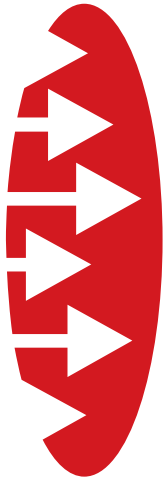
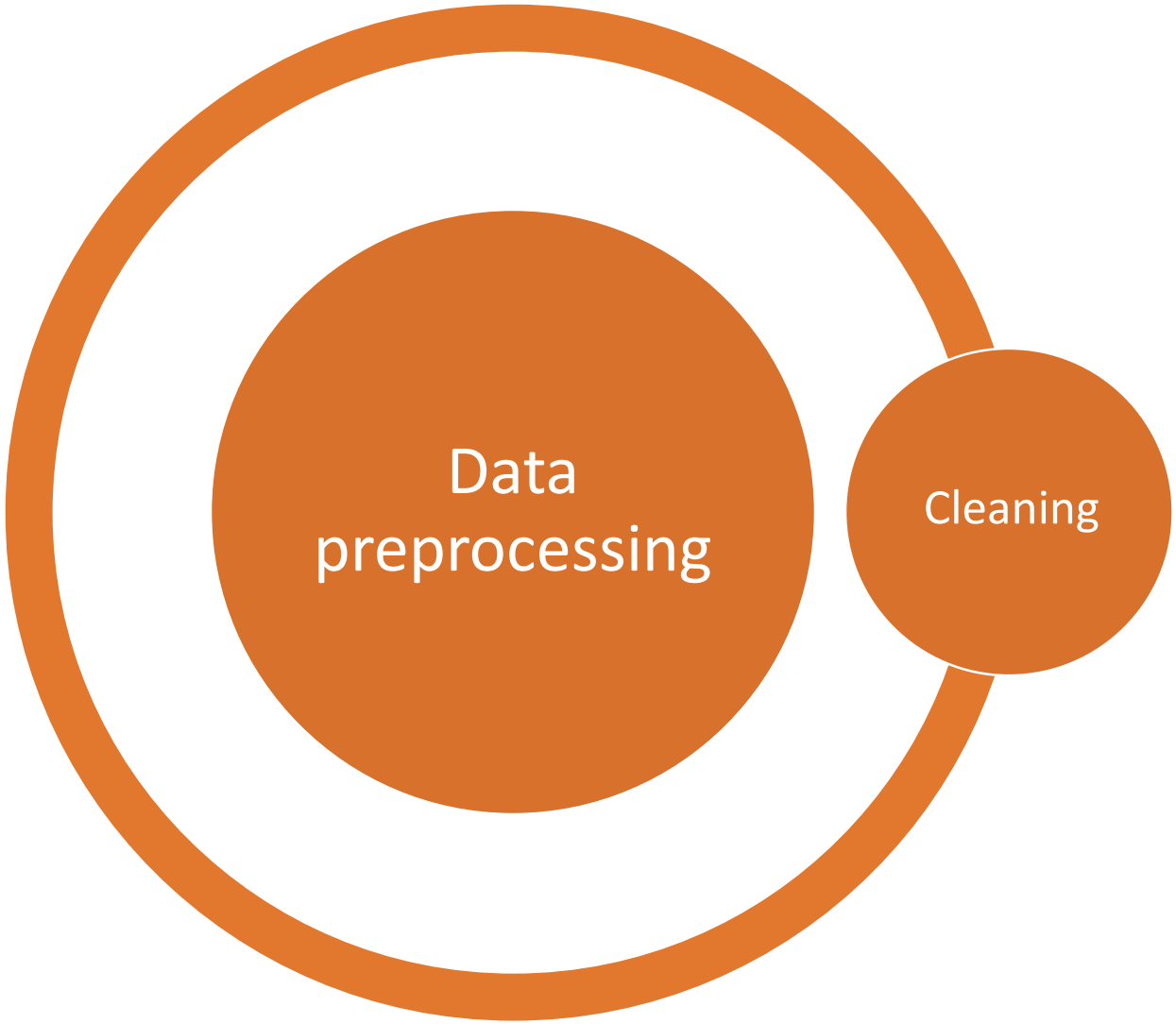


INSTITUTE OF SCIENTIFIC INSTRUMENTS
The Czech Academy of Sciences

Data Preprocessing for DL

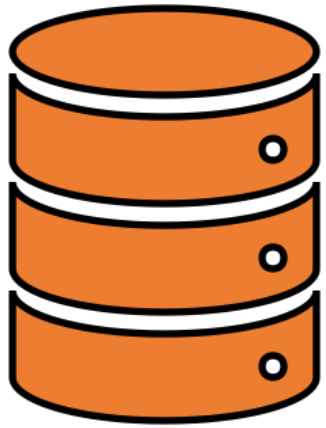
Ing. Ján Pavlus

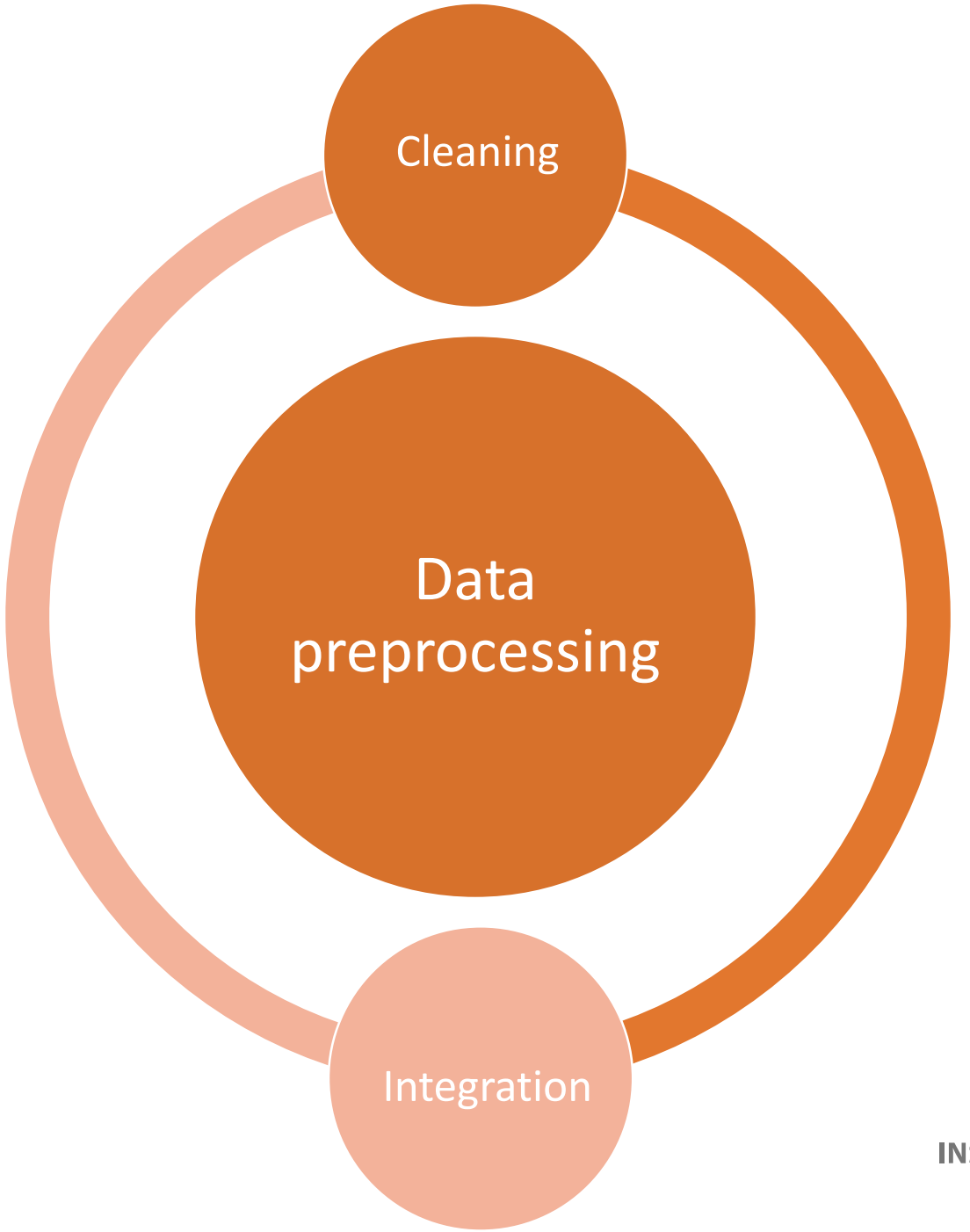

Data preprocessing
Cleaning

Cleaning

Data cleaning

- Missing values
- Noisy data
- Removing outliers
- Duplicities

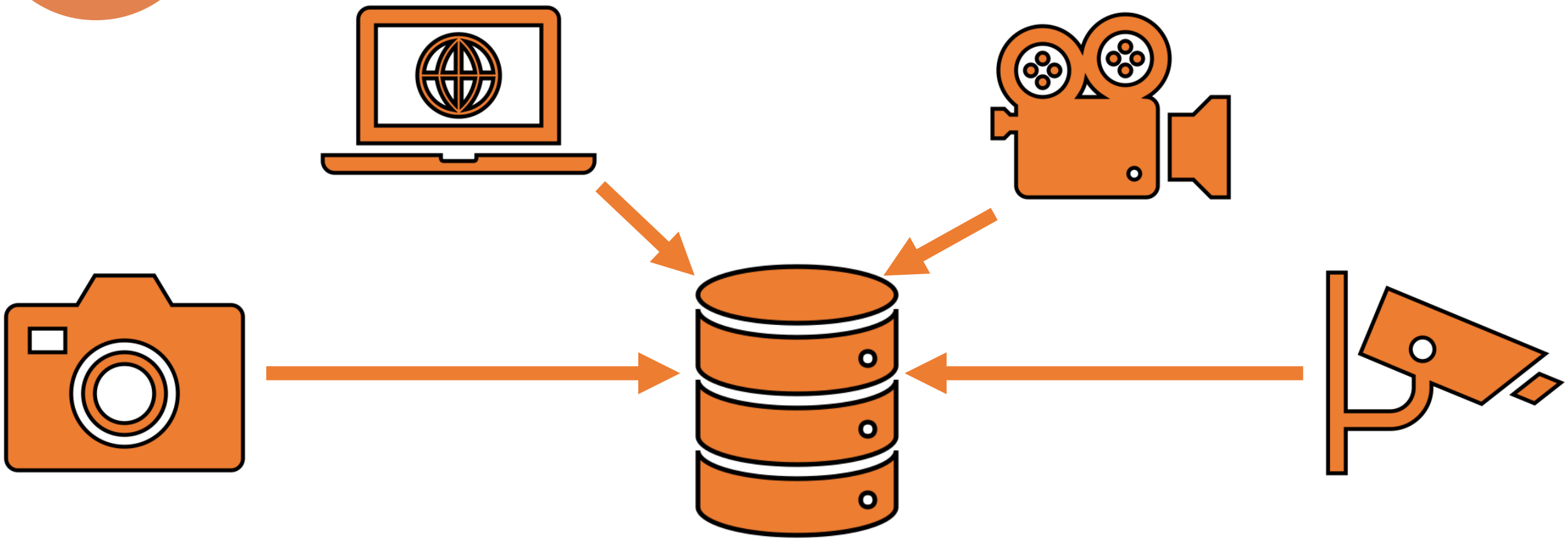


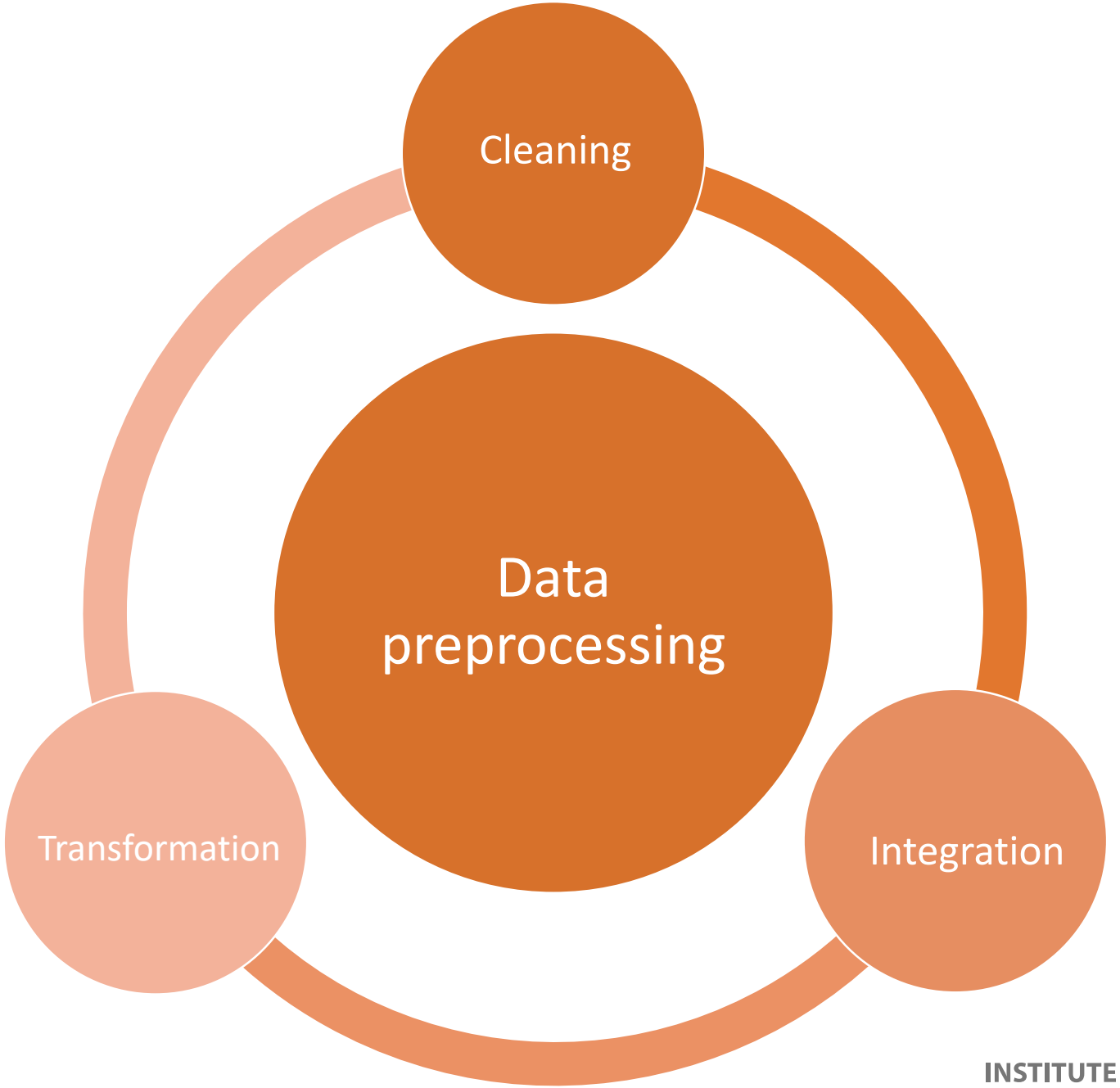


INSTITUTE OF SCIENTIFIC INSTRUMENTS
The Czech Academy of Sciences

Integration

Data integration



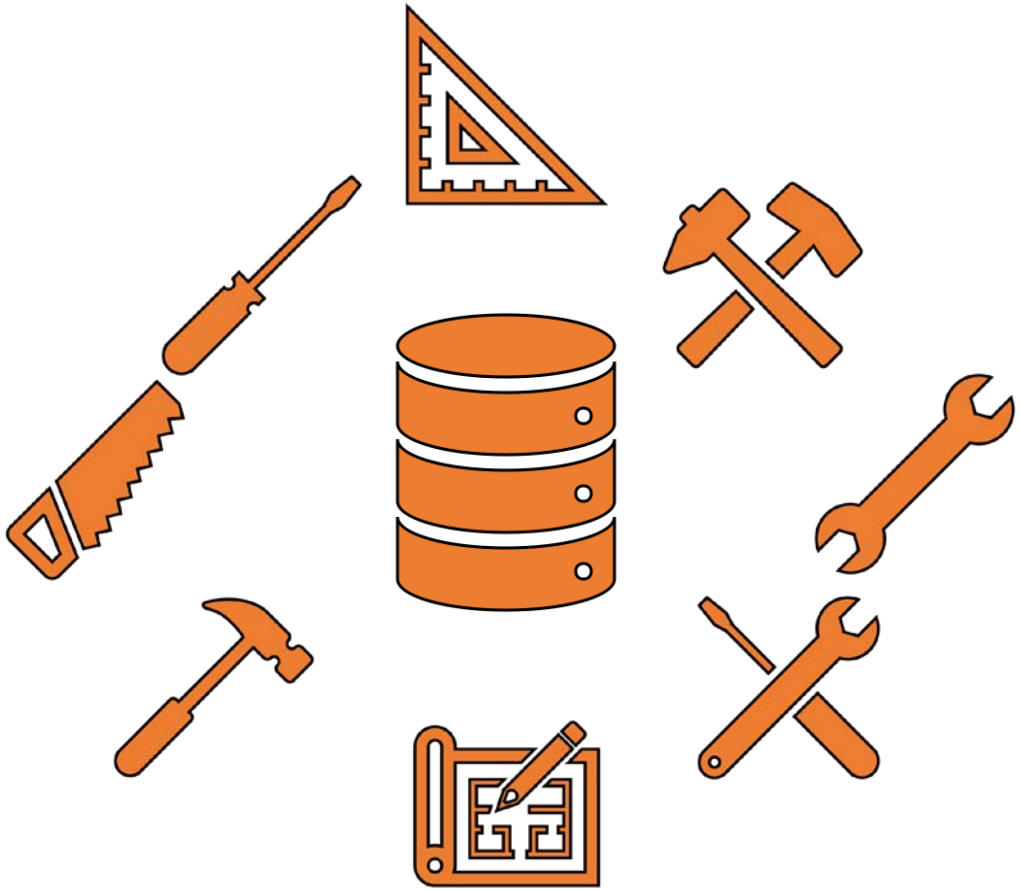


INSTITUTE OF SCIENTIFIC INSTRUMENTS
The Czech Academy of Sciences

Transformation

Data transformation

- Normalization
- Attribute selection
- Augmentations

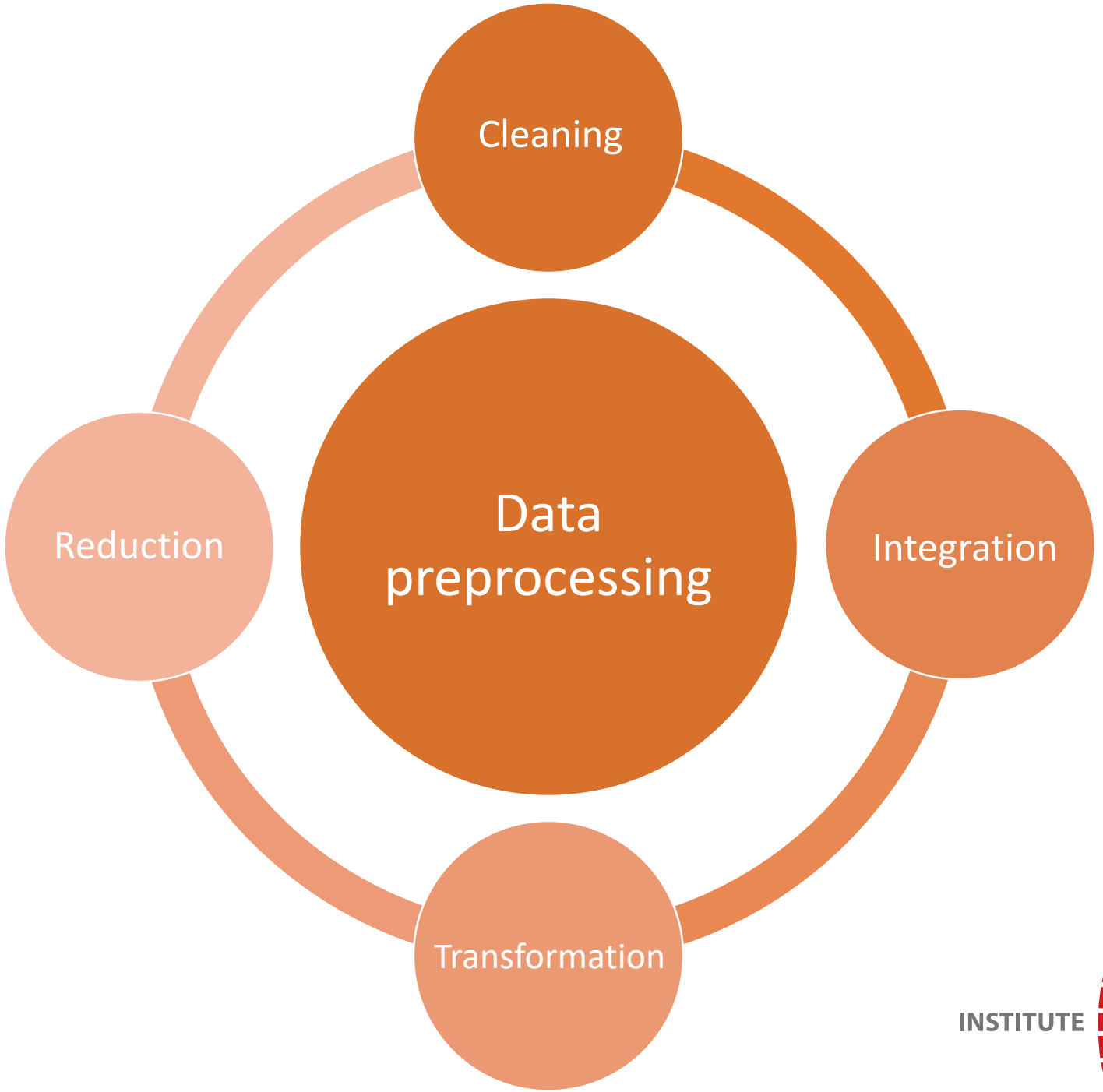


Transformation

Data transformation - Augmentations

- Extend dataset
- Extend variability
- Robust system
- Prevent from overfitting

INSTITUTE OF SCIENTIFIC INSTRUMENTS
The Czech Academy of Sciences

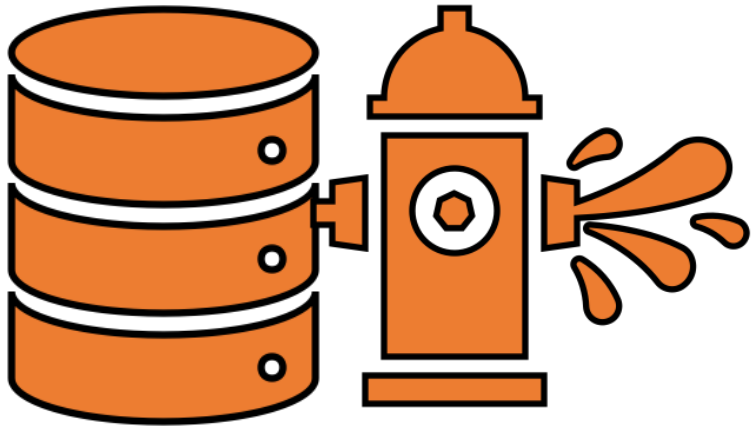


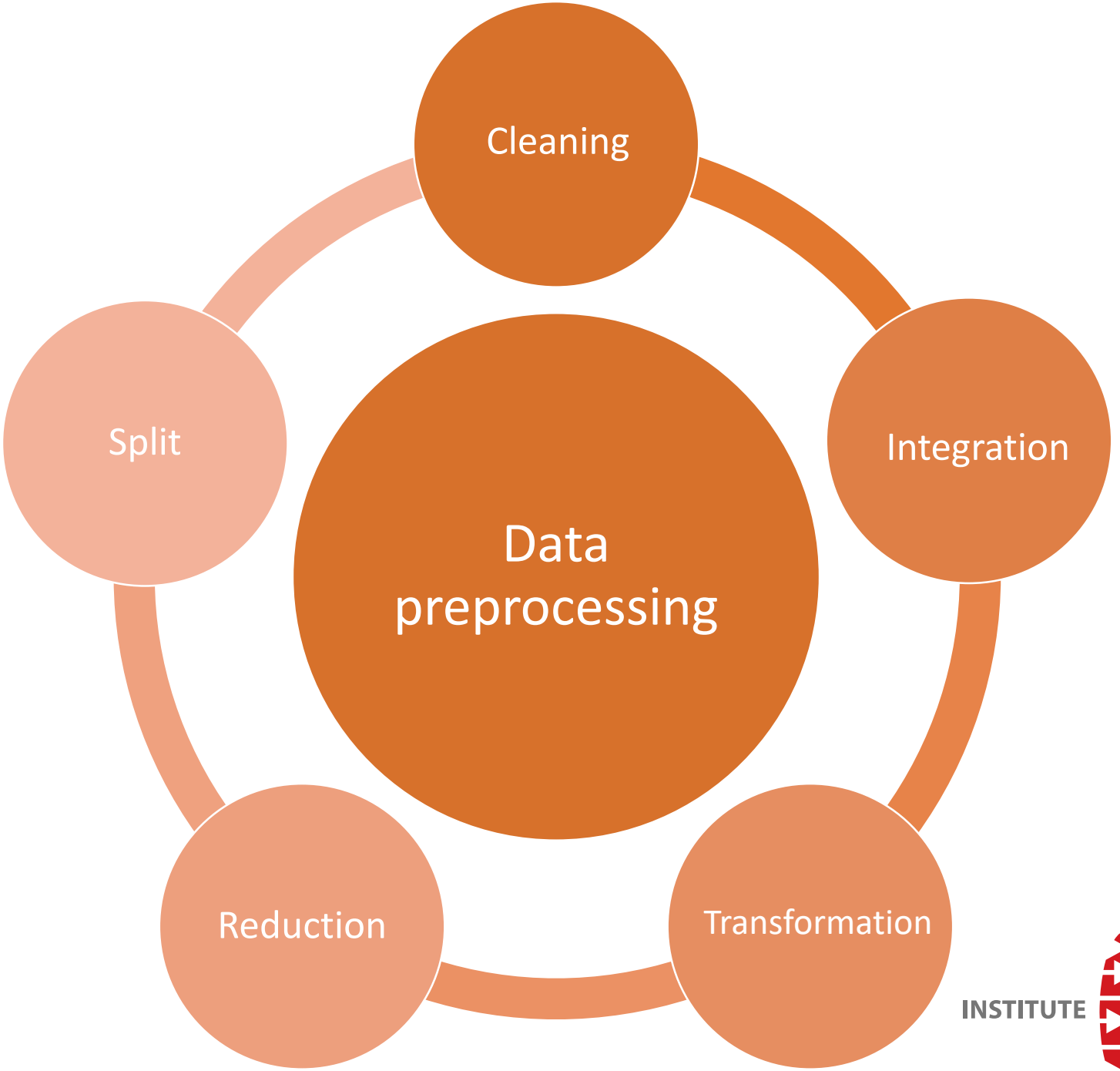
INSTITUTE OF SCIENTIFIC INSTRUMENTS
The Czech Academy of Sciences

Reduction

Data reduction

- Compress data
- Choose right representation
- Remove unnecessary attributes

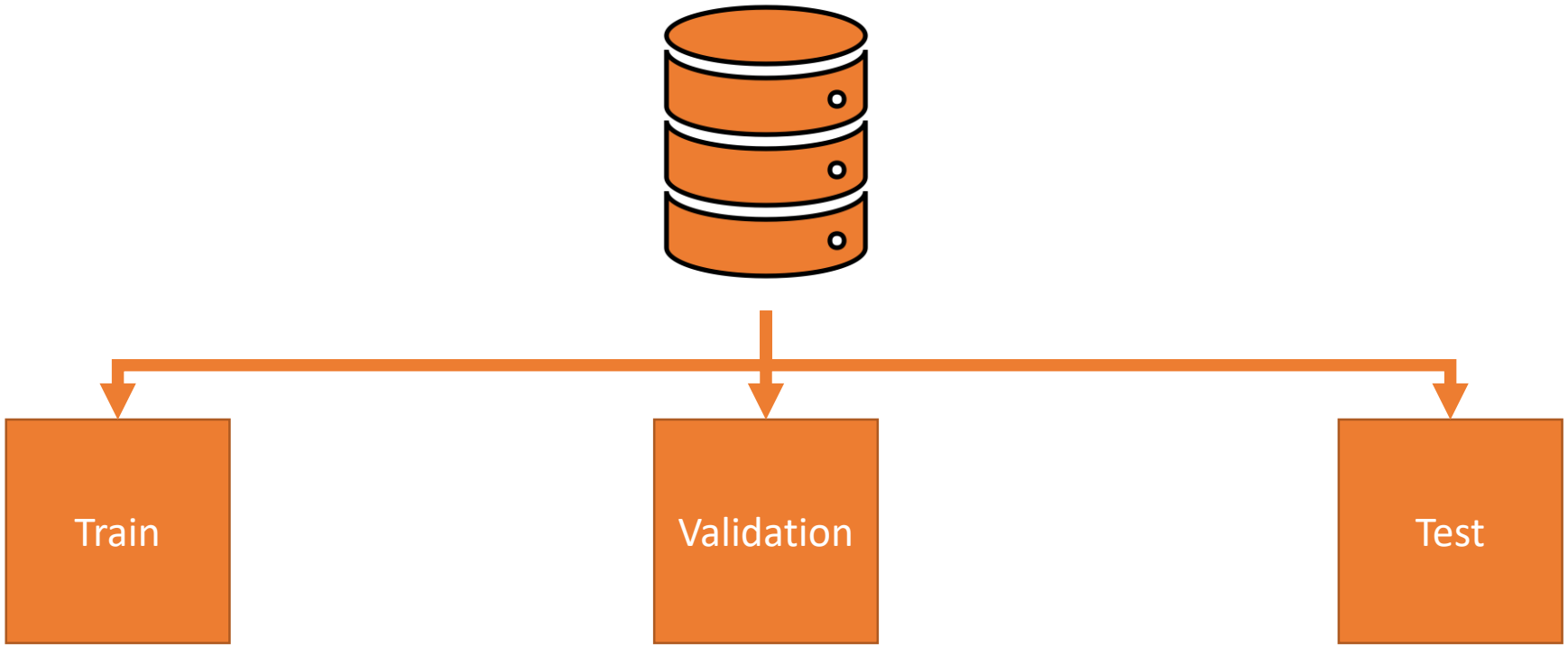




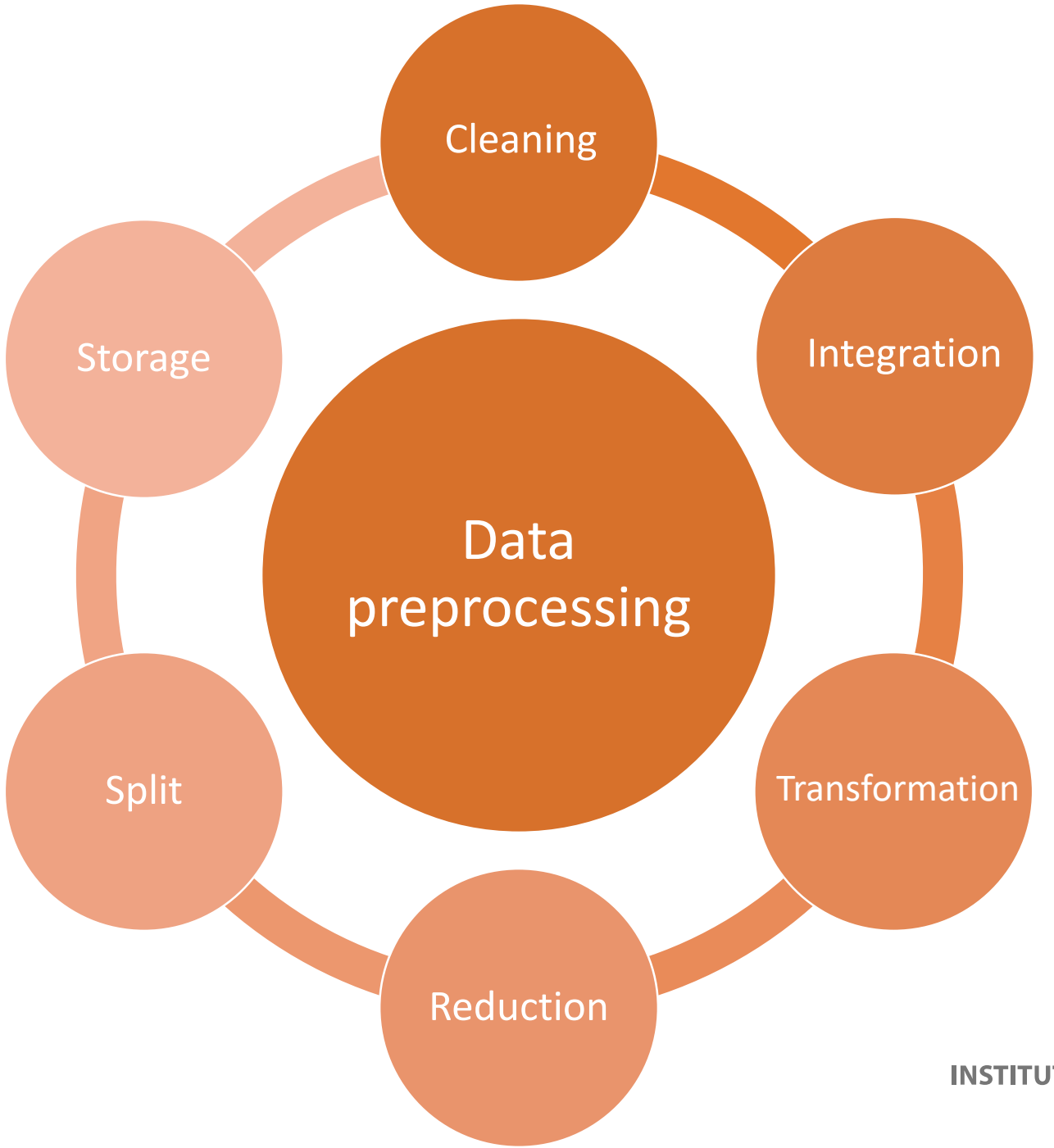
INSTITUTE OF SCIENTIFIC INSTRUMENTS
The Czech Academy of Sciences

Storage

Data split



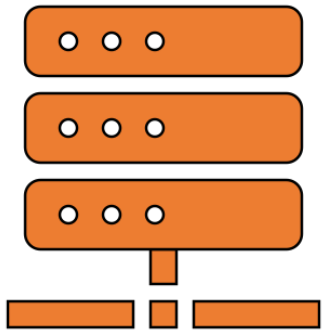
INSTITUTE OF SCIENTIFIC INSTRUMENTS
The Czech Academy of Sciences



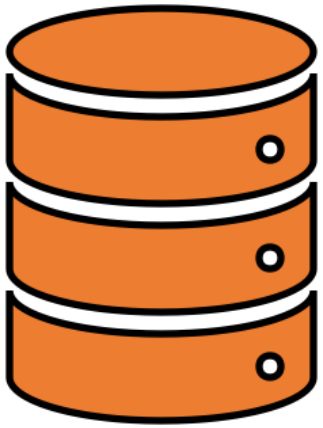
INSTITUTE OF SCIENTIFIC INSTRUMENTS
The Czech Academy of Sciences

Storage

Data storage

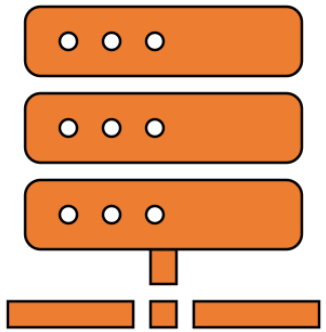


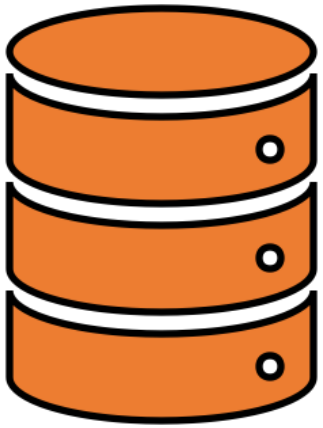
HDD



Storage

Data storage



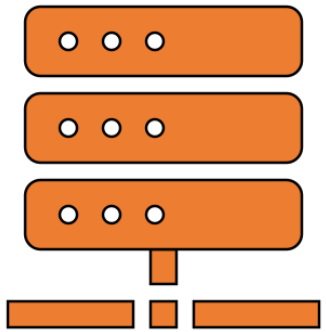


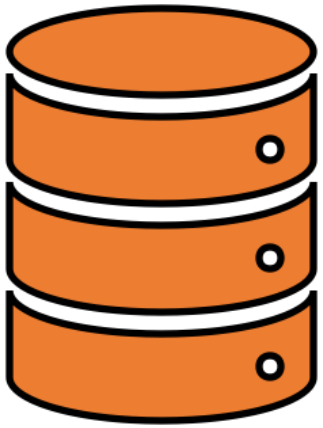
HDD

+ Cheap

Storage

Data storage



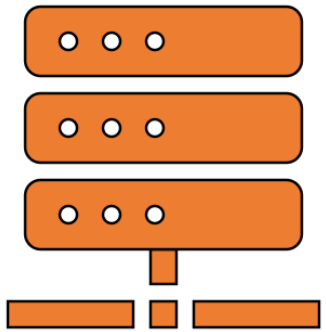


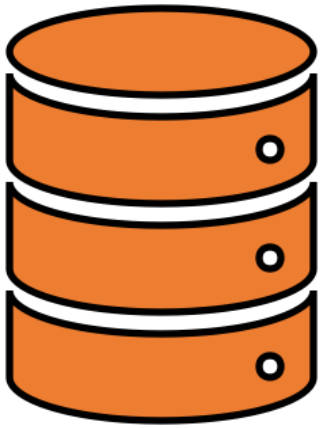
HDD

+ Cheap
+ Store big data

Storage

Data storage



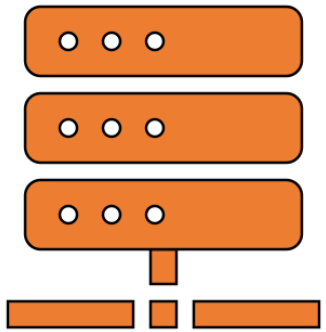


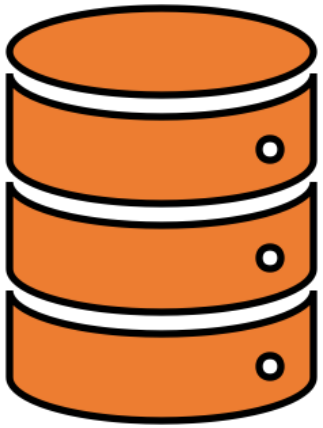
HDD

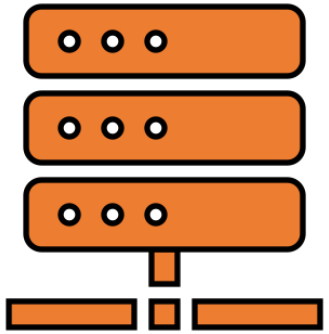
+ Cheap
+ Store big data
- Slow

Storage

Data storage







HDD

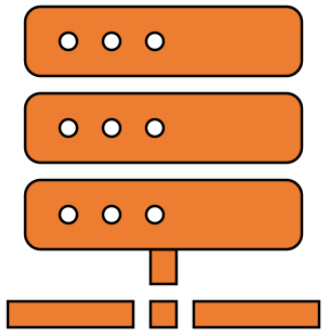
- \+ Cheap
- \+ Store big data
- \- Slow

SSD

INSTITUTE OF SCIENTIFIC INSTRUMENTS
The Czech Academy of Sciences

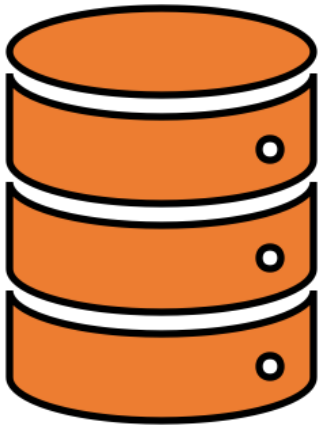
Storage

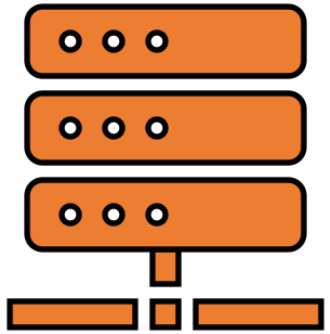
Data storage



HDD

+ Cheap
+ Store big data
- Slow





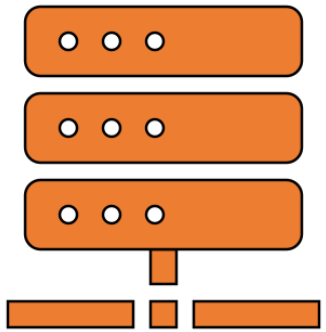
SSD

+ Really fast

INSTITUTE OF SCIENTIFIC INSTRUMENTS
The Czech Academy of Sciences

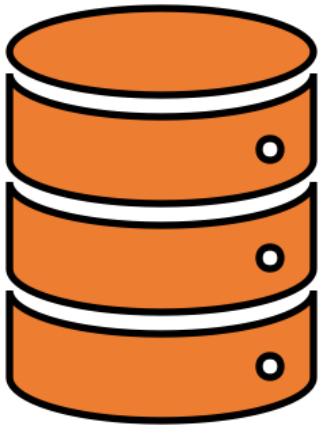
Storage

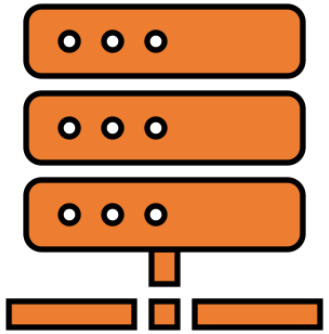
Data storage



HDD

+ Cheap
+ Store big data
- Slow





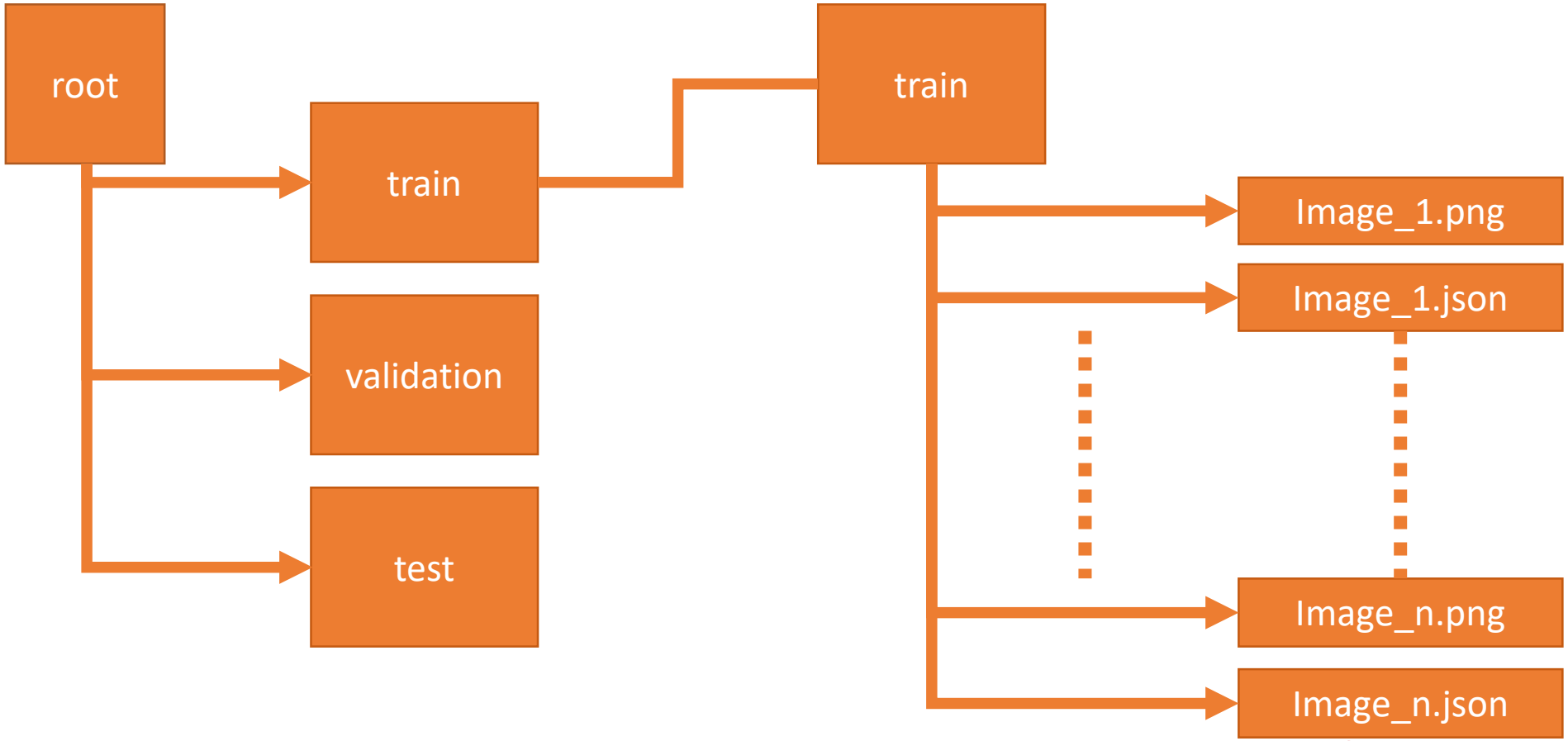
SSD

+ Really fast
- Expensive

INSTITUTE OF SCIENTIFIC INSTRUMENTS
The Czech Academy of Sciences

Storage

Data storage – folder structure



Dataloader

- Load data

Dataloader

- Load data
- Multithread

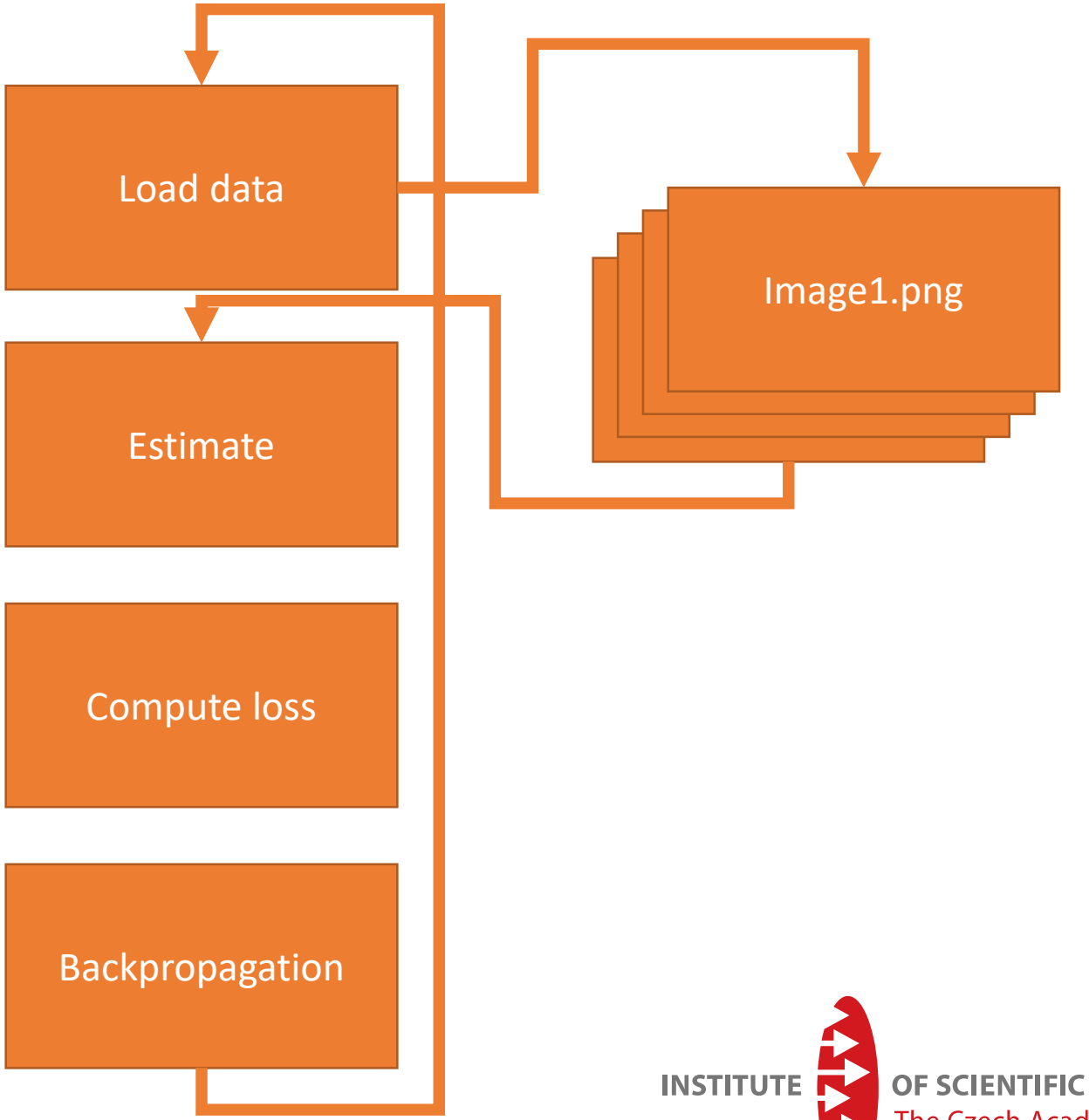
Dataloader

- Load data
- Multithread
- Provides data for the neural network

INSTITUTE OF SCIENTIFIC INSTRUMENTS
The Czech Academy of Sciences

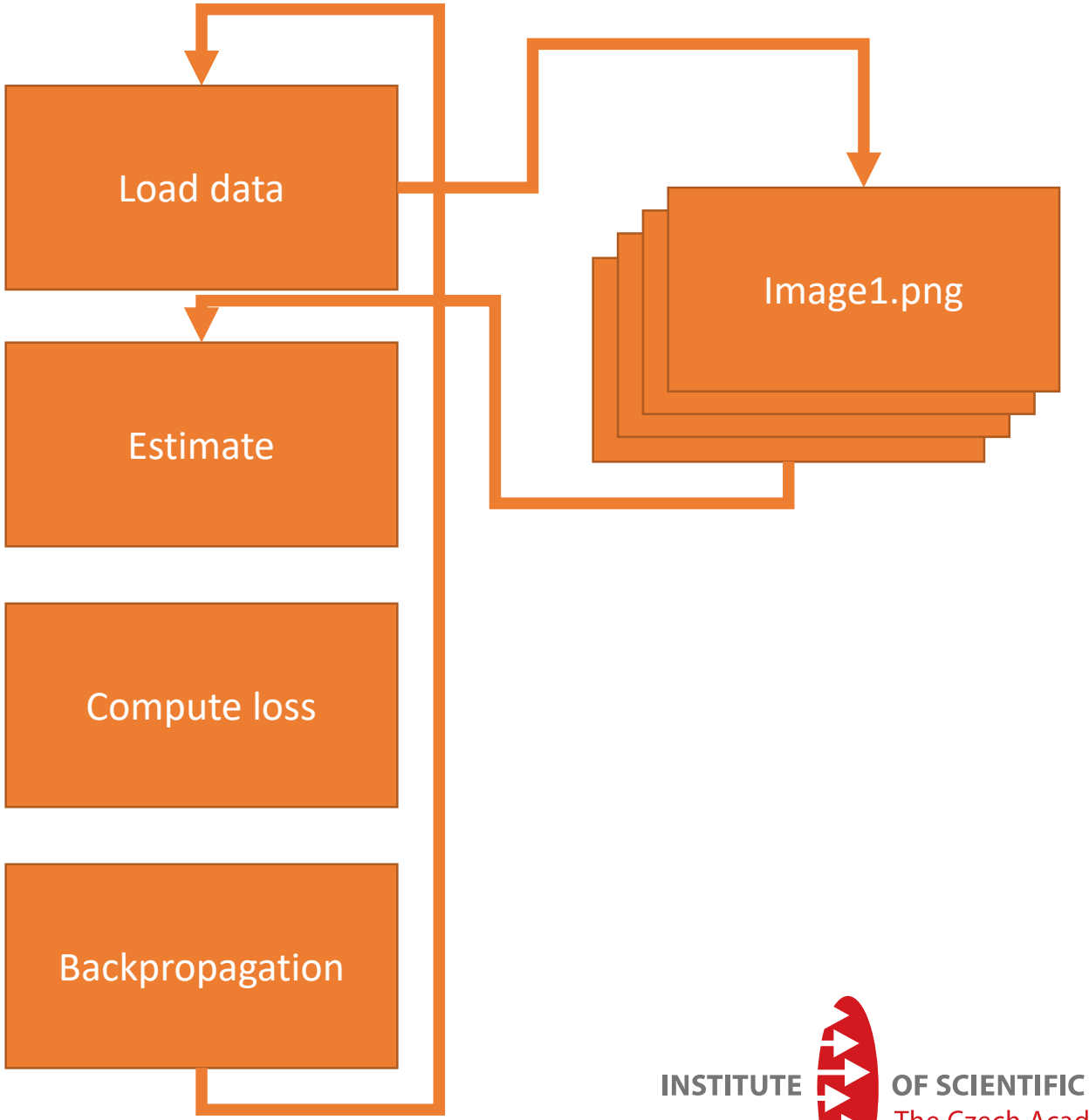
Dataloader

- Load data
- Multithread
- Provides data for the neural network
- Create batches



Dataloader

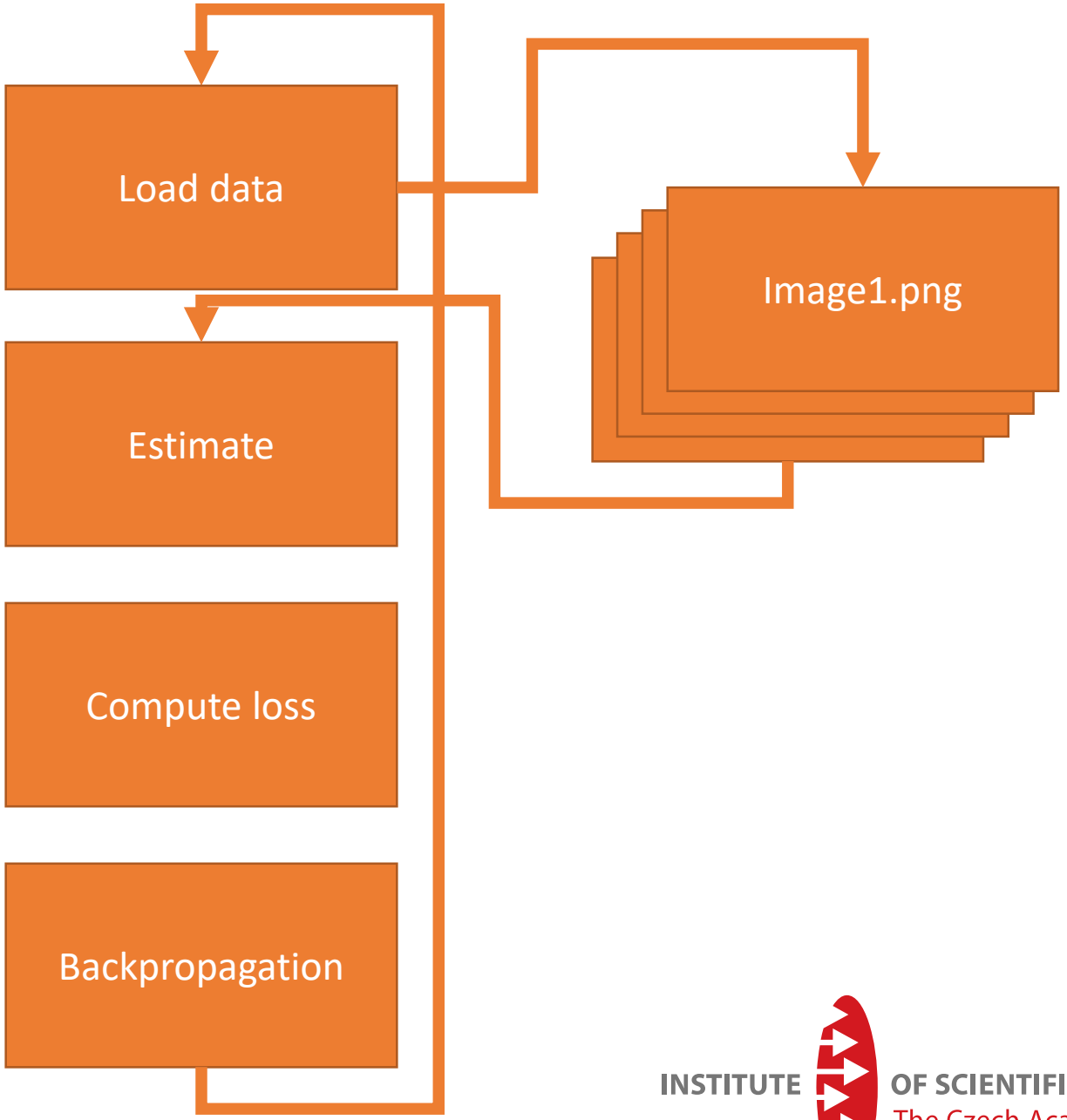
- Load data
- Multithread
- Provides data for the neural network
- Create batches
- Chunk data



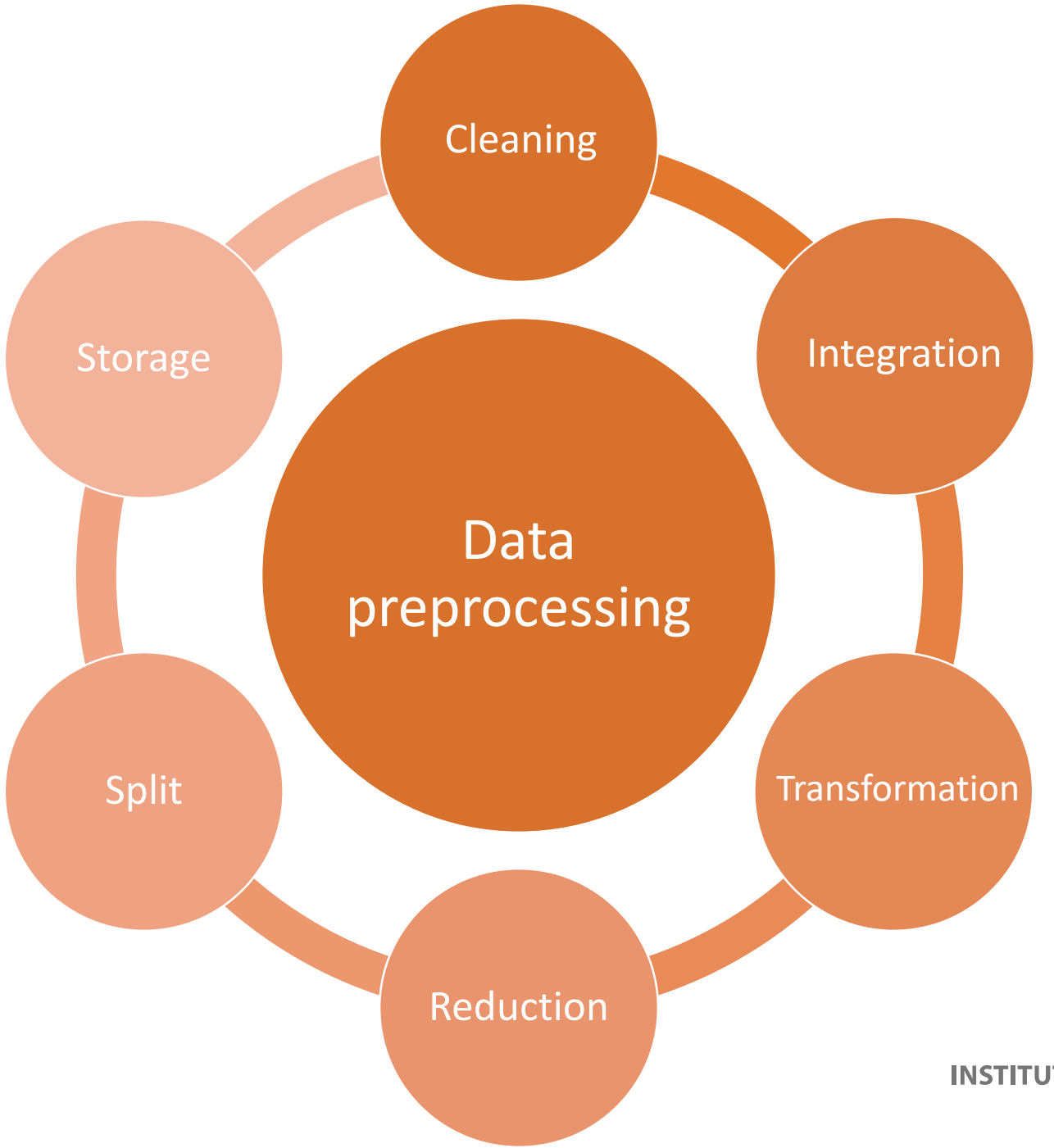
INSTITUTE OF SCIENTIFIC INSTRUMENTS
The Czech Academy of Sciences

Dataloader

- Load data
- Multithread
- Provides data for the neural network
- Create batches
- Chunk data
- Make online augmentation



INSTITUTE OF SCIENTIFIC INSTRUMENTS
The Czech Academy of Sciences



INSTITUTE OF SCIENTIFIC INSTRUMENTS
The Czech Academy of Sciences