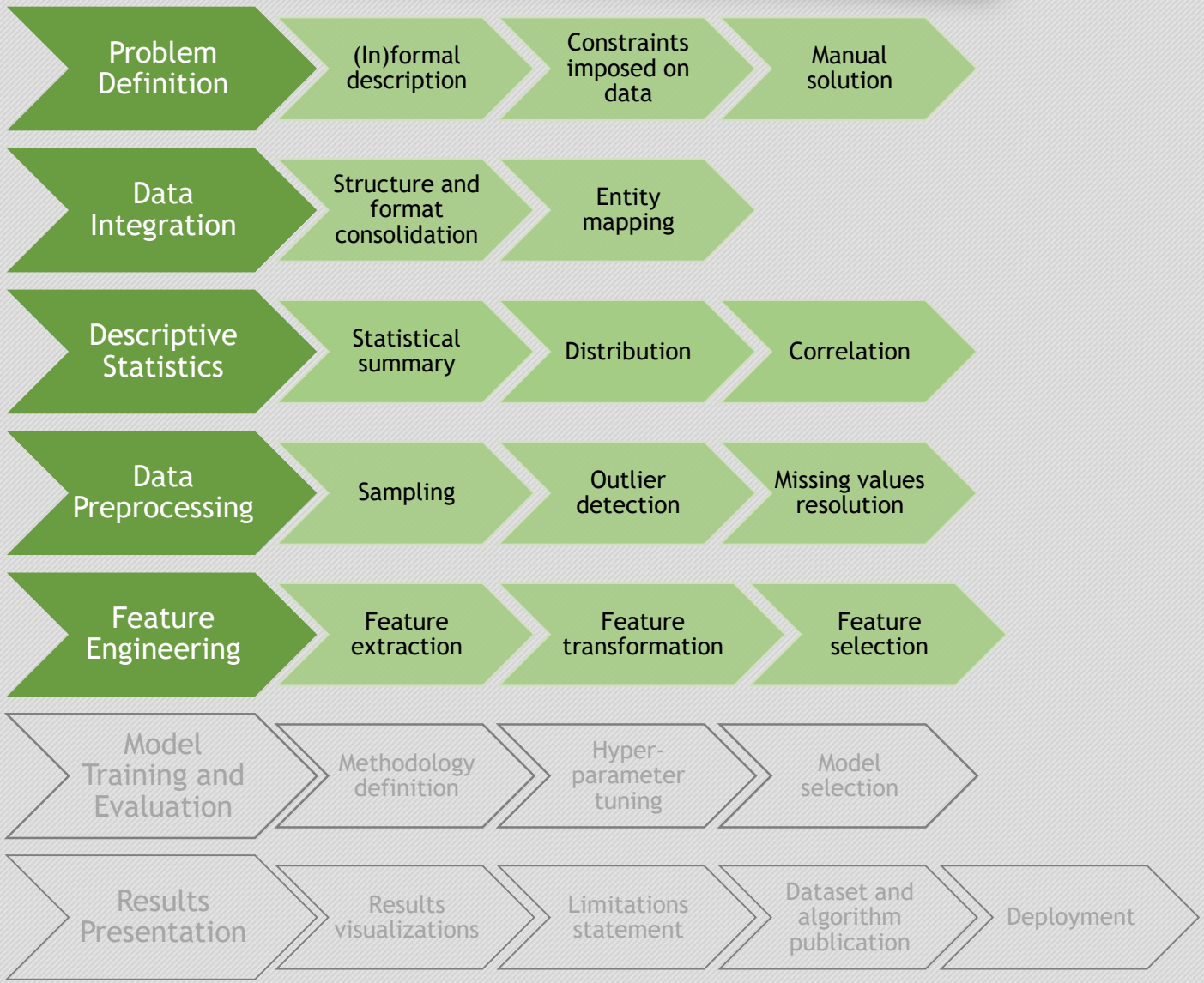
- Some features may not be suitable to be used directly by ML algorithm

- Techniques
  - Combining features
  - Splitting features (e.g. date)
  - Polynomial features



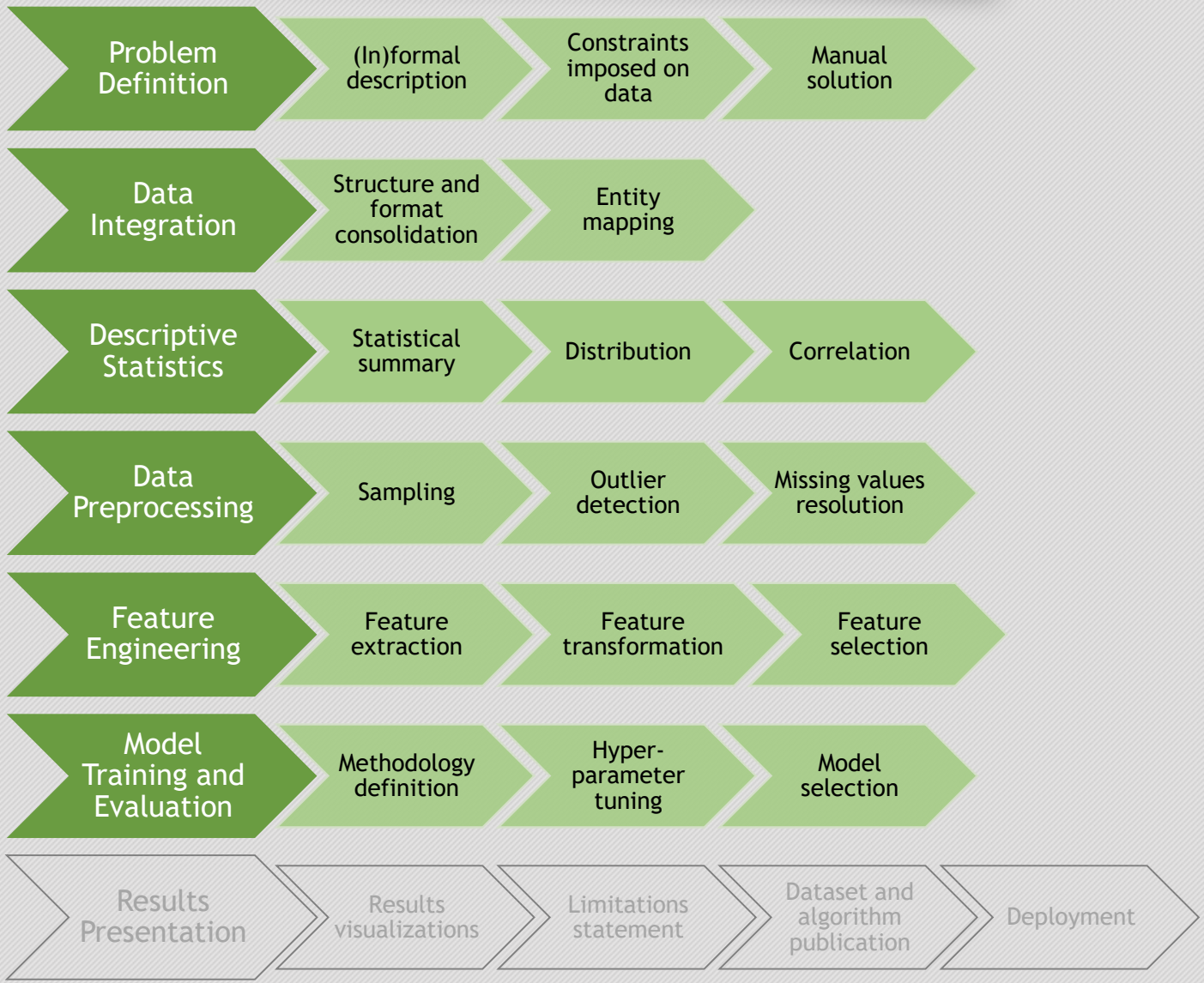
ML Workflow: Feature selection

- Feature construction can lead to huge number of features
- Techniques
  - Filter methods
  - Wrapper methods
  - Embedded methods



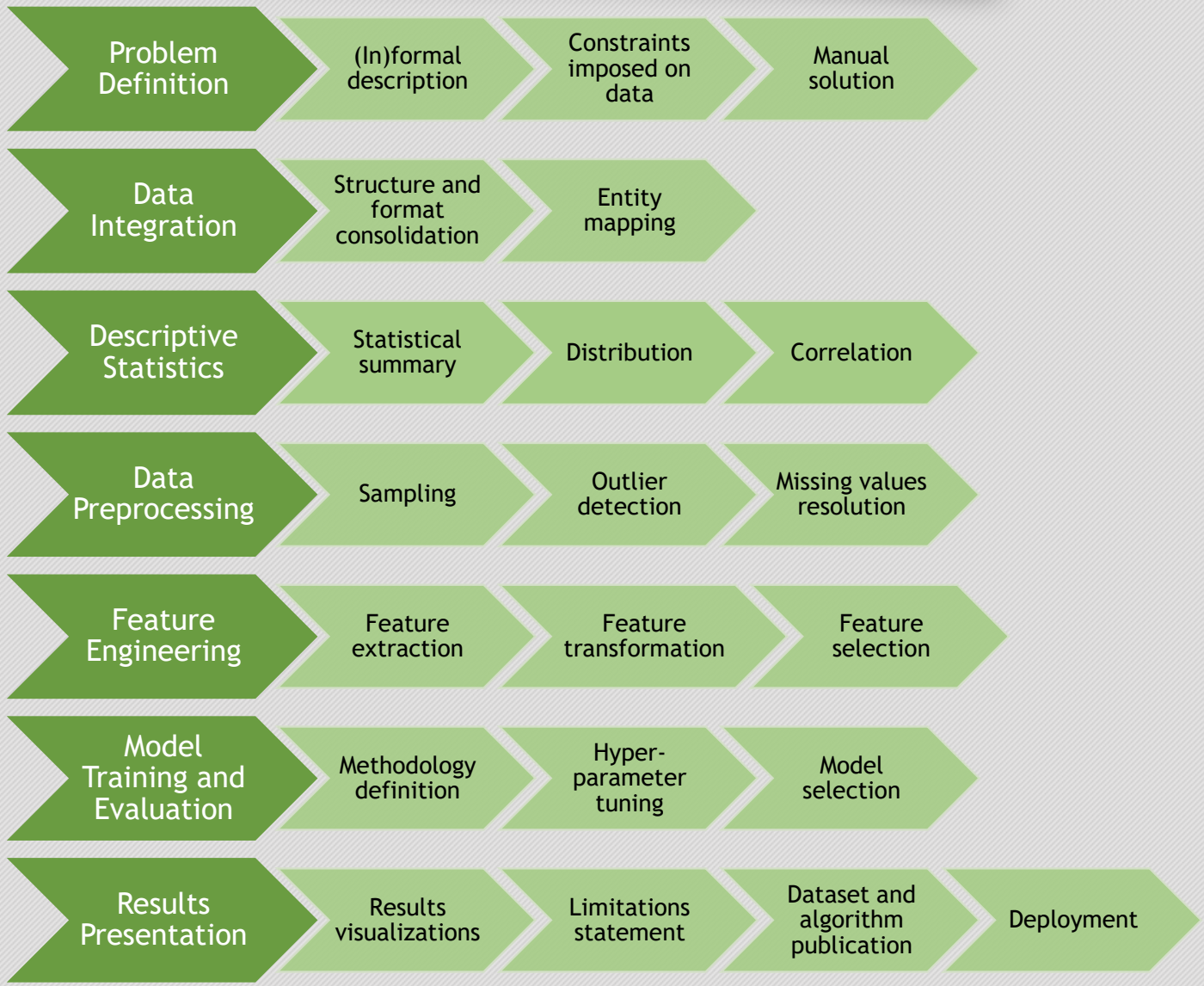
ML Workflow: Model Training and Evaluation

- Explicitly state your methodology for training and evaluation
  - Select and define metrics suitable for your ML task
  - Distinguish training, testing and validating sets
  - Use cross-validation if necessary
- Always do hyperparameter tuning
  - Never rely on default algorithm parameters
  - Grid-search, random-search, …
- Select best model according to your stated problem/goal



ML Workflow: Results Presentation

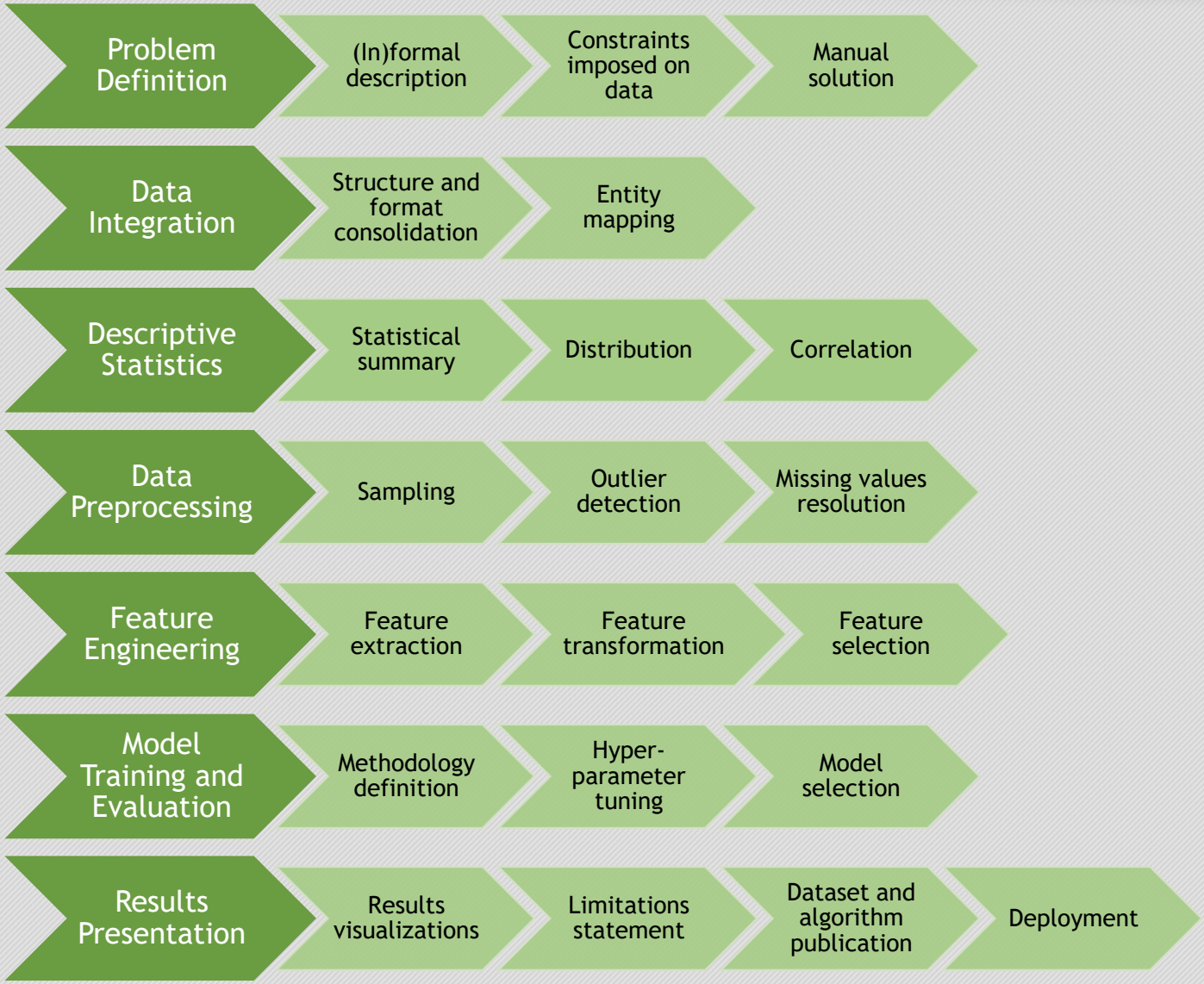
- Visualize and compare results
  - Jupiter Notebooks (Python, R, …)
- Always admit limitations of your methods
- Best results are 100% reproducible
  - Publish your dataset and algorithm (e.g. Github + Jupiter Notebook)
- If applicable, deploy the algorithm online, measure its performance and continue improving
  - Or describe typical use cases



ML Workflow: Typical Problems

- Curse of dimensionality
  - Dimensionality reduction, …
- Overfitting
  - Regularization parameters, pruning decision trees, …
- Imbalanced datasets
  - Under-sampling, over-sampling, different weights
- Concept drift

ML Workflow: Discussion



- Not all steps are necessary
  - Descriptive statistics and model evaluation will show you your way
- Never do waterfall
  - Iteratively improve your features and model
  - Incrementally try new features and algorithms
- Feature engineering is a real art and mastery of applied ML

We will discuss particular techniques and details about ML problems during next seminars…

Sources

- https://machinelearningmastery.com/4-steps-to-get-started-in-machine-learning/