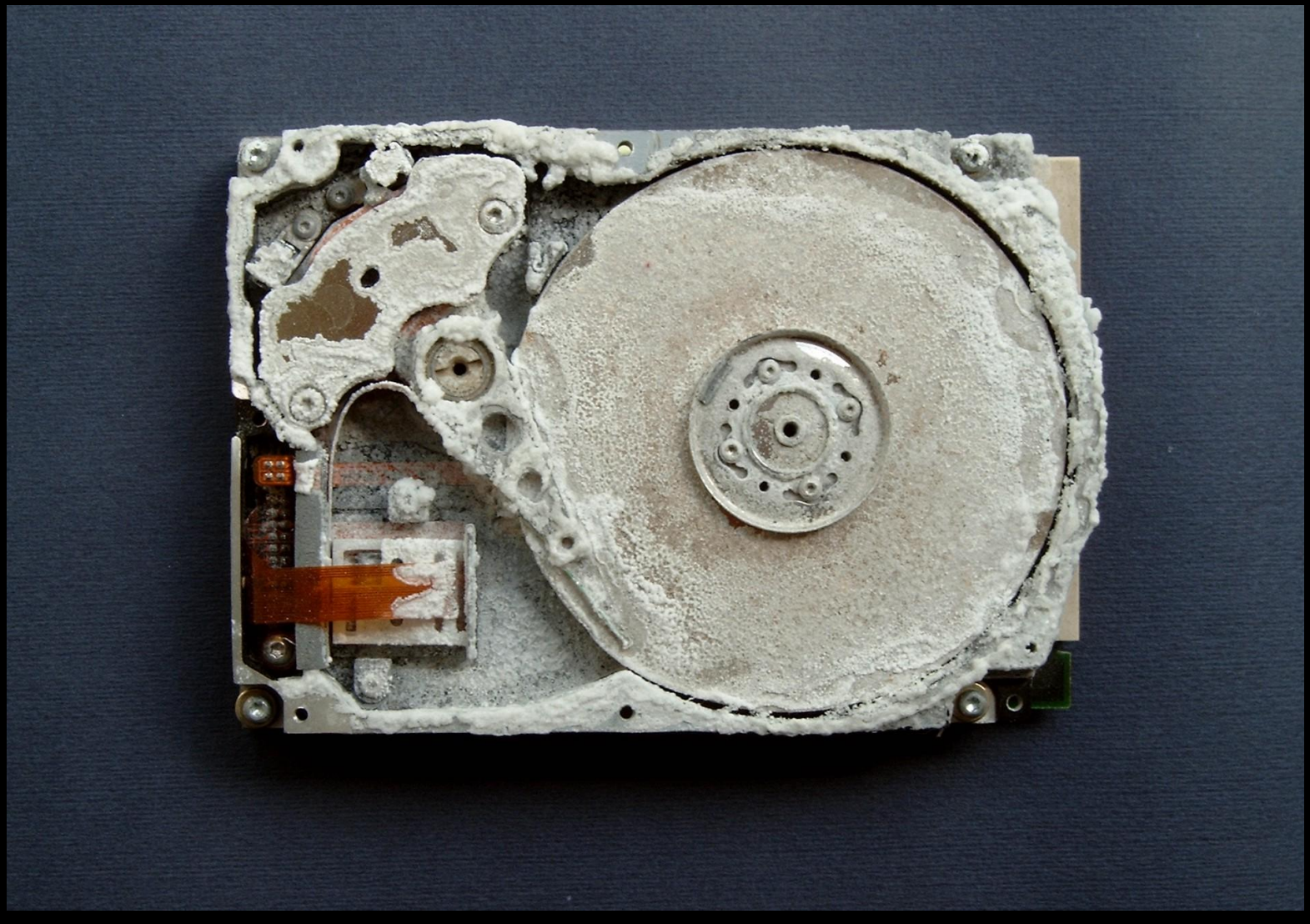
6 Tips to clean your data



Highly simplified the data science life cycle consists of 5 steps.

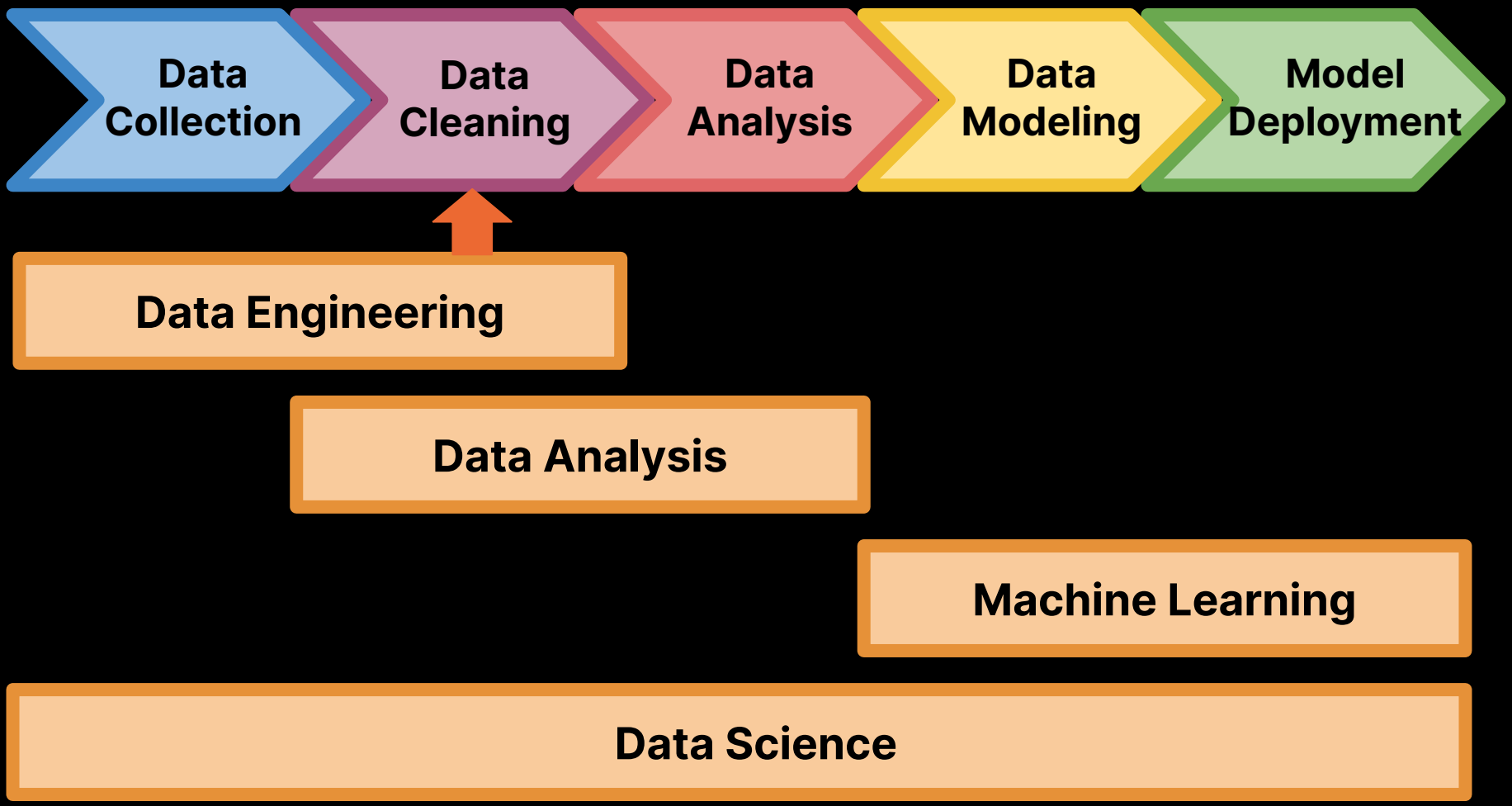
Data Engineering

Data Analysis

Machine Learning

Data Science

I want to focus on the most time consuming and often neglected step: data cleaning.



Data cleaning is a crucial part of data engineering and analysis...

Data Engineering

Data Analysis

Machine Learning

Data Science

... and is the basis for all future modeling tasks. Dirty data will result in dirty models.

Data Engineering

Data Analysis

Machine Learning

Data Science

Here are 6 tips for cleaner data.

Tip 1:
Remove
irrelevant data

When combining data sources, there are countless potentials to create data duplicates.

Identifying and removing duplicates is one of the most important steps in data cleaning.

Irrelevant data is everything that does not fit in the specific problem you are analyzing.

Irrelevant data will **muddle up** your model in the long run.

Imagine you want to model video content consumption of millennials. You might want to consider removing data from older generations.

This creates a more manageable and performant data set.

Tip 2:
Fix structural errors and missing data

Remove unusable observations from your data, including duplicates, NaNs, or irrelevant observations.

Tip 3: Remove legitimate outliers

Outliers are tricky, because outliers follow the data format but differ significantly in one or multiple measurements. Outliers can occur from measurement errors or errors in data conversion.

Beware: Outliers are innocent until proven guilty.
Never remove data points just because it a "big number".
Data can only be removed if you have legitimate reasons.

Tip 4:
Use consistent format

What happens if you add 5 and “5”?
In JavaScript you will get “55” in Python a TypeError.

My point is: use an appropriate and consistent data format.

Tip 5:
Verify your units

To which degree is your data following the same units? Is the weight given in pounds or kilograms? Are you using a consistent currency in your price model?

Convert data to a single measure unit.

Tip 6:
Normalize your data

Statistical and ML models expect data within a certain distribution.
Most commonly the data is expected to lie in the [-1,1] or the [0,1] intervall. Normalizing the data leads to better model performance.

Remember

1. Remove irrelevant data
2. Fix structural errors and missing data
3. Remove legitimate outliers
4. Use consistent format
5. Verify your units
6. Normalize your data

Feel free to reach out or to connect with me for more weekly slideshows on visualization, data science and machine learning.