Methodological topics
Data-science specifics (part 1)

datalys

Ivan Srba

30th October 2019

PeWe@FIIT
personalized web group

Warming-up

- Everything said on 9th October during presentation on *Methodological topics (Part 1)* applies perfectly also in case of all theses in data science domain

Warming-up

- Everything said on 9th October during presentation on *Methodological topics (Part 1)* applies perfectly also in case of all theses in data science domain
- Summary of gold rules
  - Search for sources (research articles), organize them by dedicated tools

Warming-up

- Everything said on 9th October during presentation on *Methodological topics (Part 1)* applies perfectly also in case of all theses in data science domain
- Summary of gold rules
    - Search for sources (research articles), organize them by dedicated tools
    - Analyze the existing solutions, write notes, compare them

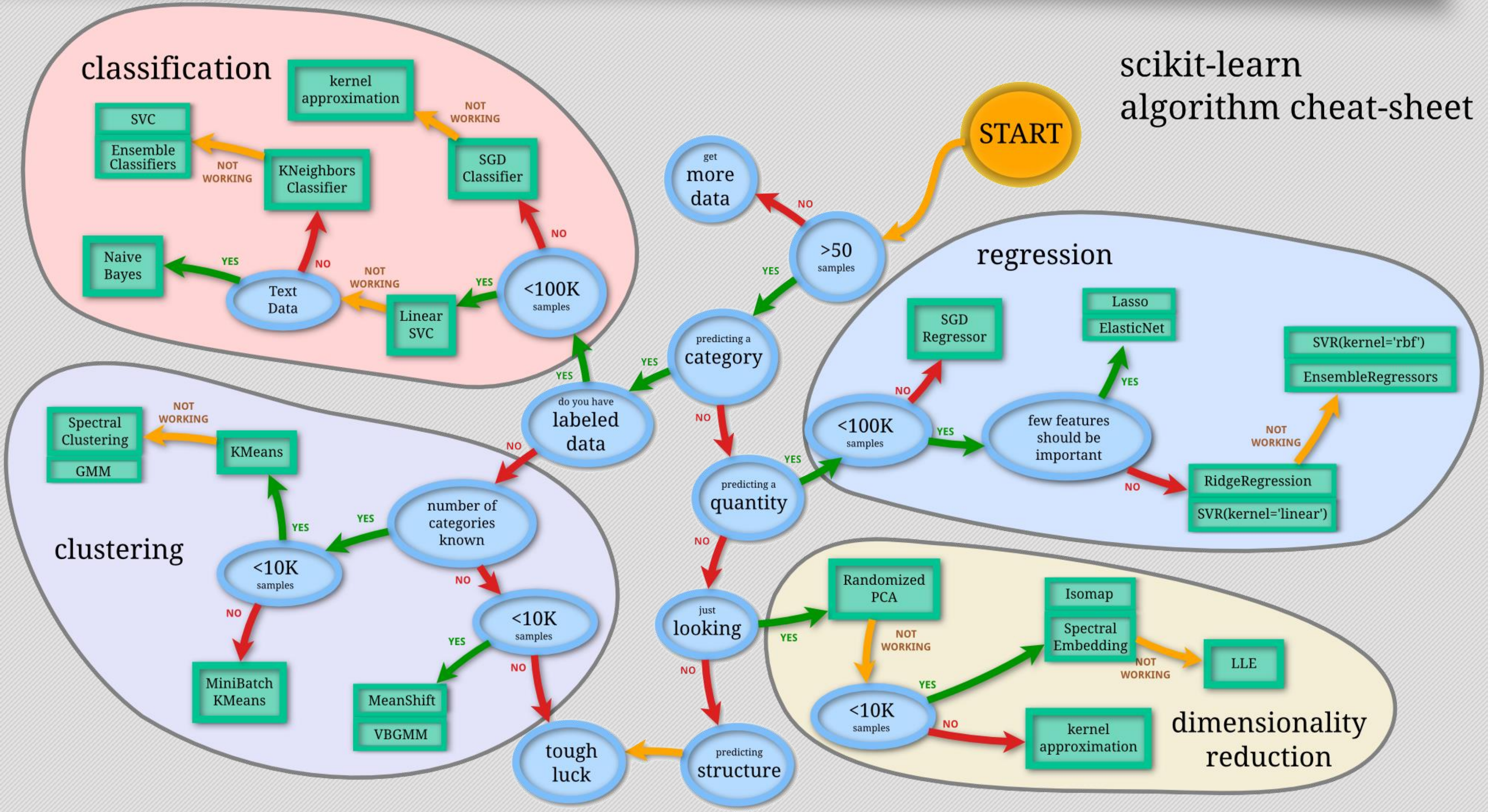
Warming-up

- Everything said on 9th October during presentation on *Methodological topics (Part 1)* applies perfectly also in case of all theses in data science domain
- Summary of gold rules
  - Search for sources (research articles), organize them by dedicated tools
  - Analyze the existing solutions, write notes, compare them
  - Select few most related articles, describe them in very details

Warming-up

- Everything said on 9th October during presentation on *Methodological topics (Part 1)* applies perfectly also in case of all theses in data science domain
- Summary of gold rules
  - Search for sources (research articles), organize them by dedicated tools
  - Analyze the existing solutions, write notes, compare them
  - Select few most related articles, describe them in very details
  - Pay a strong attention to summary/discussion at the end of analyses’ section

Scikit-learn algorithm cheat-sheet



scikit-learn
algorithm cheat-sheet

START

>50 samples

NO

get more data

YES

predicting a category

YES

do you have labeled data

YES

classification

<100K samples

YES

Linear SVC

NOT WORKING

Text Data

YES

Naive Bayes

NO

KNeighbors Classifier

NOT WORKING

SVC

Ensemble Classifiers

NO

SGD Classifier

NOT WORKING

kernel approximation

NO

clustering

number of categories known

YES

<10K samples

YES

KMeans

NOT WORKING

Spectral Clustering

GMM

NO

MiniBatch KMeans

NO

<10K samples

YES

MeanShift

VBGMM

NO

tough luck

NO

predicting a quantity

YES

regression

<100K samples

NO

SGD Regressor

YES

few features should be important

YES

Lasso

ElasticNet

NO

RidgeRegression

SVR(kernel='linear')

NOT WORKING

SVR(kernel='rbf')

EnsembleRegressors

NO

just looking

YES

dimensionality reduction

Randomized PCA

NOT WORKING

<10K samples

YES

Isomap

Spectral Embedding

NOT WORKING

LLE

NO

kernel approximation

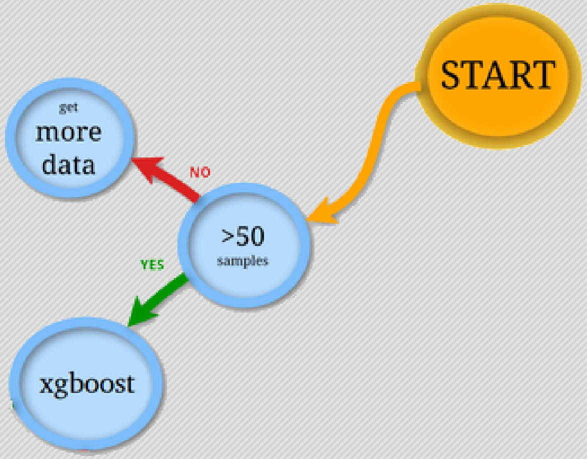
NO

predicting structure

tough luck

http://scikit-learn.org/stable/tutorial/machine_learning_map/

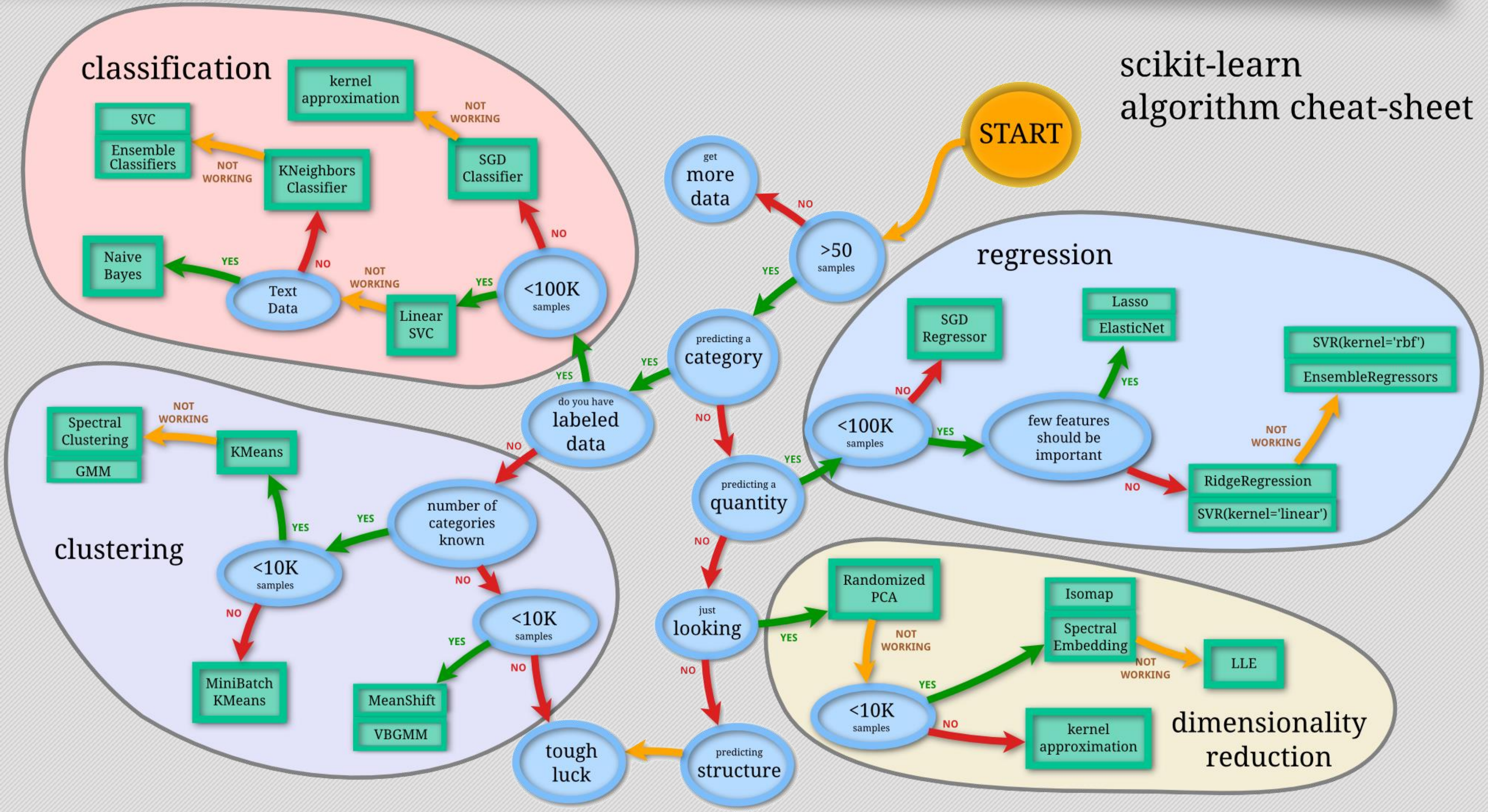
Scikit-learn algorithm cheat-sheet - simplified



scikit-learn
algorithm cheat-sheet

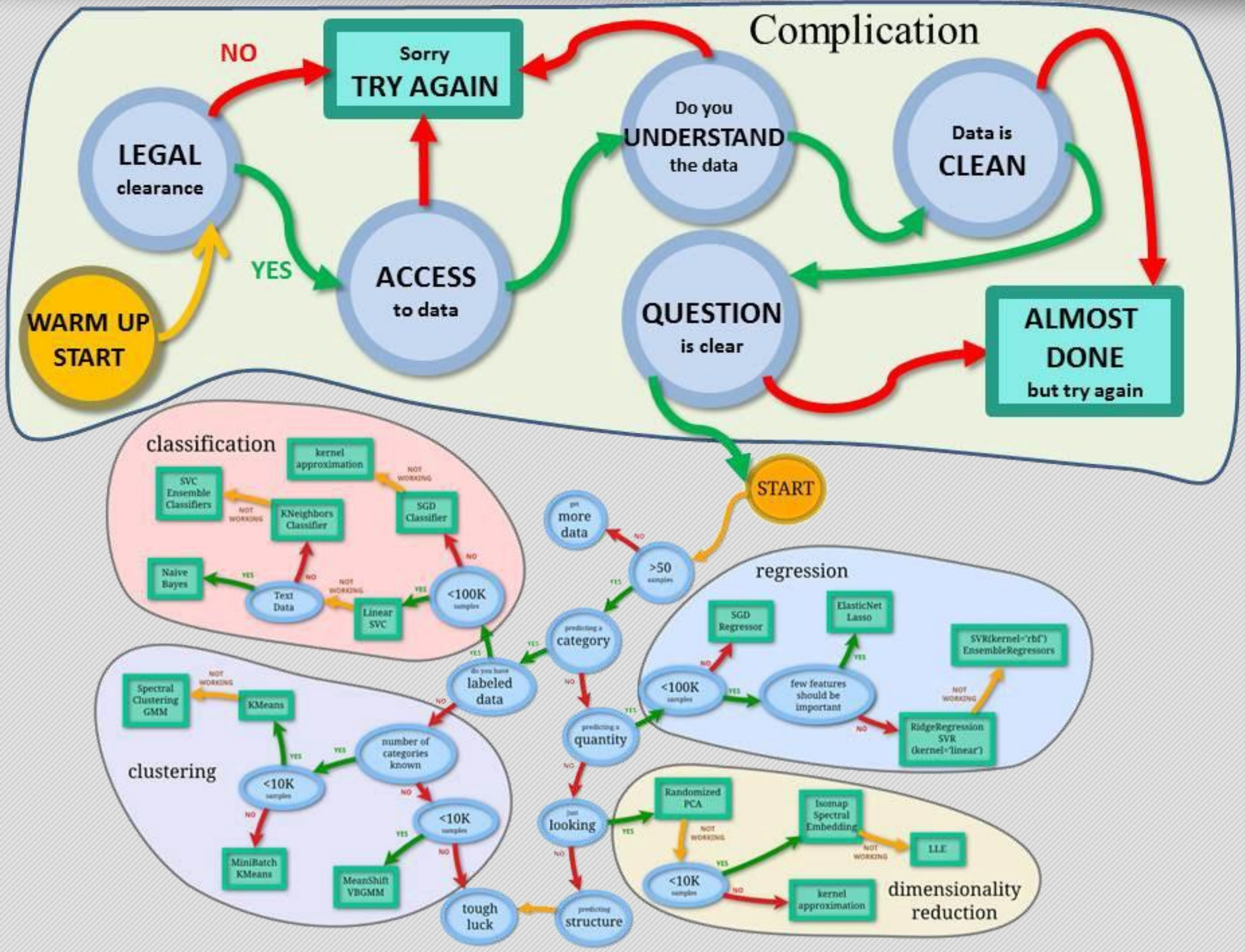
https://twitter.com/bmurr26/status/707103333408833536

Scikit-learn algorithm cheat-sheet



http://scikit-learn.org/stable/tutorial/machine_learning_map/

Scikit-learn algorithm cheat-sheet - extended

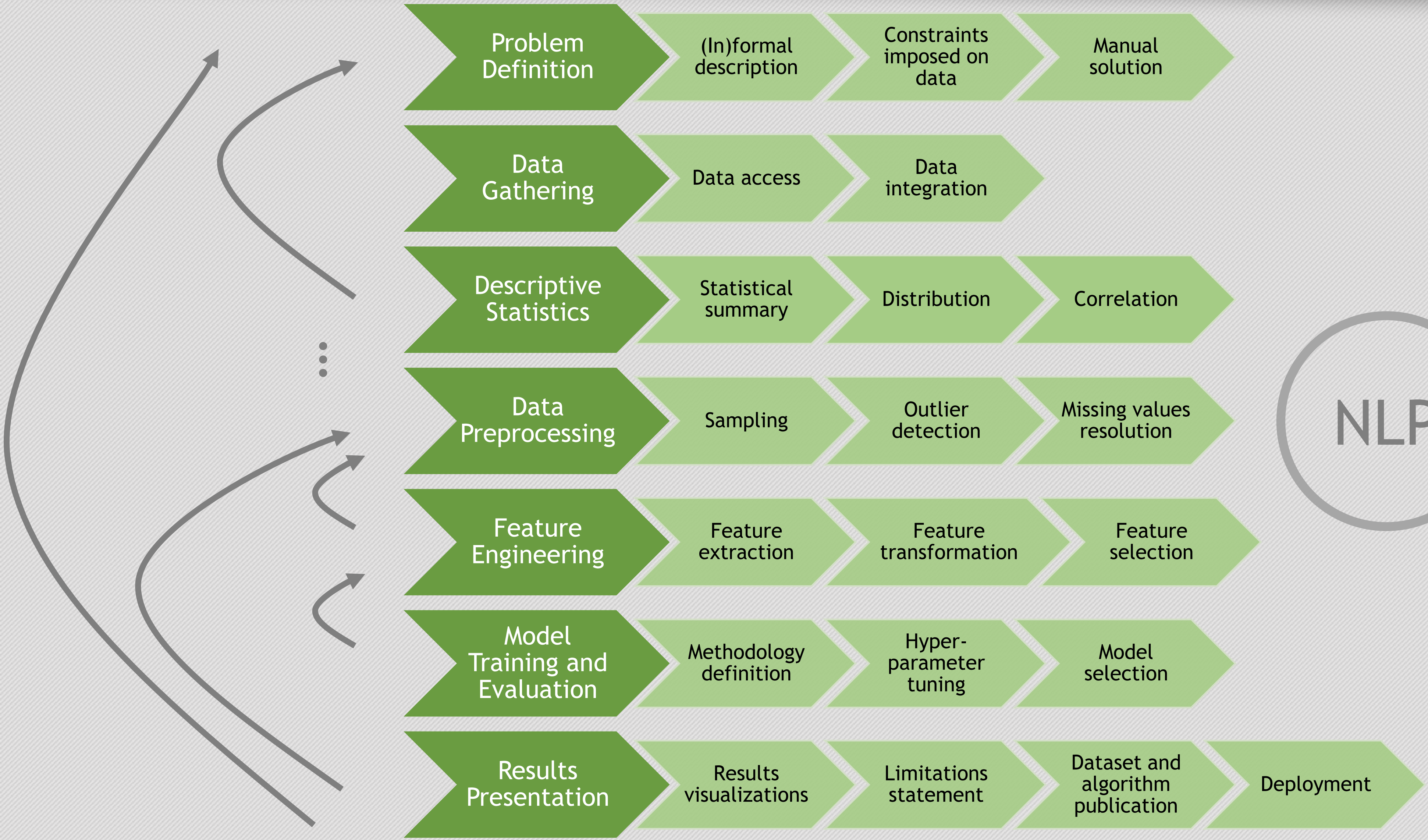


https://medium.com/@chris_bour/an-extended-version-of-the-scikit-learn-cheat-sheet-5f46efc6cbb

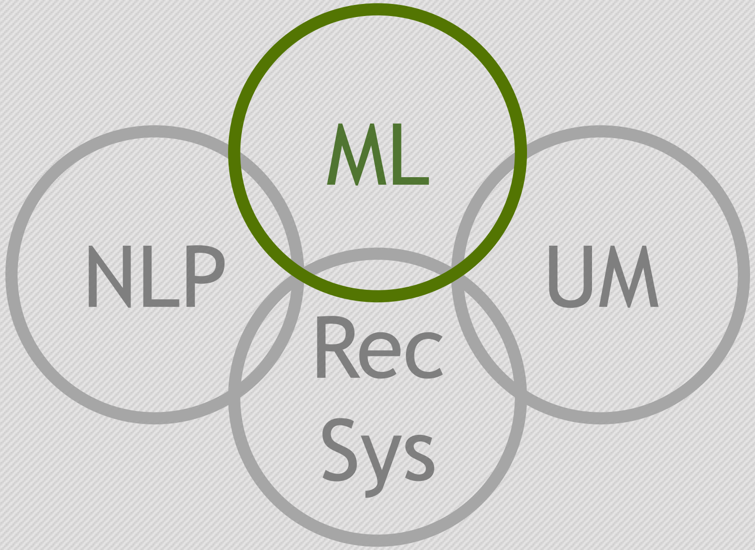
Data-science specific questions…

- … you need to answer before starting work on solution proposal and implementation:
  - How to define data-science (machine learning, …) task?
  - How to select/create appropriate dataset?
  - How to describe your dataset?
  - How to preprocess your dataset?
  - …

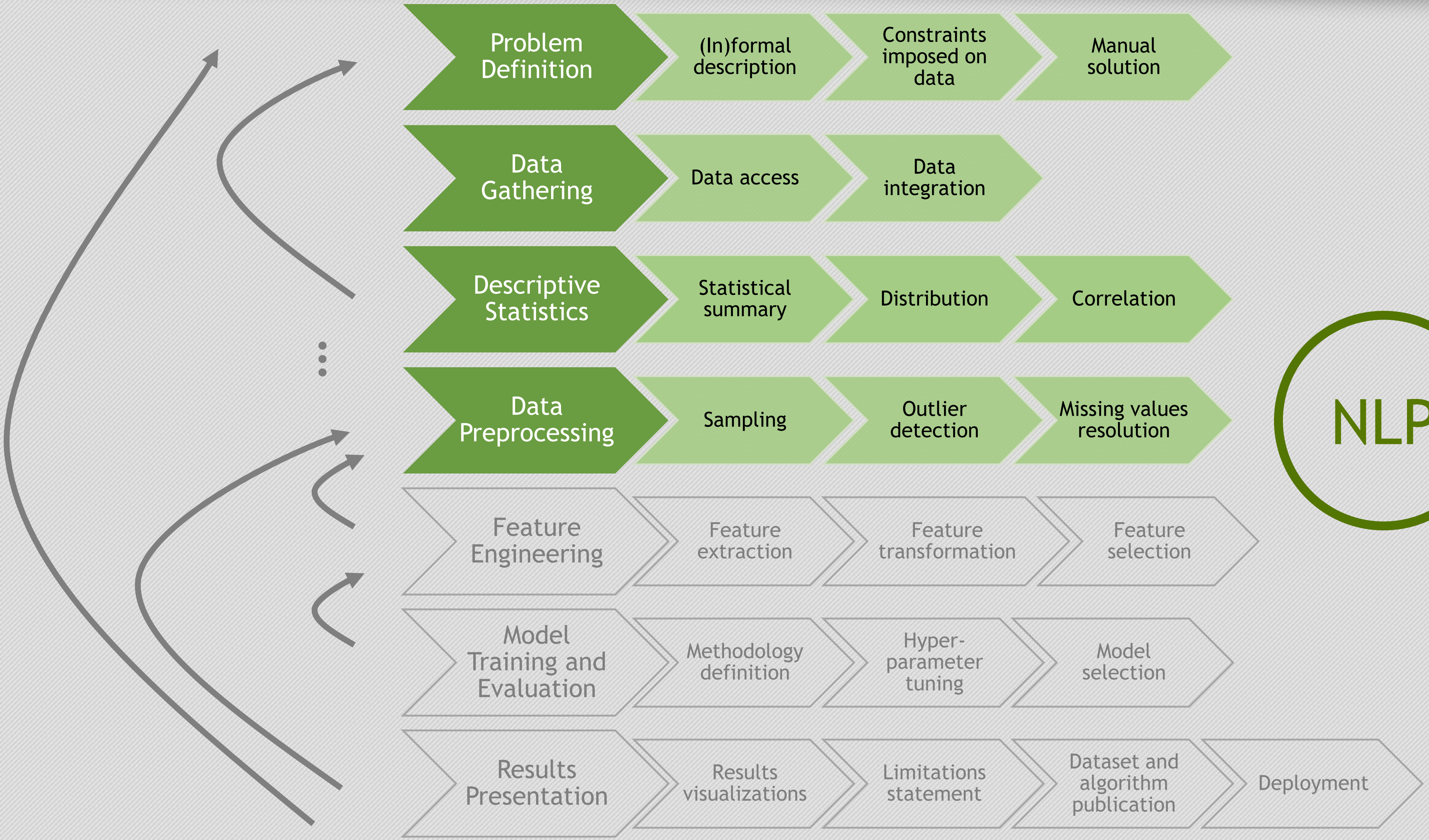
ML Workflow: Overview



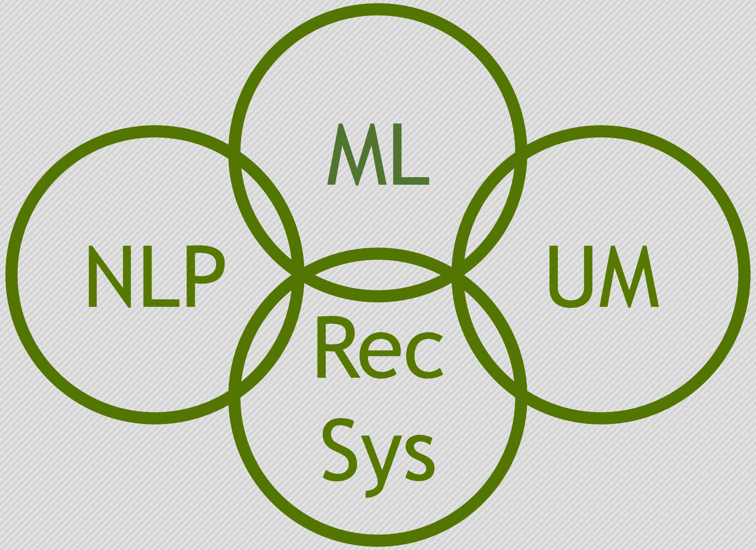
- Problem Definition
  - (In)formal description
  - Constraints imposed on data
  - Manual solution
- Data Gathering
  - Data access
  - Data integration
- Descriptive Statistics
  - Statistical summary
  - Distribution
  - Correlation
- Data Preprocessing
  - Sampling
  - Outlier detection
  - Missing values resolution
- Feature Engineering
  - Feature extraction
  - Feature transformation
  - Feature selection
- Model Training and Evaluation
  - Methodology definition
  - Hyper-parameter tuning
  - Model selection
- Results Presentation
  - Results visualizations
  - Limitations statement
  - Dataset and algorithm publication
  - Deployment



ML Workflow: Generic part

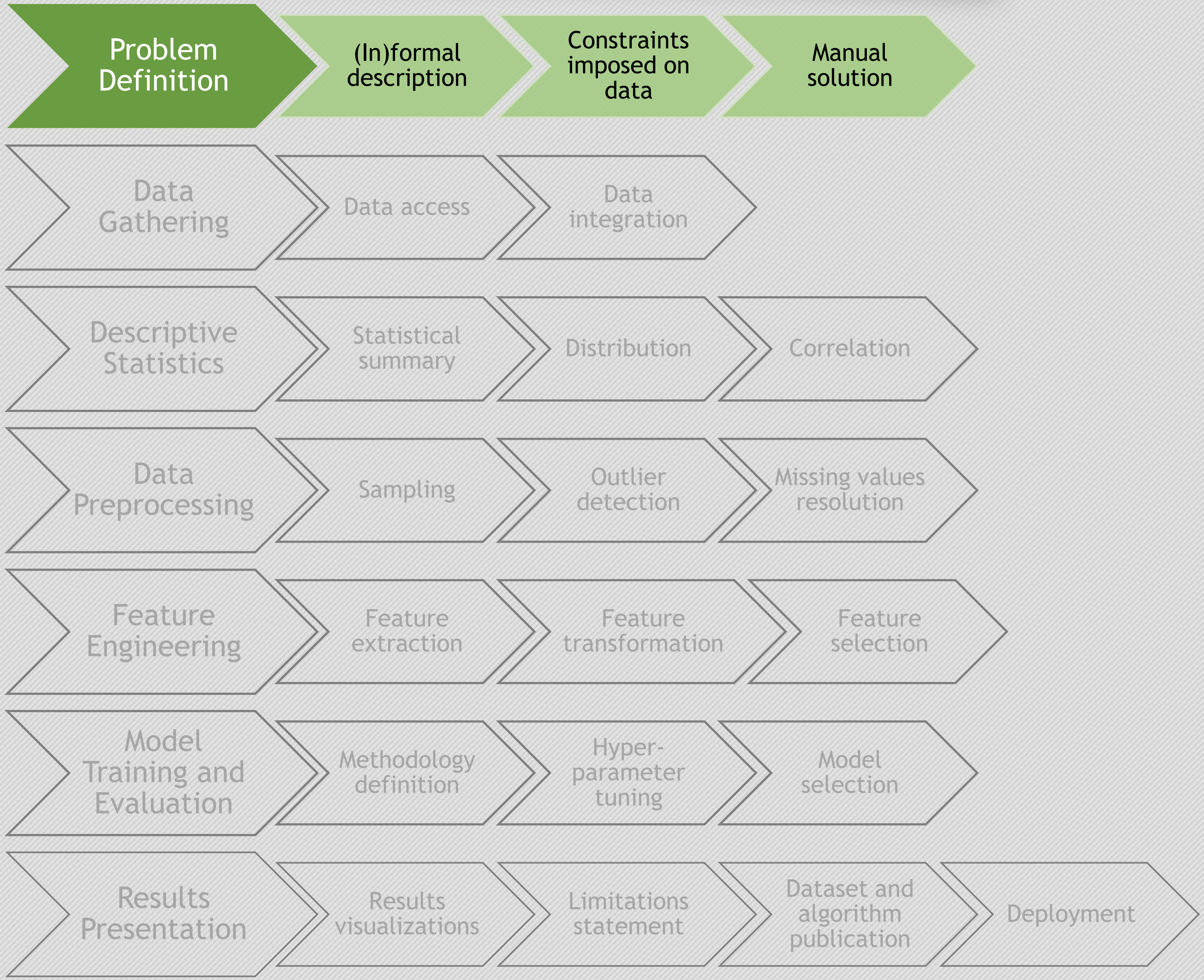


- **Problem Definition**: (In)formal description → Constraints imposed on data → Manual solution
- **Data Gathering**: Data access → Data integration
- **Descriptive Statistics**: Statistical summary → Distribution → Correlation
- ⋮
- **Data Preprocessing**: Sampling → Outlier detection → Missing values resolution
- Feature Engineering: Feature extraction → Feature transformation → Feature selection
- Model Training and Evaluation: Methodology definition → Hyper-parameter tuning → Model selection
- Results Presentation: Results visualizations → Limitations statement → Dataset and algorithm publication → Deployment



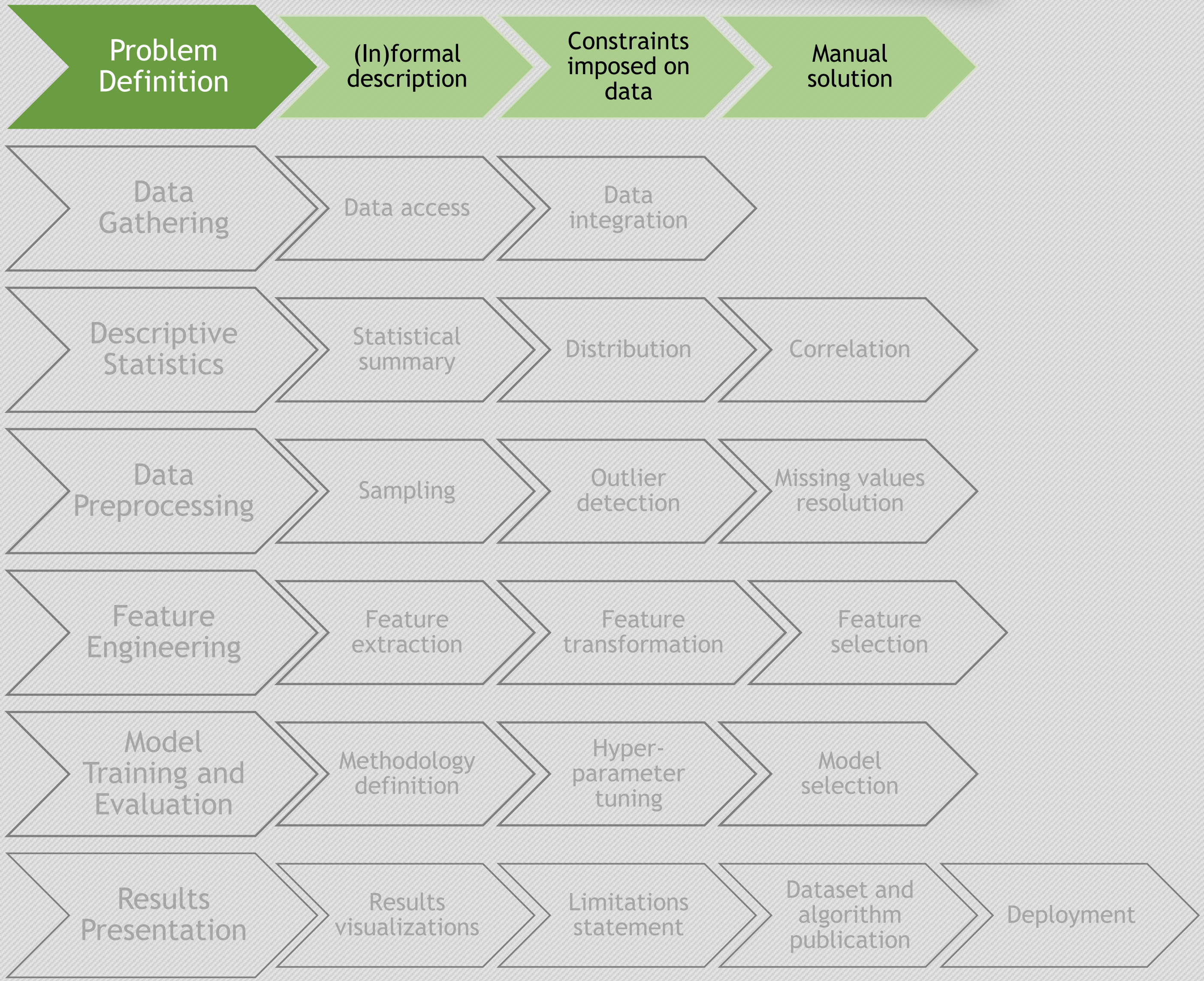
Step 1: Problem Definition

- Clearly identify your problem you are trying to solve
  - Informal description
    - As you would explain it to your friends
    - Refer back to motivation stated in analyses' summary/discussion
  - Formal description
    - As a research question
    - As a hypothesis
    - As a machine learning task



Step 1: Problem Definition

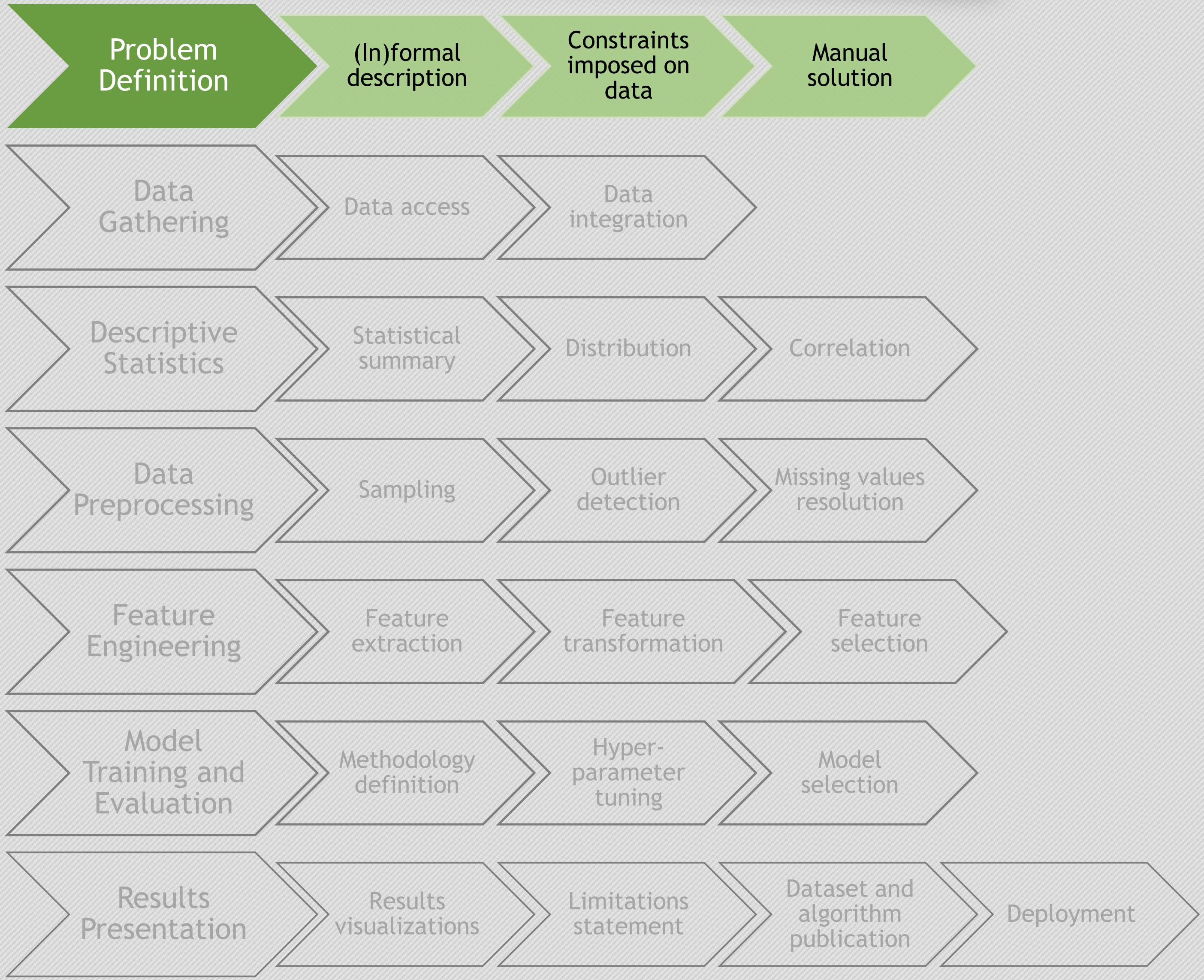
- Clearly identify your problem you are trying to solve
  - Informal description
    - As you would explain it to your friends
    - Refer back to motivation stated in analyses' summary/discussion
  - Formal description
    - As a research question
    - As a hypothesis
    - As a machine learning task
- Identify constraints imposed on required dataset



Problem Definition	(In)formal description	Constraints imposed on data	Manual solution	
Data Gathering	Data access	Data integration		
Descriptive Statistics	Statistical summary	Distribution	Correlation	
Data Preprocessing	Sampling	Outlier detection	Missing values resolution	
Feature Engineering	Feature extraction	Feature transformation	Feature selection	
Model Training and Evaluation	Methodology definition	Hyper-parameter tuning	Model selection	
Results Presentation	Results visualizations	Limitations statement	Dataset and algorithm publication	Deployment

Step 1: Problem Definition

- Clearly identify your problem you are trying to solve
  - Informal description
    - As you would explain it to your friends
    - Refer back to motivation stated in analyses’ summary/discussion
  - Formal description
    - As a research question
    - As a hypothesis
    - As a machine learning task
- 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)



Problem Definition → (In)formal description → Constraints imposed on data → Manual solution

Data Gathering → Data access → Data integration

Descriptive Statistics → Statistical summary → Distribution → Correlation

Data Preprocessing → Sampling → Outlier detection → Missing values resolution

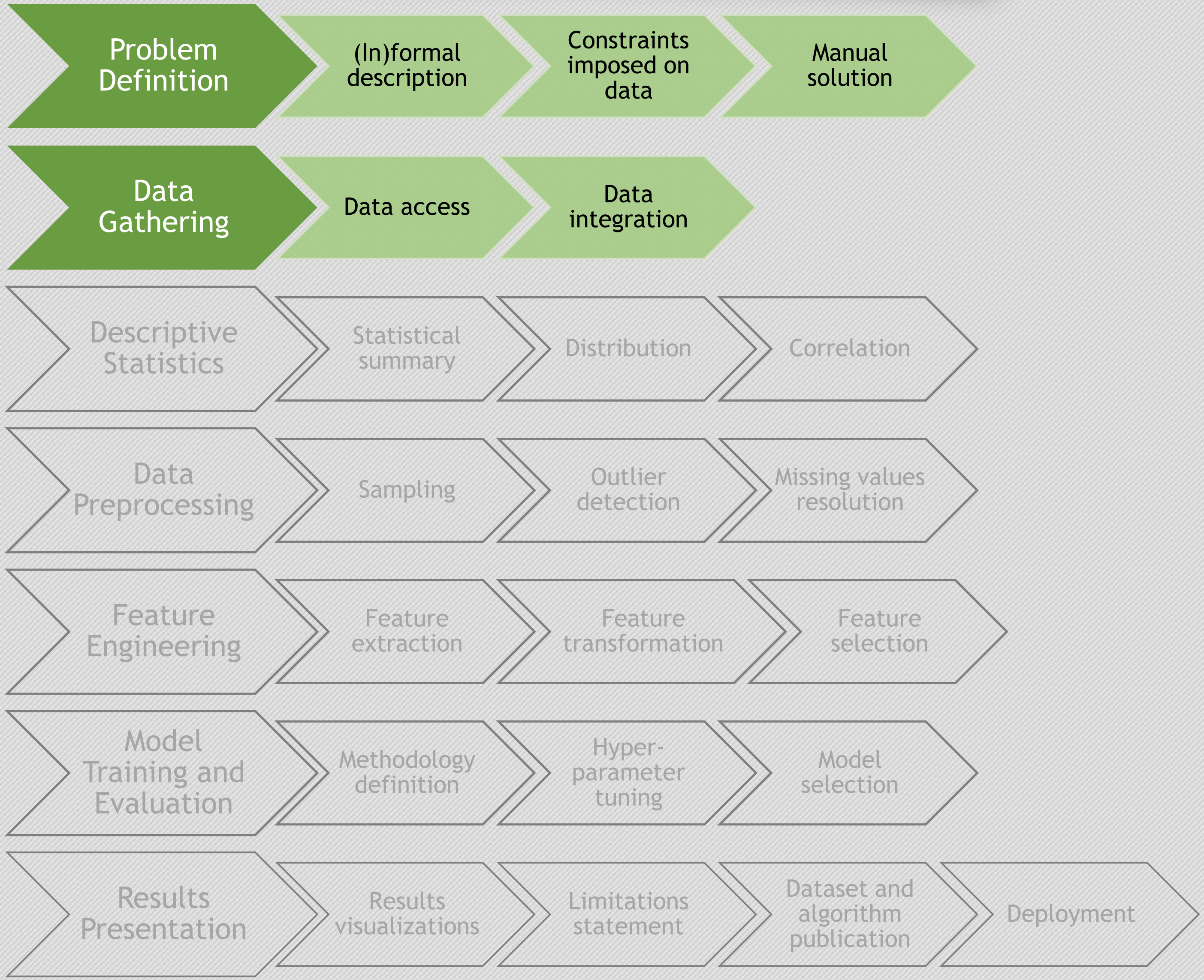
Feature Engineering → Feature extraction → Feature transformation → Feature selection

Model Training and Evaluation → Methodology definition → Hyper-parameter tuning → Model selection

Results Presentation → Results visualizations → Limitations statement → Dataset and algorithm publication → Deployment

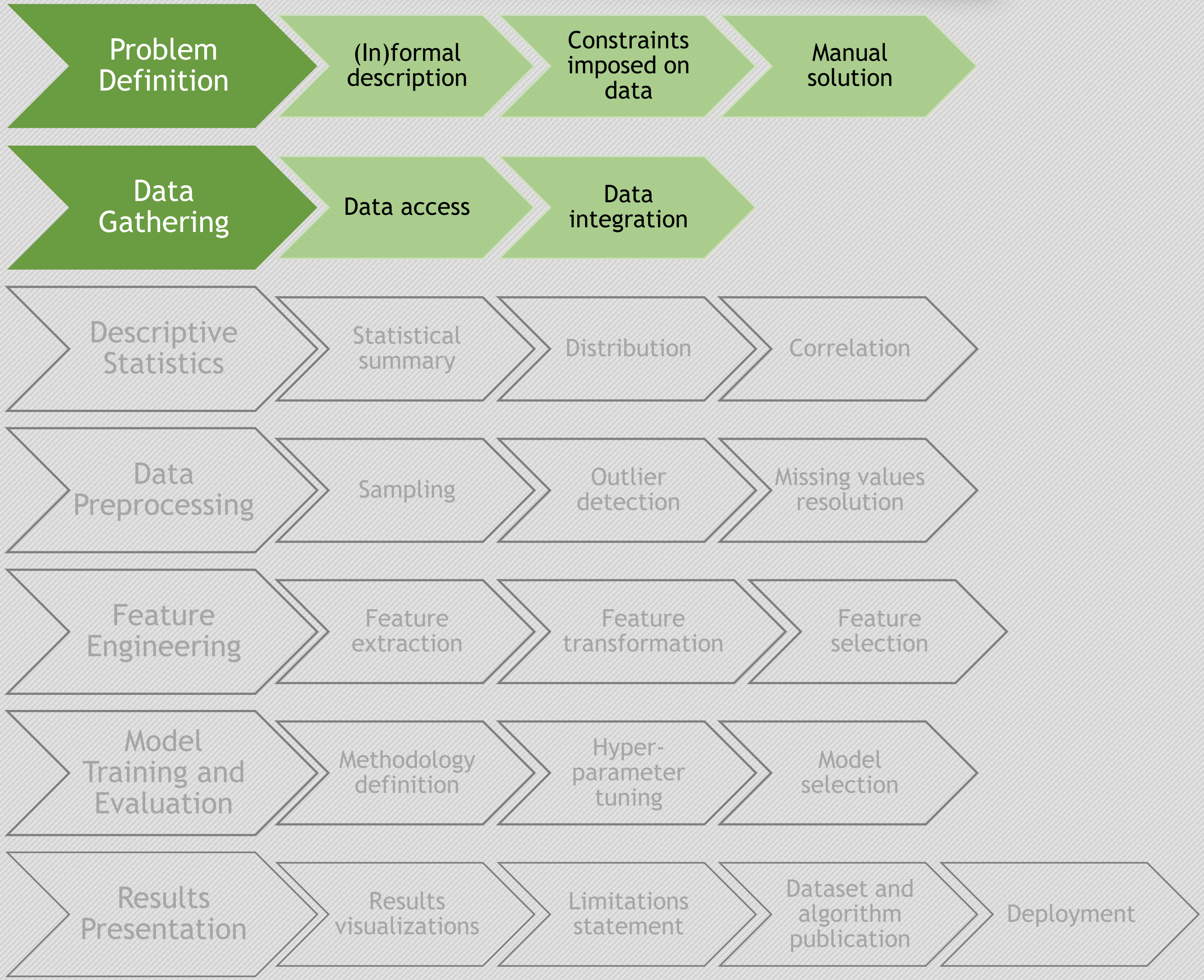
Step 2: Data Gathering

- Data access
  - Prepared datasets, crawling, API
  - Legal issues



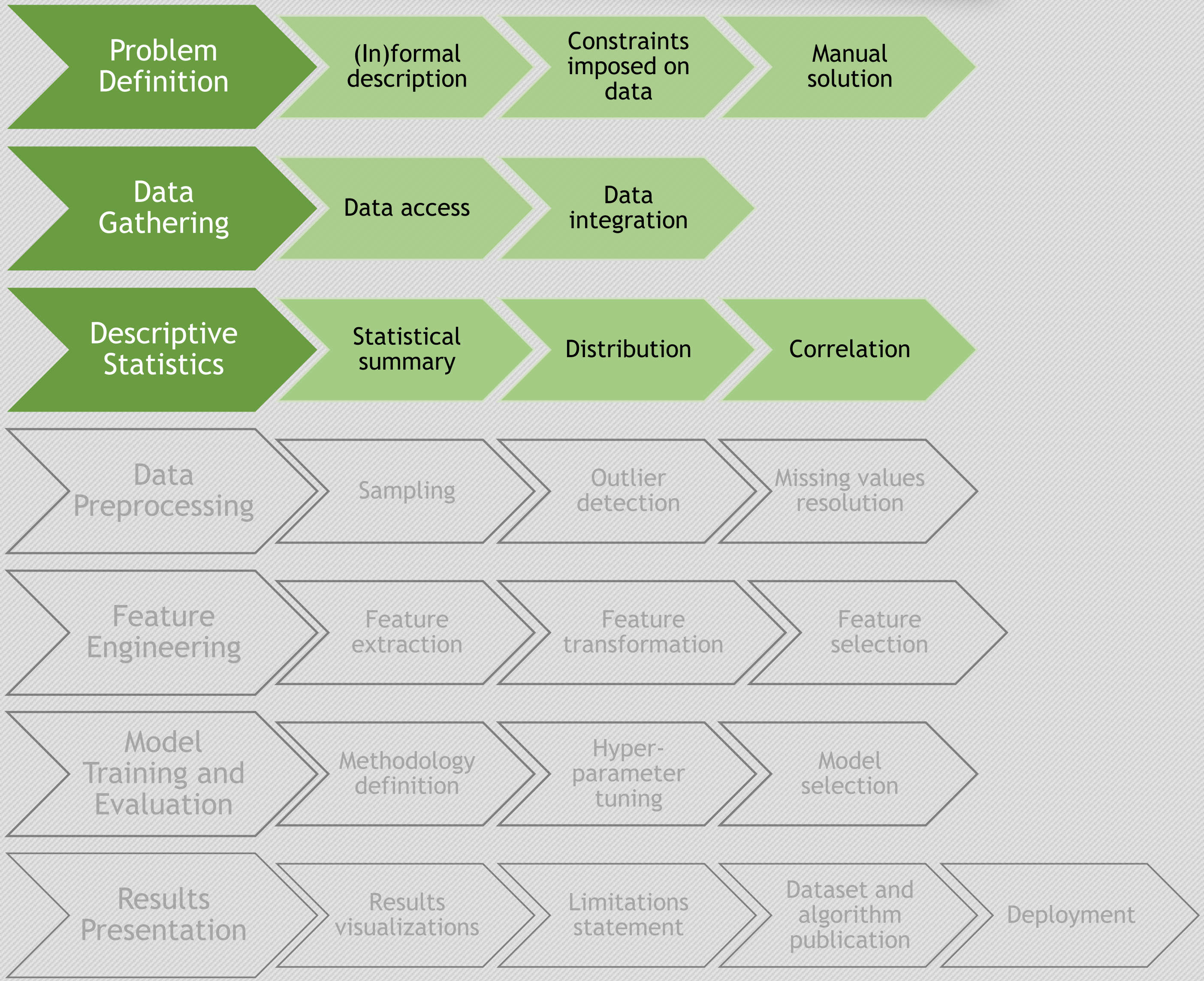
Step 2: Data Integration

- Data access
  - Prepared datasets, crawling, API
  - Legal issues
- Data integration
  - Some ML task can be solved only when you integrate data from several sources
    - Different sources = different structure and format
  - Data consolidation
    - Entity mapping
    - User IDs (email, DB ID, cookie, username, …)
    - Item IDs (code, name, …)



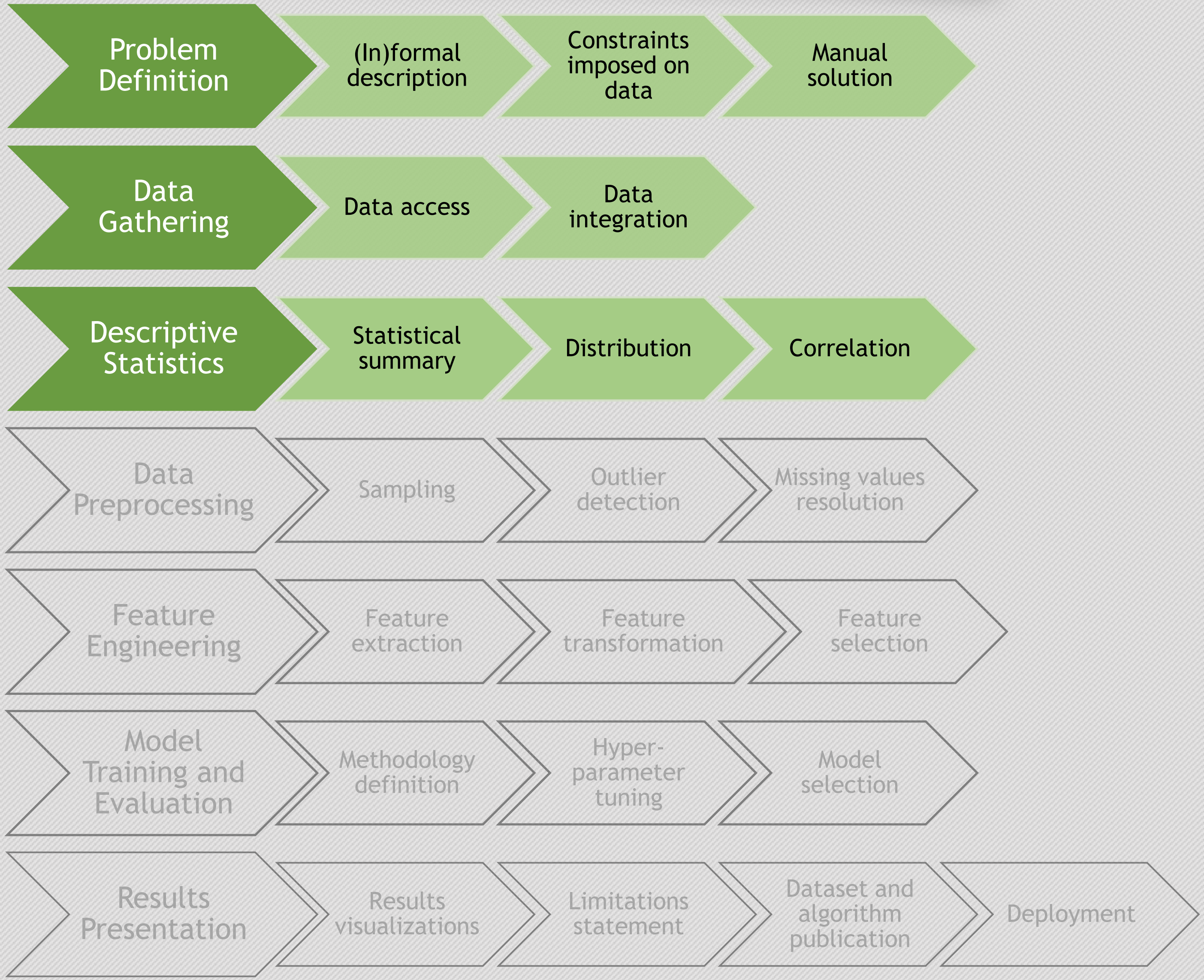
Step 3: Descriptive Statistics

- Known your data!
  - Otherwise, you are just guessing…



Step 3: 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



Problem Definition → (In)formal description → Constraints imposed on data → Manual solution

Data Gathering → Data access → Data integration

Descriptive Statistics → Statistical summary → Distribution → Correlation

Data Preprocessing → Sampling → Outlier detection → Missing values resolution

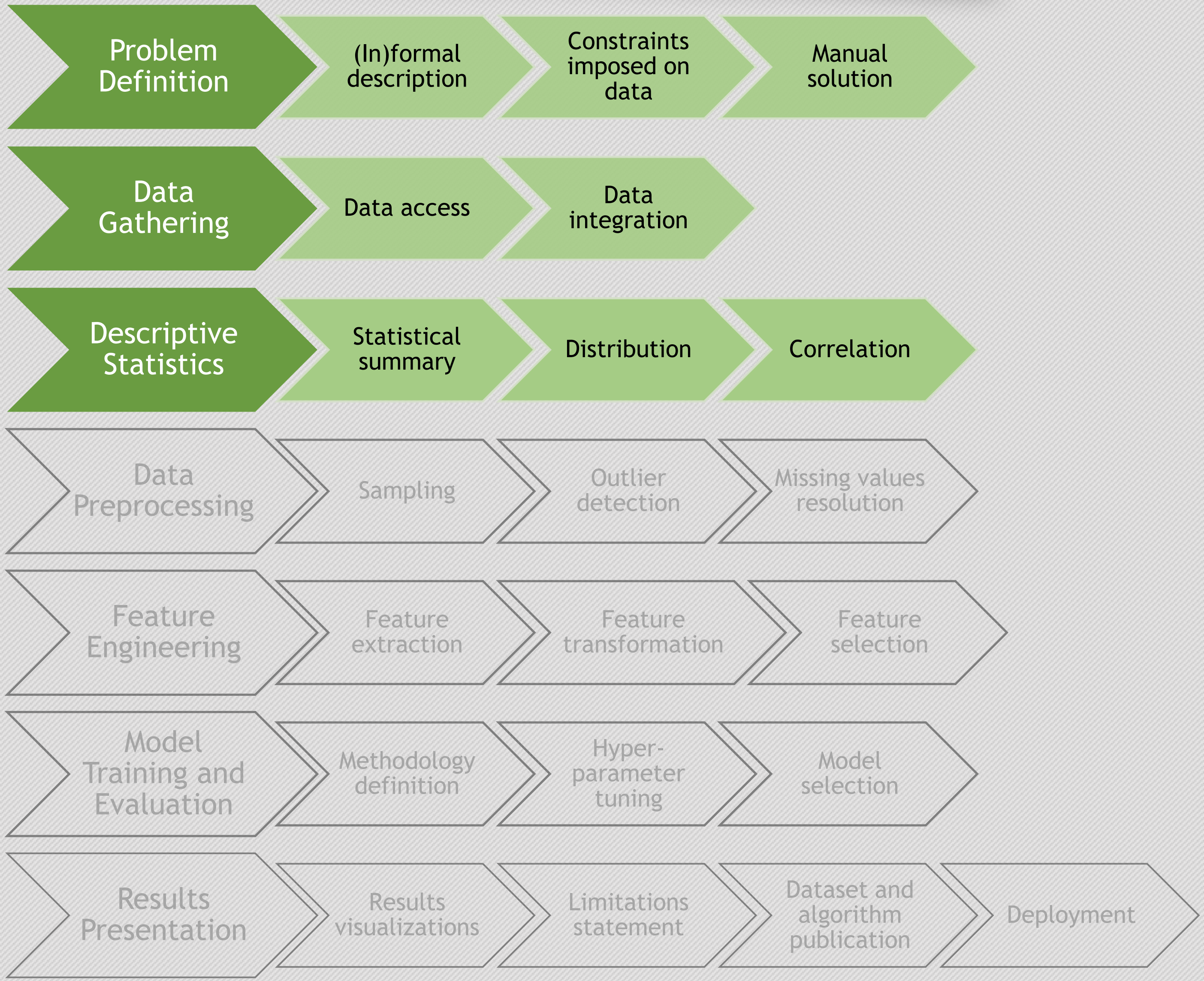
Feature Engineering → Feature extraction → Feature transformation → Feature selection

Model Training and Evaluation → Methodology definition → Hyper-parameter tuning → Model selection

Results Presentation → Results visualizations → Limitations statement → Dataset and algorithm publication → Deployment

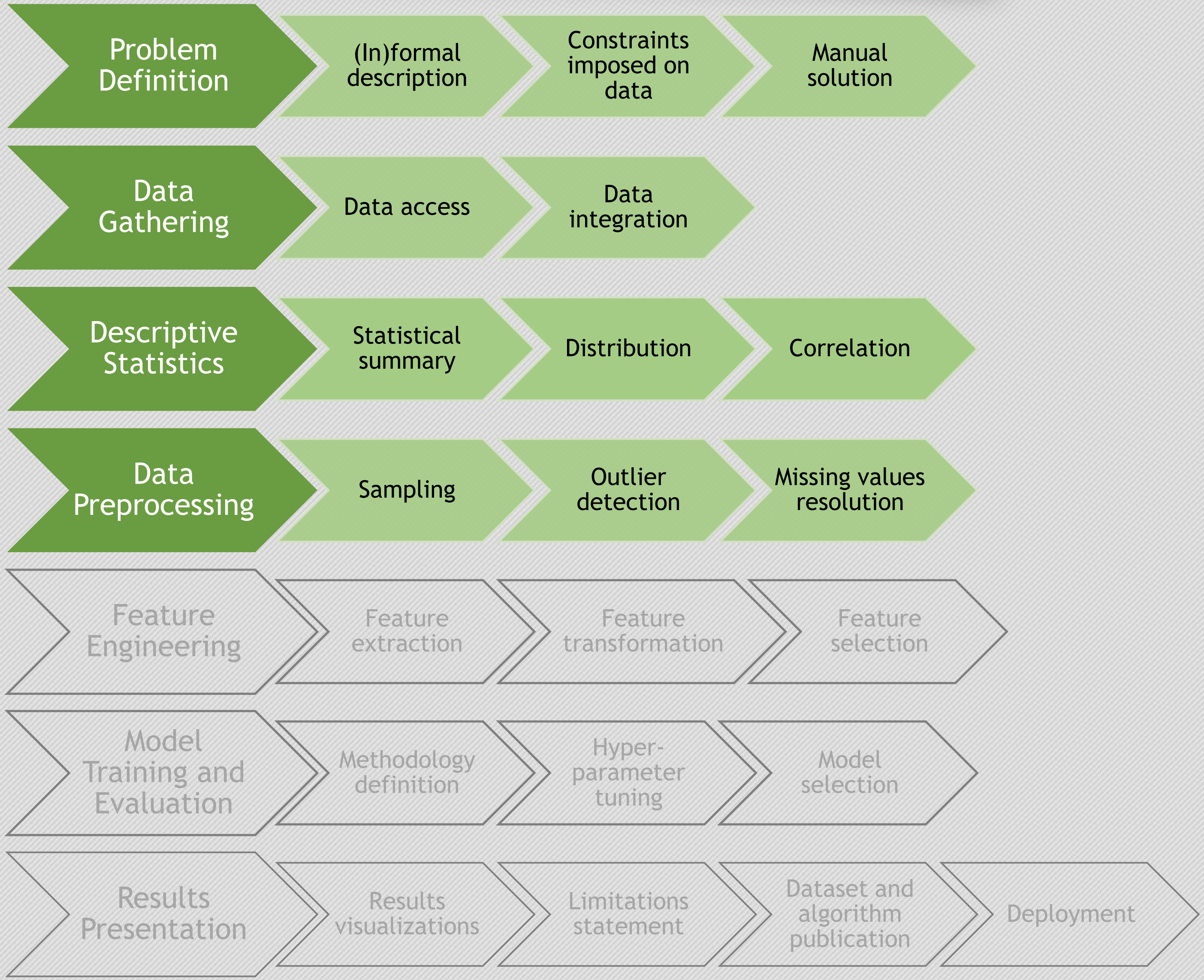
Step 3: Descriptive Statistics

- 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



Step 4: 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



Problem Definition – (In)formal description – Constraints imposed on data – Manual solution

Data Gathering – Data access – Data integration

Descriptive Statistics – Statistical summary – Distribution – Correlation

Data Preprocessing – Sampling – Outlier detection – Missing values resolution

Feature Engineering – Feature extraction – Feature transformation – Feature selection

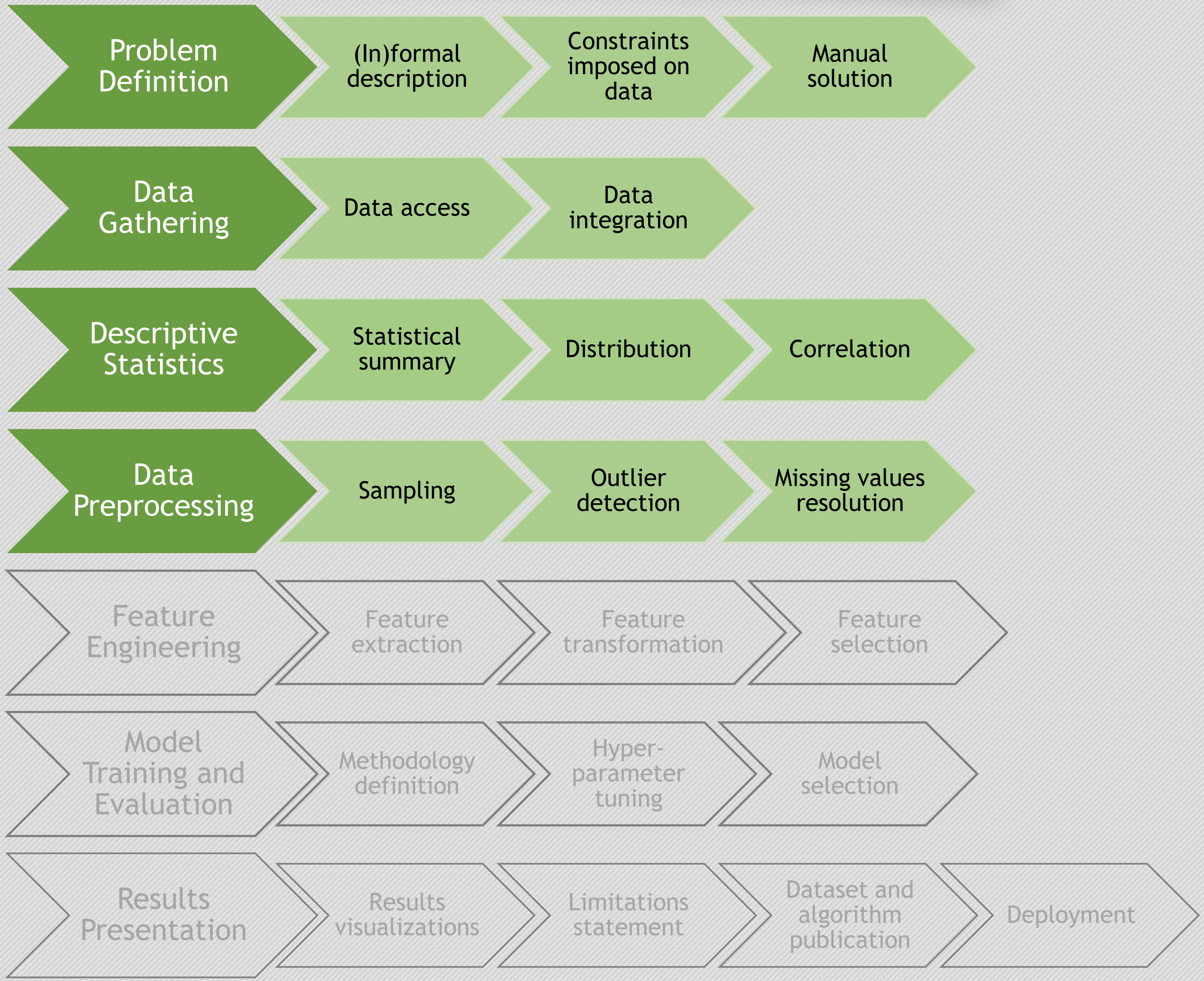
Model Training and Evaluation – Methodology definition – Hyper-parameter tuning – Model selection

Results Presentation – Results visualizations – Limitations statement – Dataset and algorithm publication – Deployment

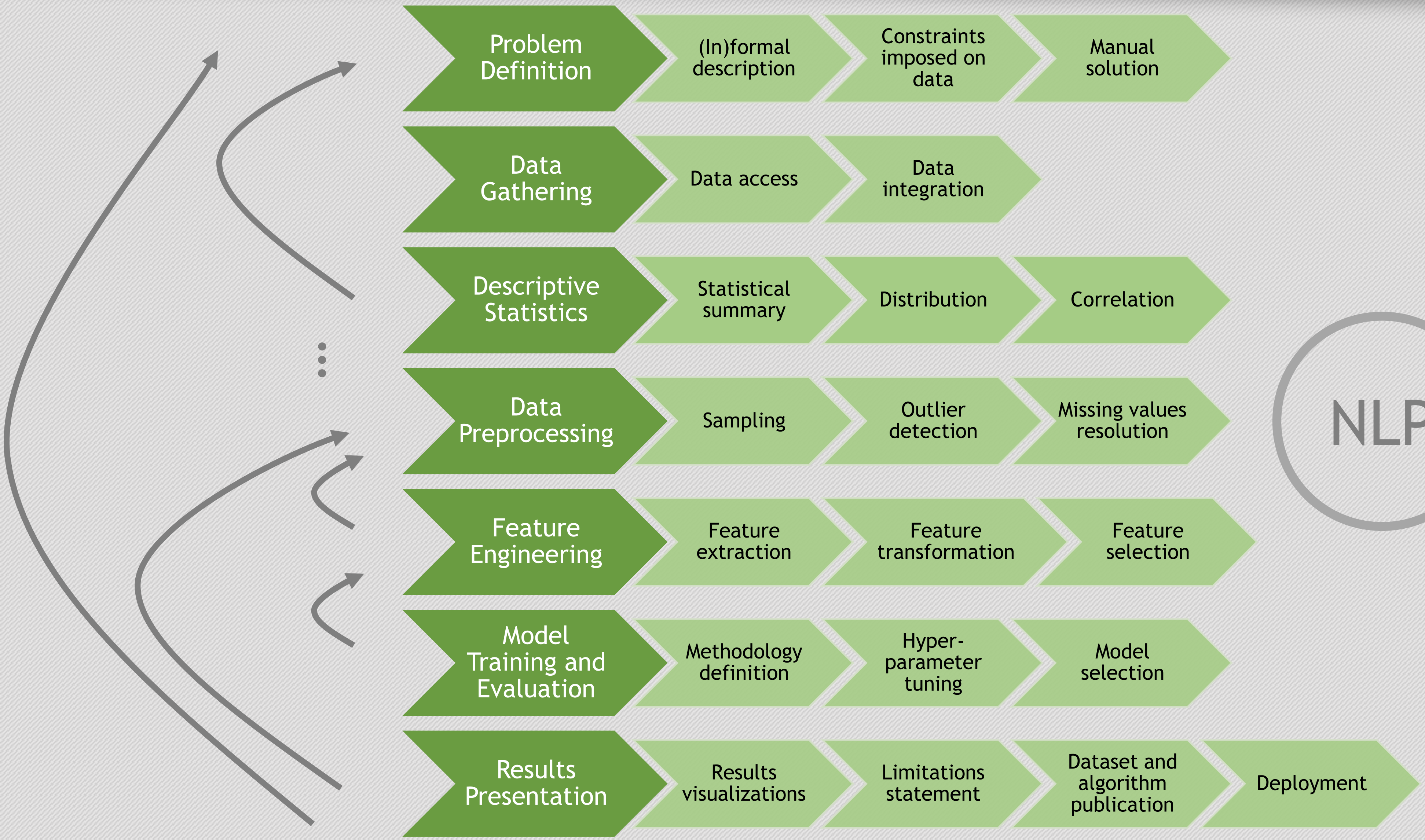
Step 4: Data Preprocessing

- Have large data? Do sampling!
  - Less data result in shorter training times
  - You can still finally run the model on larger portion of data

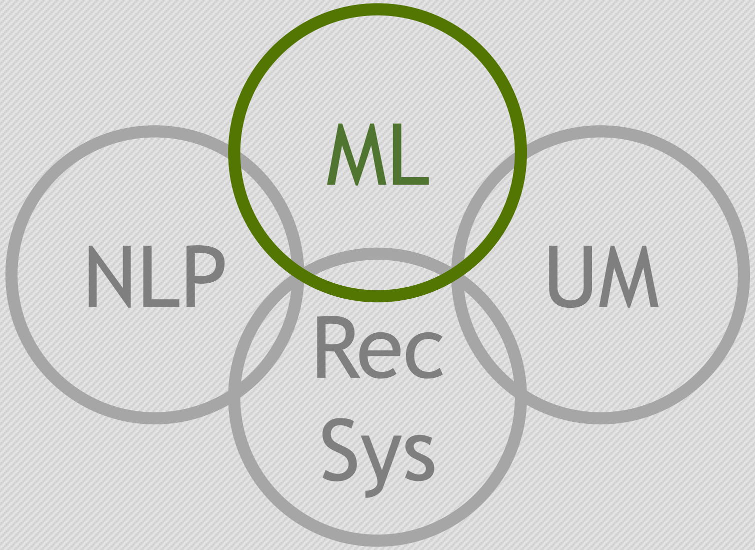
- Machine learning requires well-prepared data
  - Detect outliers
  - Replace missing values



ML Workflow: Overview



- **Problem Definition**: (In)formal description; Constraints imposed on data; Manual solution
- **Data Gathering**: Data access; Data integration
- **Descriptive Statistics**: Statistical summary; Distribution; Correlation
- **Data Preprocessing**: Sampling; Outlier detection; Missing values resolution
- **Feature Engineering**: Feature extraction; Feature transformation; Feature selection
- **Model Training and Evaluation**: Methodology definition; Hyper-parameter tuning; Model selection
- **Results Presentation**: Results visualizations; Limitations statement; Dataset and algorithm publication; Deployment



Sources

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