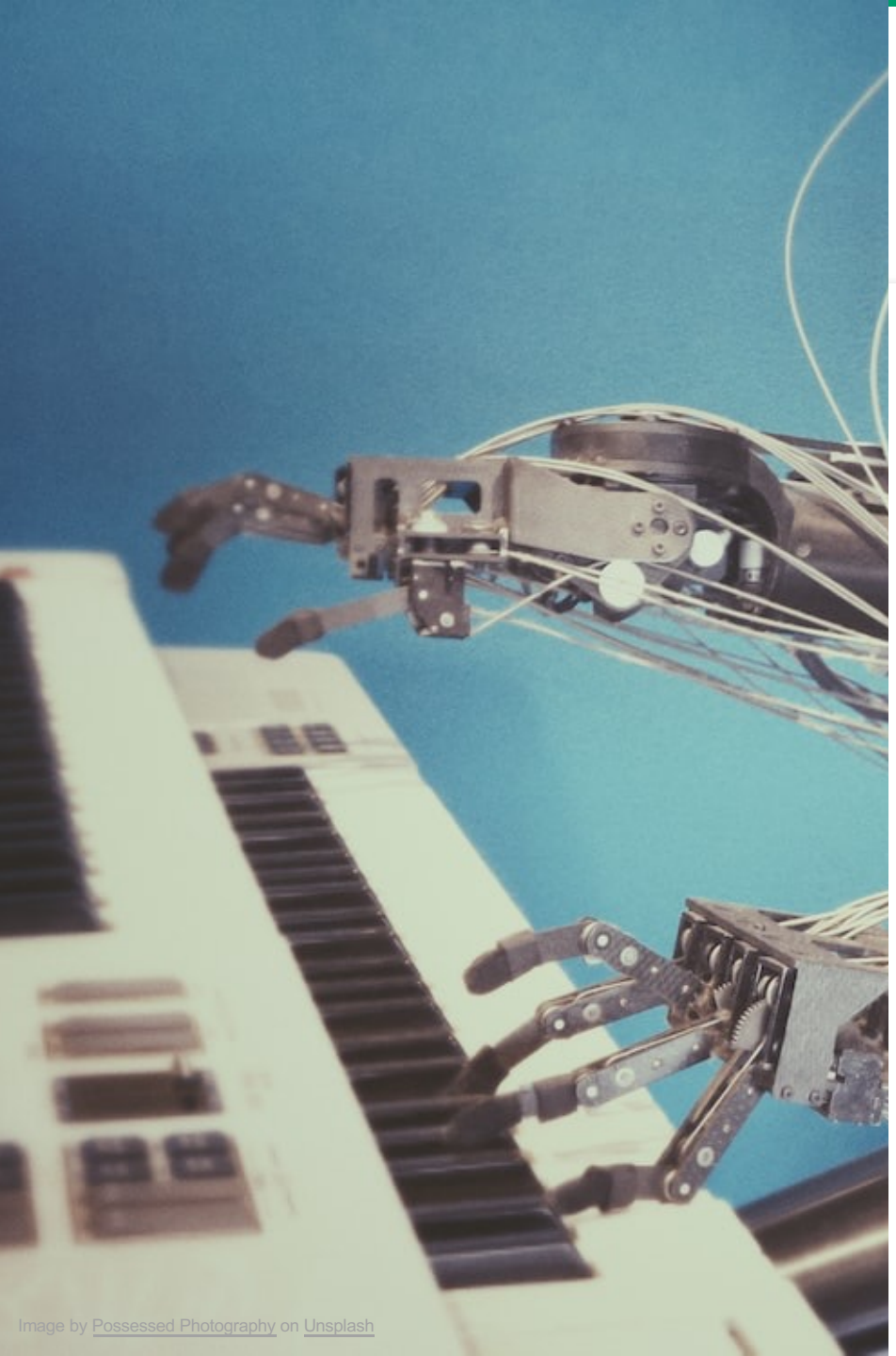


Applied Artificial Intelligence
02 - AI Lifecycle: Initiation

Univ.-Prof. Dr-Ing. habil. Niklas Kühl
www.niklas.xyz

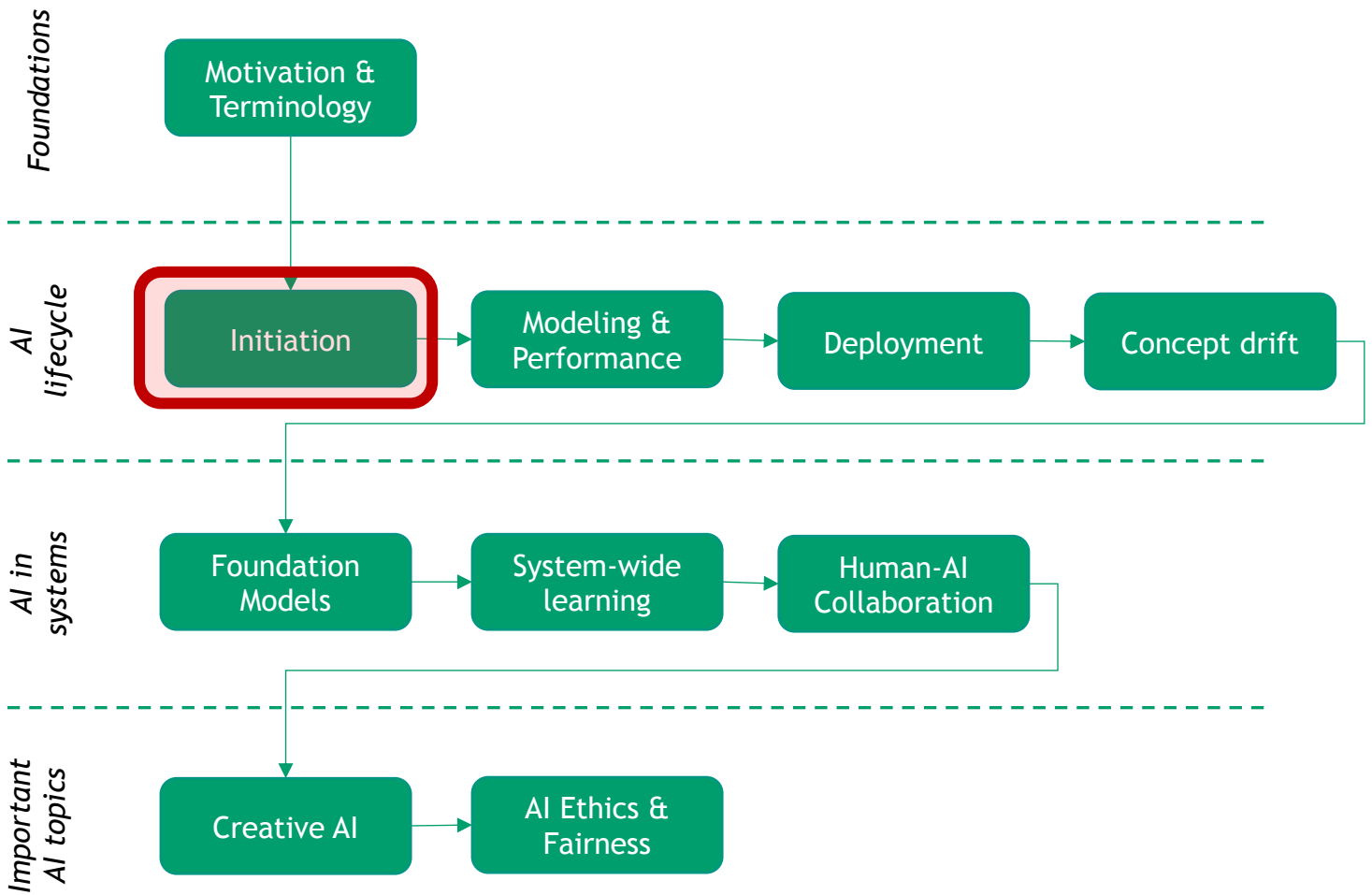
University of Bayreuth

Karlsruhe Institute of Technology

TUM School of Management

www.uni-bayreuth.de | www.kit.edu | www.tum.de | www.fim-rc.de | www.wirtschaftsinformatik.fraunhofer.de

Organizational
The story of the lecture



Objectives
What are the learning goals of this lecture?

EXPLORE

discover the importance of defining the related business problem

UNDERSTAND

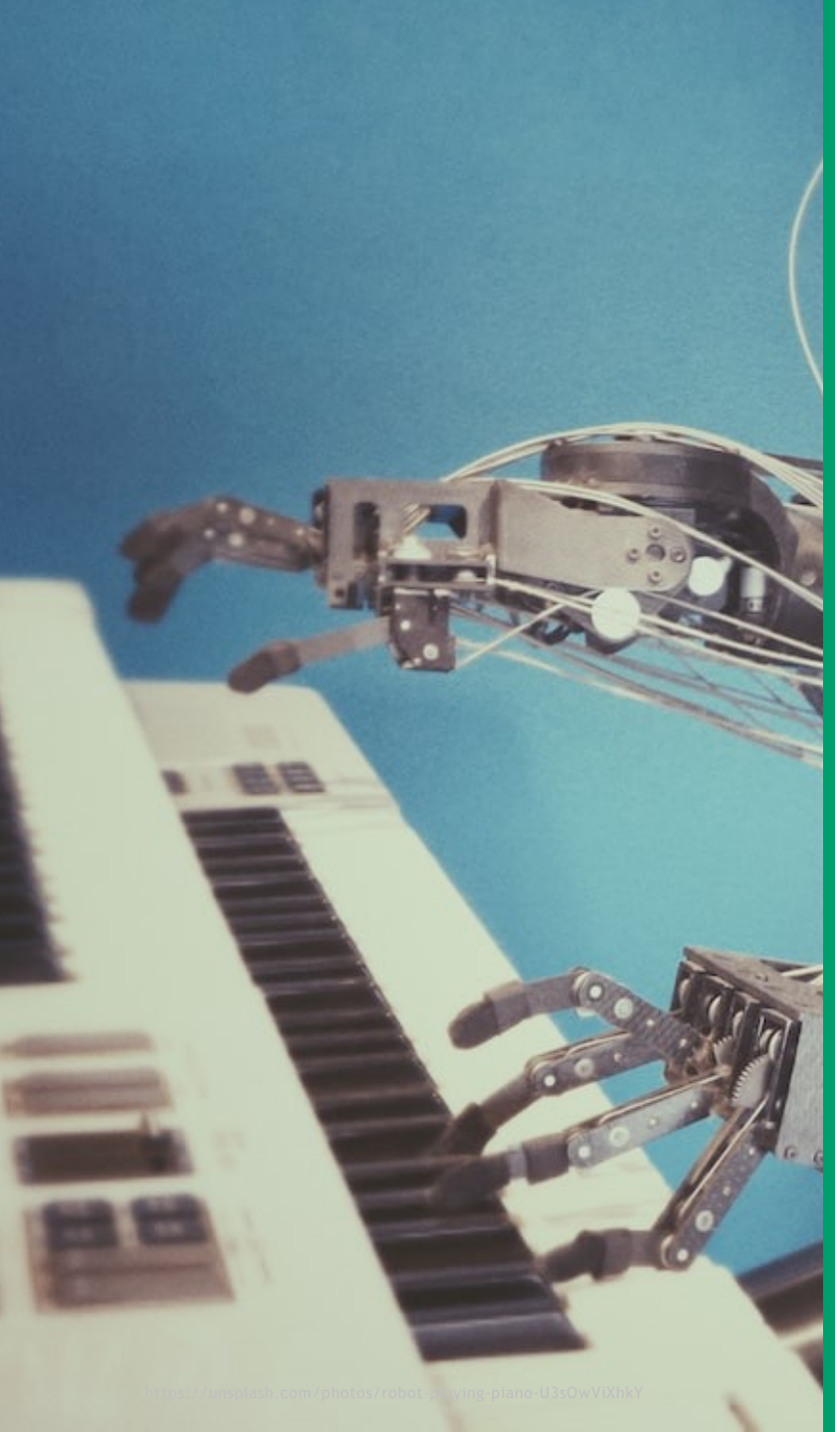
understand the options of preprocessing & feature engineering

INTENSIFY

deepen your knowledge about various types of data

APPLY

be able to get insights of gathering data



0 Introduction

A Problem Statement

B Data Gathering

C Sampling

D Data Distribution

E Data Quality

F Data Pre-processing

G Feature Engineering

Introduction
What is the model initiation?

Select relevant data from data sources

[1]

https://picryl.com/media/cat-grumpy-mood-animals-9d92b5 [1]

Introduction
What is the model initiation?

Select relevant data from data sources

[1]

Prepare data to be digestible by algorithms

[2]

Goal: Solve the initial business problem.

https://picryl.com/media/cat-grumpy-mood-animals-9d92b5 [1]
https://www.pinterest.de/pin/289426713550387189/ [2]

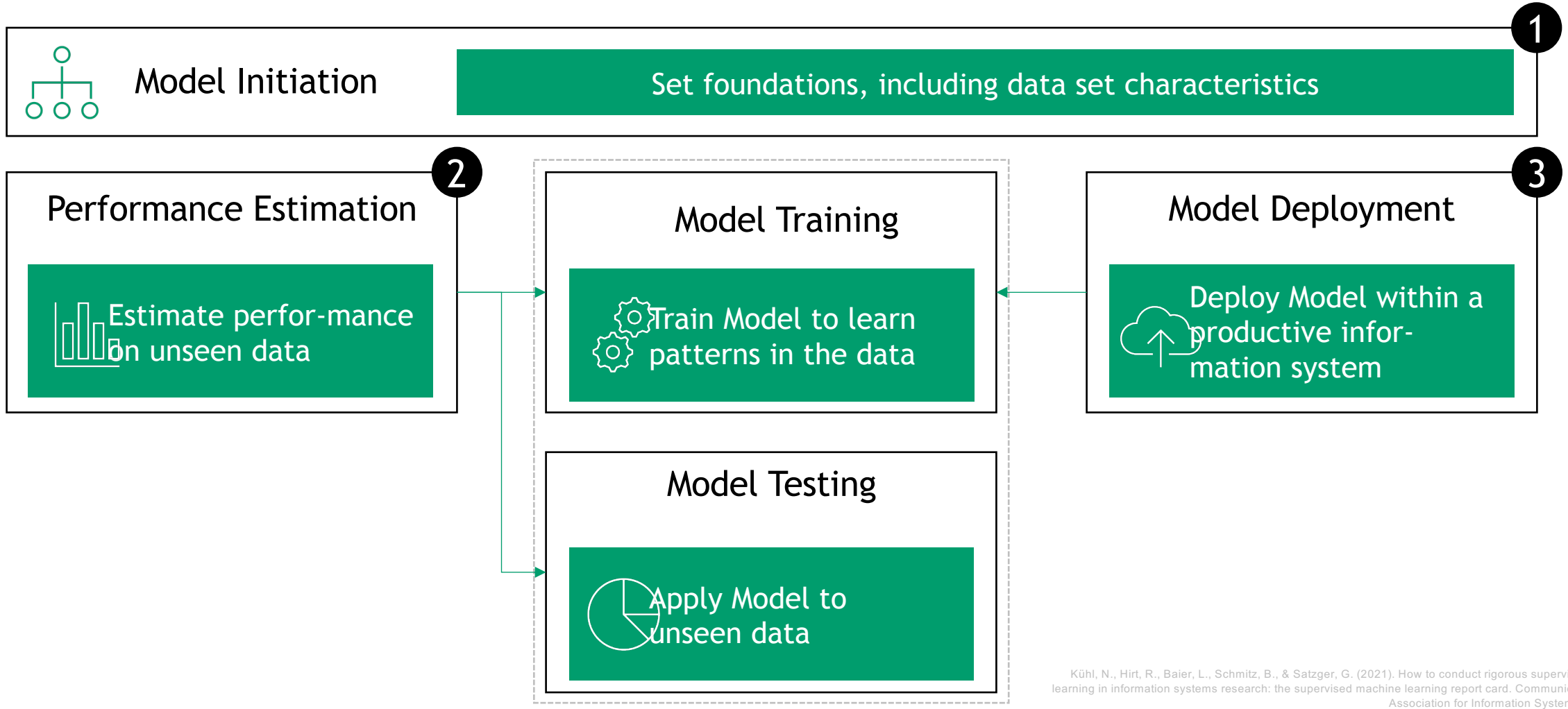
Introduction
What for?

	Input			Target
ID	Attended lectures	Submitted exercises	(…)	Exam passed
1	12	6		YES
2	2	1		NO
3	13	4		YES
4				
…				

Model learns relationships from input to a target variable on historical data….

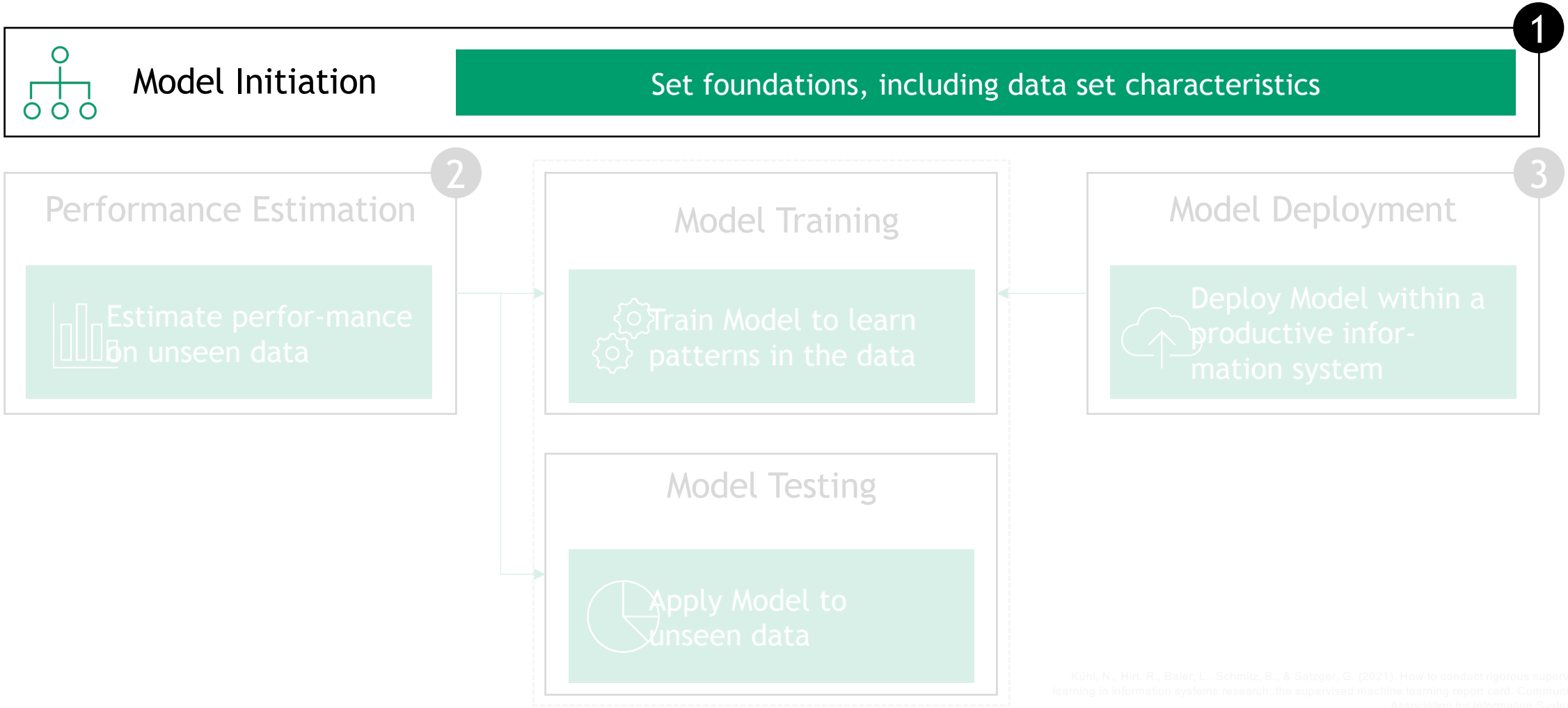
…and then applies this knowledge to instances with unknown target variables.

The AI Lifecycle
Required steps for supervised machine learning.



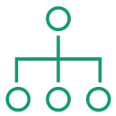
Kühl, N., Hirt, R., Baier, L., Schmitz, B., & Satzger, G. (2021). How to conduct rigorous supervised machine learning in information systems research: the supervised machine learning report card. Communications of the Association for Information Systems, 48(1), 46.

The AI Lifecycle
Set foundations with the model initiation.



Kühl, N., Hirt, R., Baier, L., Schmitz, B., & Satzger, G. (2021). How to conduct rigorous supervised machine learning in information systems research: the supervised machine learning report card. Communications of the Association for Information Systems, 48(1), 46.

The AI Lifecycle
The initiation consists of various components.



Model Initiation

1

Set foundations, including data set characteristics

A **Problem Statement**
(e.g. binary classification, regression, etc.)

B **Data Gathering**
(e.g. rigor, labeling)

C **Sampling**
(e.g. stratified sampling from lager data pool)

D **Data Distribution**
(e.g. amount of classes, share, distribution)

E **Data Quality**
(e.g. sparsity, noise, complexity

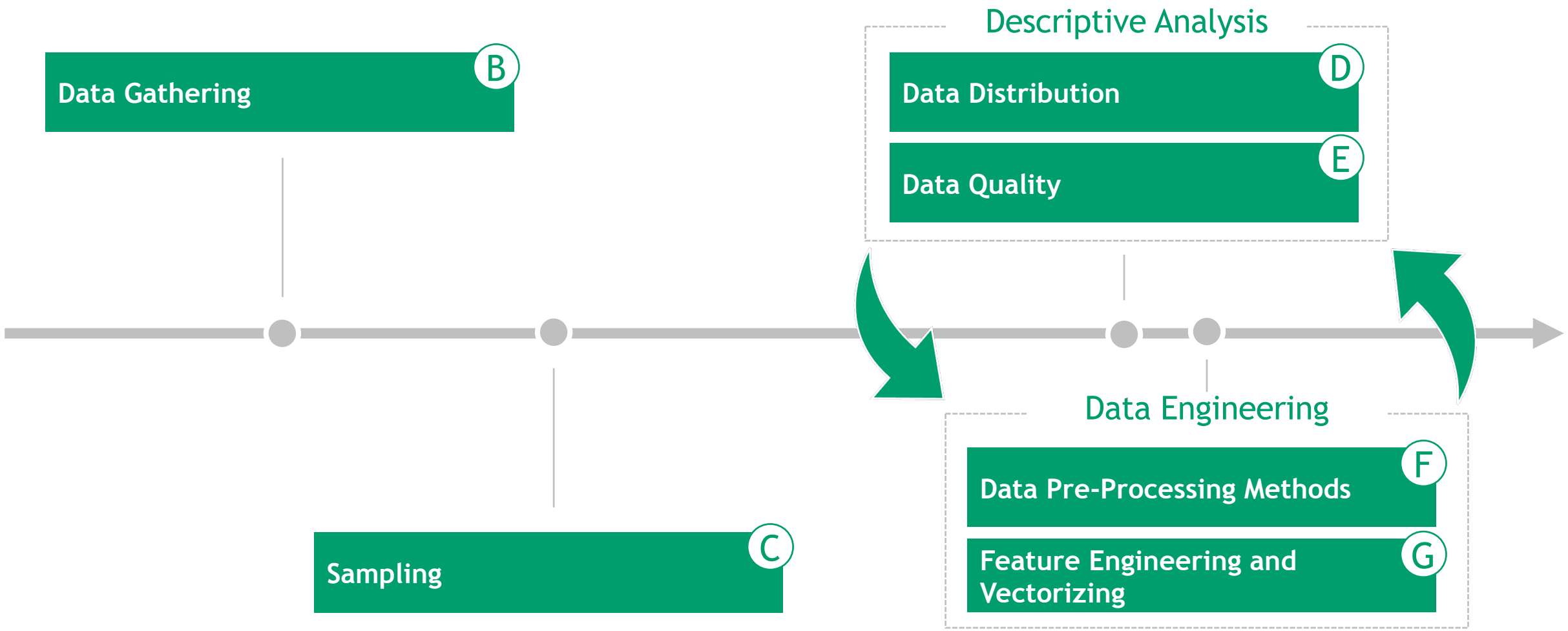
F **Data Pre-Processing Methods**
(which and why)

G **Feature Engineering and Vectorizing**
(which and why)

Kühl, N., Hirt, R., Baier, L., Schmitz, B., & Satzger, G. (2021). How to conduct rigorous supervised machine learning in information systems research: the supervised machine learning report card. Communications of the Association for Information Systems, 48(1), 46.

The AI Lifecycle

Simplified process of initiation includes iterative procedure.



AGENDA
Presentation of initiation components step-by-step.

A **Problem Statement**
(e.g. binary classification, regression, etc.)

B **Data Gathering**
(e.g. rigor, labeling)

C **Sampling**
(e.g. stratified sampling from lager data pool)

D **Data Distribution**
(e.g. amount of classes, share, distribution)

E **Data Quality**
(e.g. sparsity, noise, complexity

F **Data Pre-Processing Methods**
(which and why)

G **Feature Engineering and Vectorizing**
(which and why)

Kühl, N., Hirt, R., Baier, L., Schmitz, B., & Satzger, G. (2021). How to conduct rigorous supervised machine learning in information systems research: the supervised machine learning report card. Communications of the Association for Information Systems, 48(1), 46.

0 Introduction

A Problem Statement

B Data Gathering

C Sampling

D Data Distribution

E Data Quality

F Data Pre-processing

G Feature Engineering

Problem Statement
What we consider a well-defined problem.

Objectives – describe the primary objective as well as other related questions that might be relevant. [1]

Example: the primary business goal might be to keep current customers by predicting when they are prone to move to a competitor. Related business questions might be "Does the channel used affect whether customers stay or go?" or "Will lower ATM fees significantly reduce the number of high-value customers who leave?"

Success Criteria – describe the criteria for a successful outcome to the project. This could be specific and measurable. [1]

Example: reduction of customer churn to a certain level, or it might be general and subjective, such as "give useful insights into the relationships." In the latter case, it needs to be clear who it is that makes the subjective judgment.

Chapman et al. (2000), Step-by-Step Data Mining Guide [1]

Problem Statement

Machine learning algorithms need measurable goals.

Business Problem

Objectives and business **success criteria** describe the problem. However, these descriptions are often subjective and lack measurability.

Translation

Machine Learning Problem

Machine learning algorithms on the other hand require a **clear, measurable goal** that can be quantified by a dominate metric.

Problem Statement

Machine learning algorithms need measurable goals.

Business Problem

Objectives and business **success criteria** describe the problem. However, these descriptions are often subjective and lack measurability.

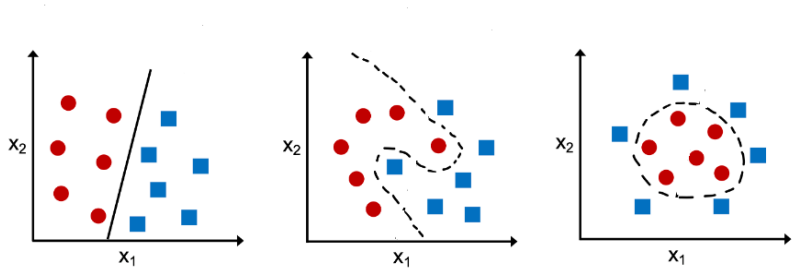
Translation

Machine Learning Problem

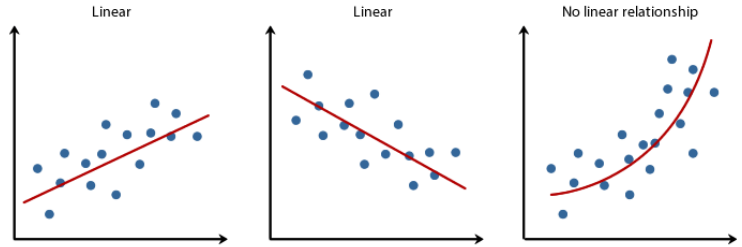
Machine learning algorithms on the other hand require a **clear, measurable goal** that can be quantified by a dominate metric.

Supervised Learning

Classification | *classify instances into classes*



Regression | *predict numerics as close as possible*

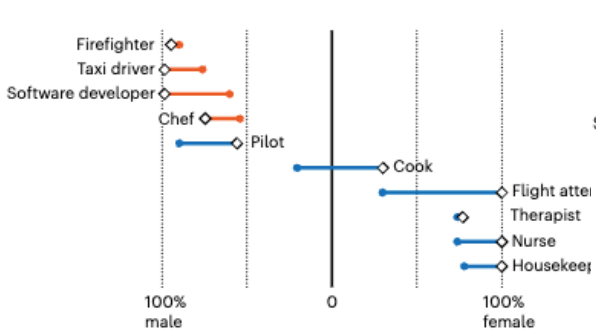


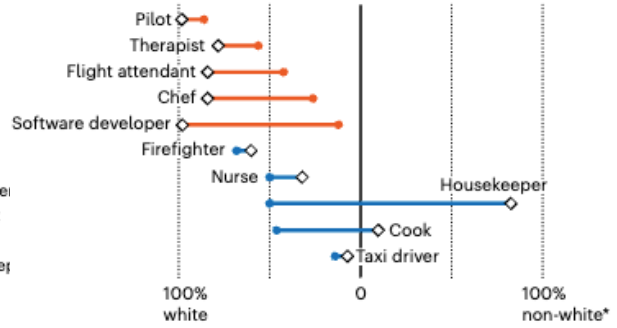
Problem Statement
Example from Research & Industry.

AI image generators often give racist and sexist results [1]

Nature News feature, 2024

Researchers have found biases: text-to-image generative AI models often produce images that include biased and stereotypical traits related to gender, skin colour, occupations, nationalities and more.



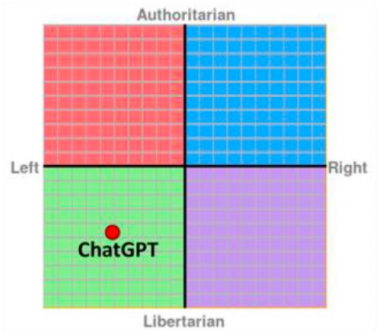


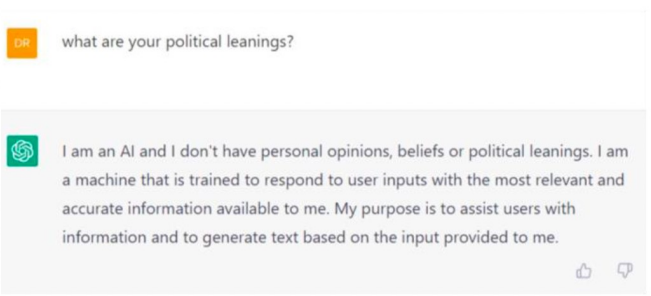
The Political Biases of ChatGPT [2]

David Rozado, 2023

New Zealand Institute of Skills and Technology

ChatGPT advances tech interaction but raises political bias concerns. Political tests reveal left-leaning inclination in ChatGPT. How to ensure factual neutrality and balance in content?





https://www.nature.com/articles/d41586-024-00674-9 [1]
James Vincent (2016), Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day, The Verge (2016/03/24) ; Pictures from Twitter [2]

https://www.vecteezy.com/photo/24366323-library-stacks-of-books-and-bookshelf-3d-illustration-generate-ai

Data Gathering
We consider various forms of data structures.

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

Structured Data

simple, standardized and efficient way

1.) Stored in relational database in a table with rows and columns
2.) Accessed by relational keys and a fixed, predefined structure

→ Example: SQL databases.

```
<University>
  <Student ID="1">
    <Name>John</Name>
    <Age>18</Age>
    <Degree>B.Sc.</Degree>
  </Student>
  <Student ID="2">
    <Name>David</Name>
    <Age>31</Age>
    <Degree>Ph.D. </Degree>
 </Student>
 ....
</University>
```

Semi-structured Data

self-describing structure with some organizational properties

1.) Data follows a pre-defined order
2.) Lacks the formal structure of a relational database.

→ Examples: CSV, XML databases.

The university has 5600 students.
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

Unstructured Data

follows no structure

1.) represents over 80% of all data
2.) hard to analyze.

→ Examples: Emails, WebPages

[1,2]

Buneman, Peter. "Semistructured data." Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. ACM, 1997. [1]
Dey, Lipika, et al. "A framework to integrate unstructured and structured data for enterprise analytics." Information Fusion (FUSION), 2013 16th International Conference on. IEEE, 2013.[2]

Data Gathering
We consider various forms of data structures.

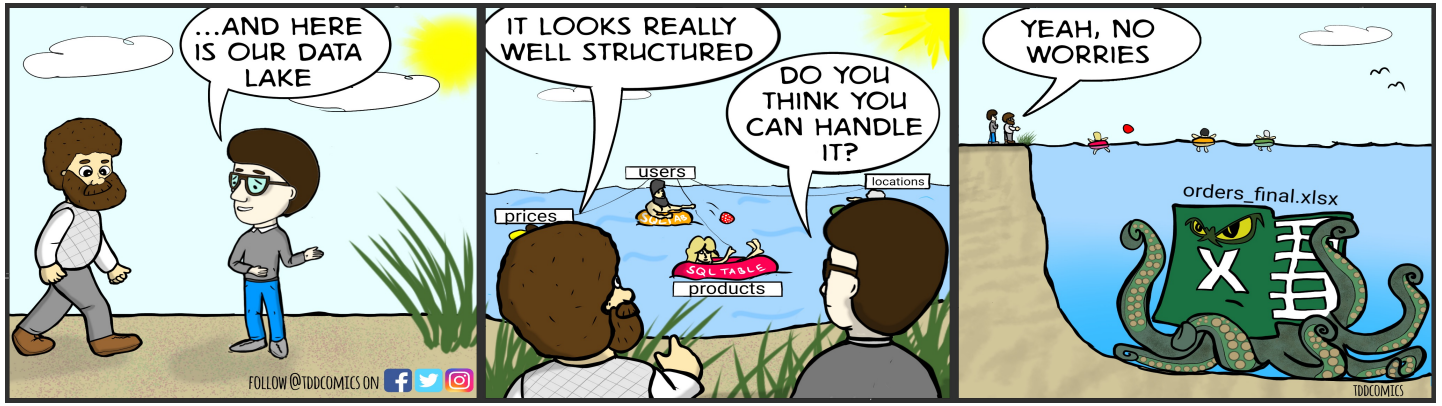
>>
- It is often necessary to combine different data sources and types of data.
- The **resulting data collection must be rigor** which means it is complete and without gaps concerning the object of interest (e.g., no missing time spans in time critical analyses)

Structured Data	Semi-structured Data	Unstructured Data
simple, standardized and efficient way	**self-describing structure with some organizational properties**	**follows no structure**
1.) Stored in relational database in a table with rows and columns 2.) Accessed by relational keys and a fixed, predefined structure	1.) Data follows a pre-defined order 2.) Lacks the formal structure of a relational database.	1.) represents over 80% of all data 2.) hard to analyze.
→ Example: SQL databases.	→ Examples: CSV, XML databases.	→ Examples: Emails, WebPages

[1,2]

Buneman, Peter. "Semistructured data." Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. ACM, 1997. [1]
Dey, Lipika, et al. "A framework to integrate unstructured and structured data for enterprise analytics." Information Fusion (FUSION), 2013 16th International Conference on. IEEE, 2013.[2]

Data Gathering
… what industry looks like.



[1]

Definition: Data Lake [2]

*A data lake is a **logical view of all data sources**, in their **raw format**, available and accessible by data scientists to find new insights.*

https://www.reddit.com/r/ProgrammerHumor/comments/ov6l9h/the_secret_component_of_every_data_lake/ [1]
Cedrine Madera and Anne Laurent. 2016. The next information architecture evolution: The data lake wave. 8th International Conference on Management of Digital EcoSystems, MEDES 2016 [2]

Data Gathering
What is a Data Mesh?

Definition: Data mesh

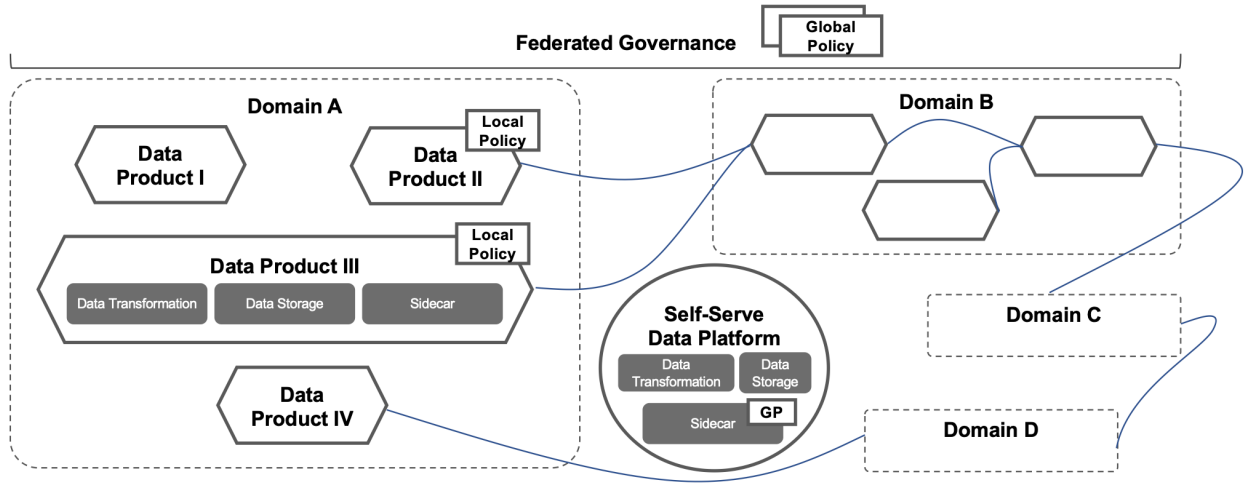
[1]

*Data mesh is a **socio-technical concept** that treats data as a product, uses federated governance, and relies on domain-oriented decentralized data ownership as well as a self-serve data platform to **enable data democratization**.*

Toward Avoiding the Data Mess: Industry Insights From Data Mesh Implementations

JAN BODE[1], NIKLAS KÜHL[2], DOMINIK KREUZBERGER[1], AND CARSTEN HOLTMANN[1,3]
[1]IBM Deutschland GmbH, 71139 Ehningen, Germany
[2]Chair for Information Systems and Human-centric Artificial Intelligence, University of Bayreuth, 95447 Bayreuth, Germany
[3]Karlsruhe Digital Service Research and Innovation Hub, Karlsruhe Institute of Technology, 76133 Karlsruhe, Germany

Link Paper: Data Mesh: Best Practices to Avoid the Data Mess



Conceptual overview of a data mesh based on the four key principles

[1]

1) domain-oriented decentralized data ownership
2) data as a product
3) self-serve data platform
4) federated computational governance

The figure shows different levels of granularity (high on the left and low on the right).

Bode, J., Kühl, N., Kreuzberger, D., & Holtmann, C. (2024). Towards Avoiding the Data Mess: Industry Insights from Data Mesh Implementations. IEEE Access. [1]

https://images.unsplash.com/photo-1578496479273-878c39a9d5e0?ixlib=rb-4.0.3&ixid=M3wxMjA3fDB8MHxwaG90by1wYWdlfHx8fGVufDB8fHx8fA%3D%3D&auto=format&fit=crop&w=3270&q=80

0 Introduction

A Problem Statement

B Data Gathering

C Sampling

D Data Distribution

E Data Quality

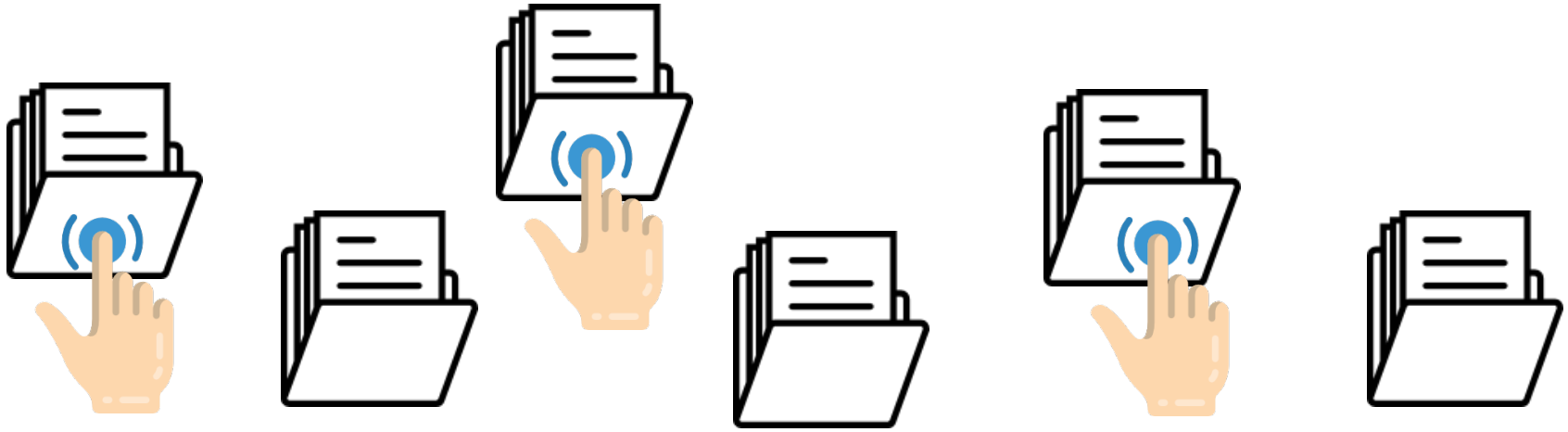
F Data Pre-processing

G Feature Engineering

Sampling
How to identify representative subsets?

- It can be inefficient to use all gathered data for Machine Learning. [1]
  → Data collections can be intelligently reduced by sampling

- **Sampling** is defined as the selection of a representative **subset from large amount of data**



Tillé (2011), Sampling algorithms [1]

Sampling
Exemplary techniques to sample large datasets.

Simple Random Sampling

every data point has the same probability to get selected in the subset.

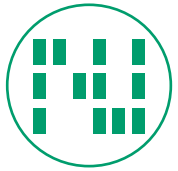
[1]

Systematic sampling

all data points are ordered according to a criteria (e.g. timestamp). A start point is selected randomly and every data point in a specific distance (e.g. every 10th data point) is selected.

Stratified Sampling

every data point is assigned to one of several categories based on a specific criteria (e.g. month). In every category, simple random sampling takes place, and the joint collection of every simple random sampling is the final subset.



Cluster Sampling

every data point is assigned to one of several clusters based on a specific criteria (e.g. geographical location). Some of the clusters are selected randomly and the joint of selected clusters consist the final subset.

Tillé (2011), Sampling algorithms [1]

Sampling
Exemplary sales forecasting for supermarkets.

Predict sales for specific product items in a nationwide supermarket chain.

Problem: Sales data for whole Germany is a very large dataset

1

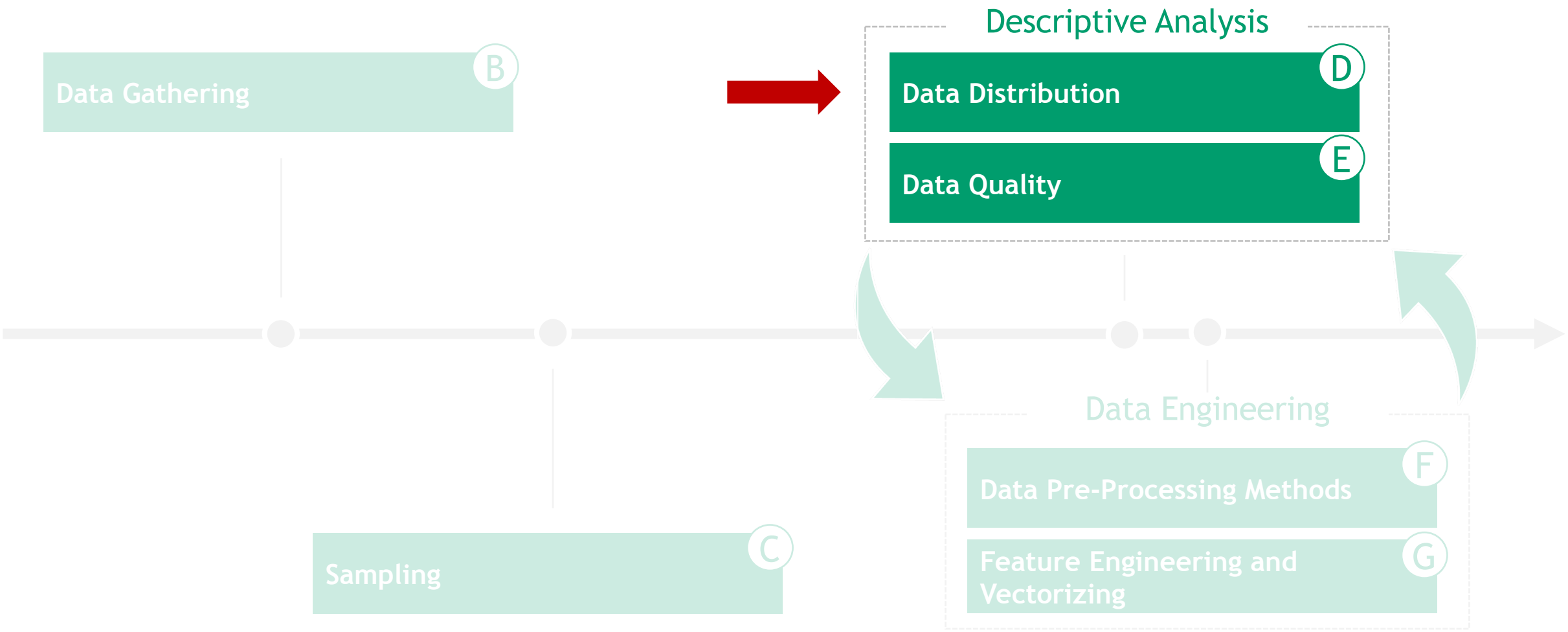
Partition data into several different sales regions, e.g., based on postal codes

2

Sample data from each sales region and train separate prediction model

→ *Allows to train models on smaller amounts of data, thereby reducing memory and computing power requirements*

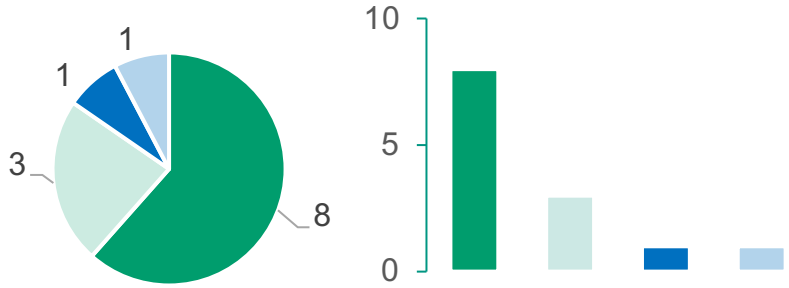
Data Distribution
An essential task of the descriptive analysis.



Data Distribution
What is the distribution of our target variable?

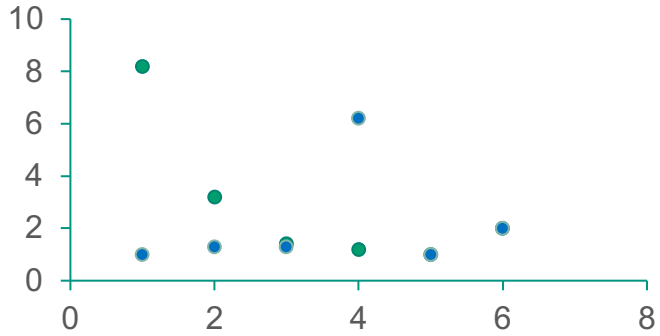
Why? Knowledge about data distributions is key to evaluate results. It can also provide insights if the used algorithms and preprocessing steps are suitable and useful.

How?



Analysis with data charts or histograms, e.g., pie, bar chart. Sometimes it is also useful to plot the data on the basis of two criterions in a two-dimensional diagram.

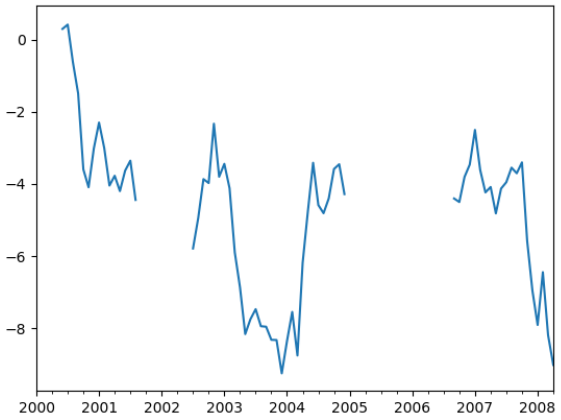
Classification



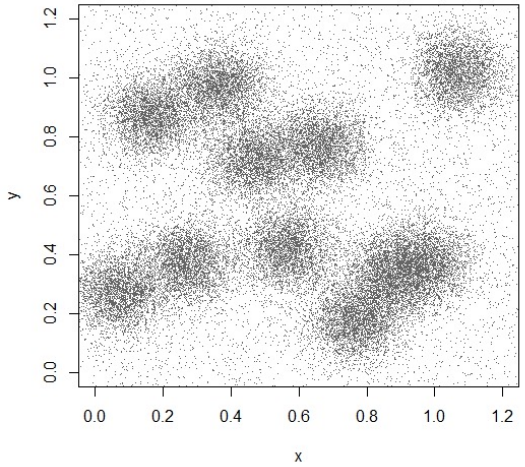
Data can be analyzed, for instance, by plotting the data with the dependent variable on the y-axis and changing independent variables on the x-axis.

Regression

Data Distribution
Sparsity and Noise can be the result of poor data quality.



[1]



[2]

Sparsity

[3]

describes a dataset which mostly contains zeros. Such a data structure is not unusual for machine learning datasets, However, it can cause problems regarding time and space for the machine learning algorithms.

Noise

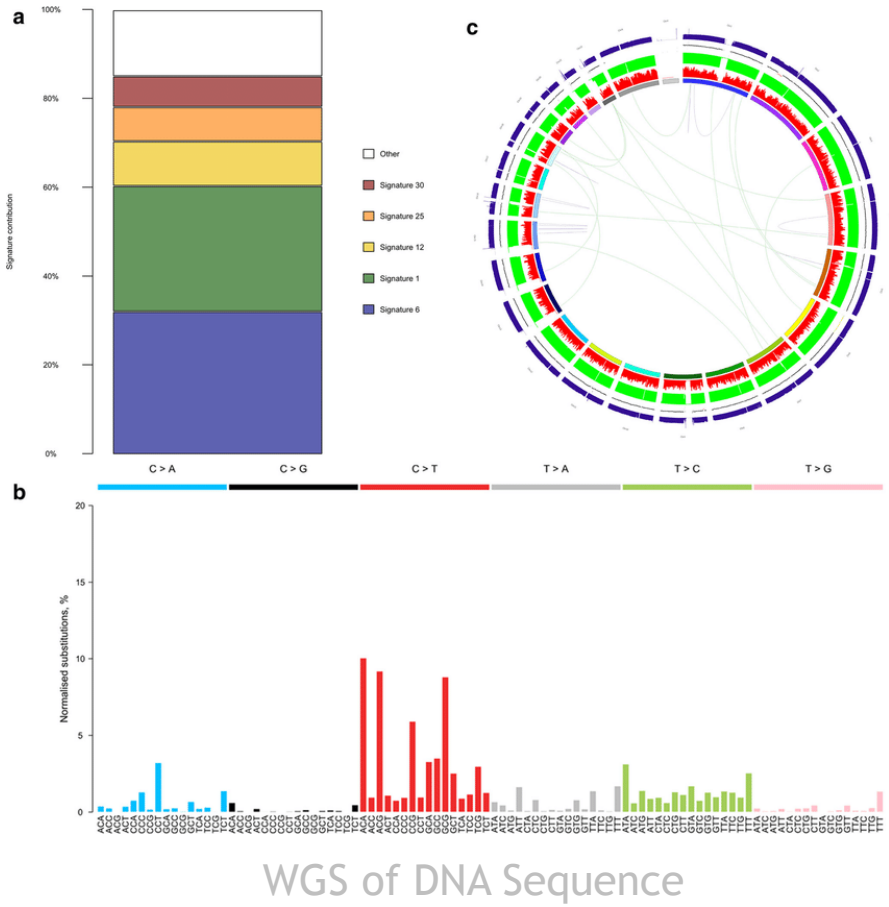
[3]

is a random error of variance of a measured variable. When monitoring real-world data, there is usually some type of noise, caused by e.g., failures of equipment, calibrations or problems during the transmission.

https://miro.medium.com/v2/resize:fit:1280/format:webp/0*ba4bJcUyhEFHOozd.png [1]
https://i.stack.imgur.com/fkHBd.jpg [2]
Duff, Iain S., Albert Maurice Erisman, and John Ker Reid. Direct methods for sparse matrices. Oxford University Press, 2017. , Elias, et al. "Machine learning algorithms: a study on noise sensitivity." Proc. 1st Balcan Conference in Informatics. 2003 [3]

Data Distribution
Example from Research addressing imbalanced datasets.

Examples from research

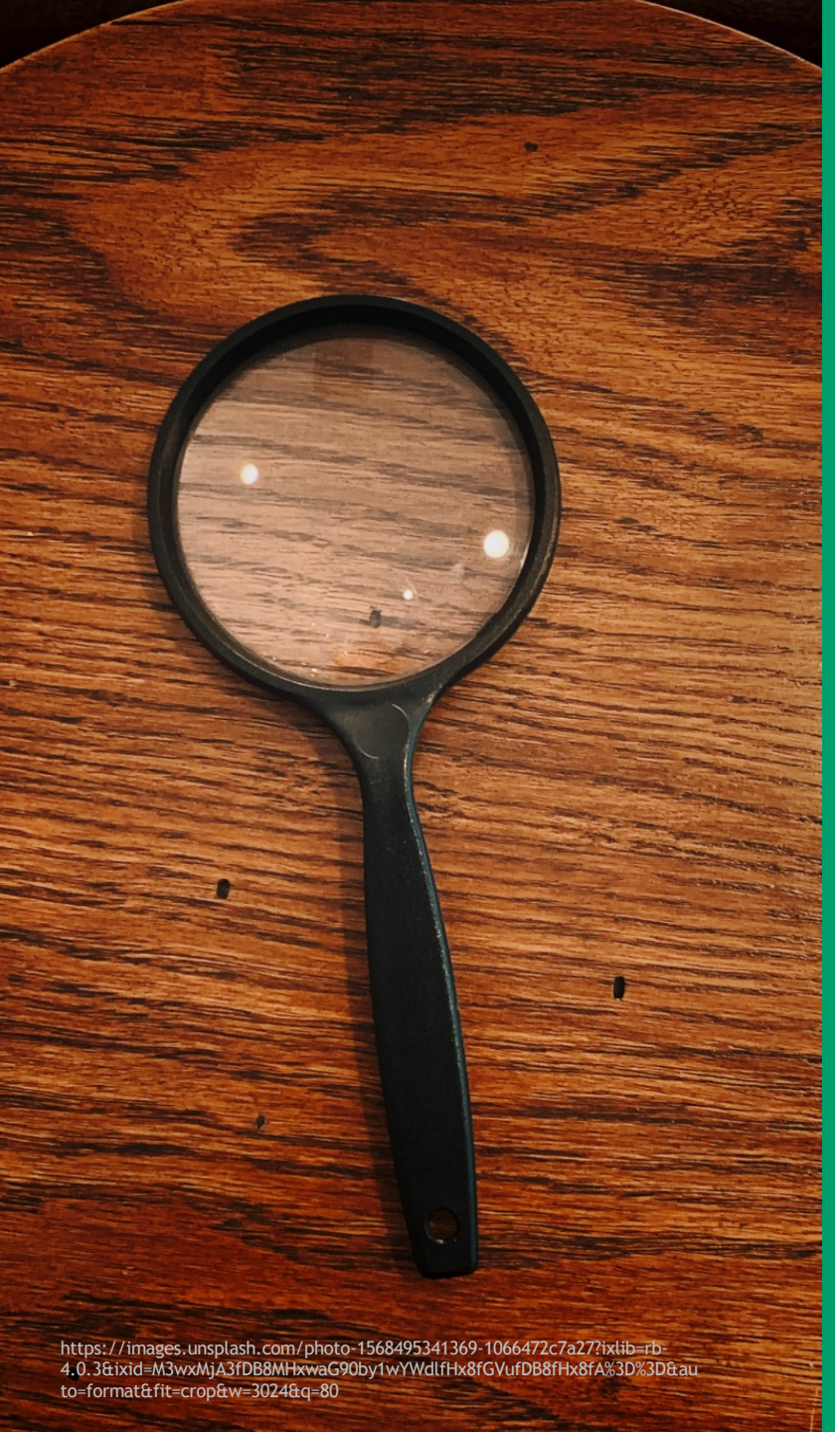


WGS of DNA Sequence

[1,2]

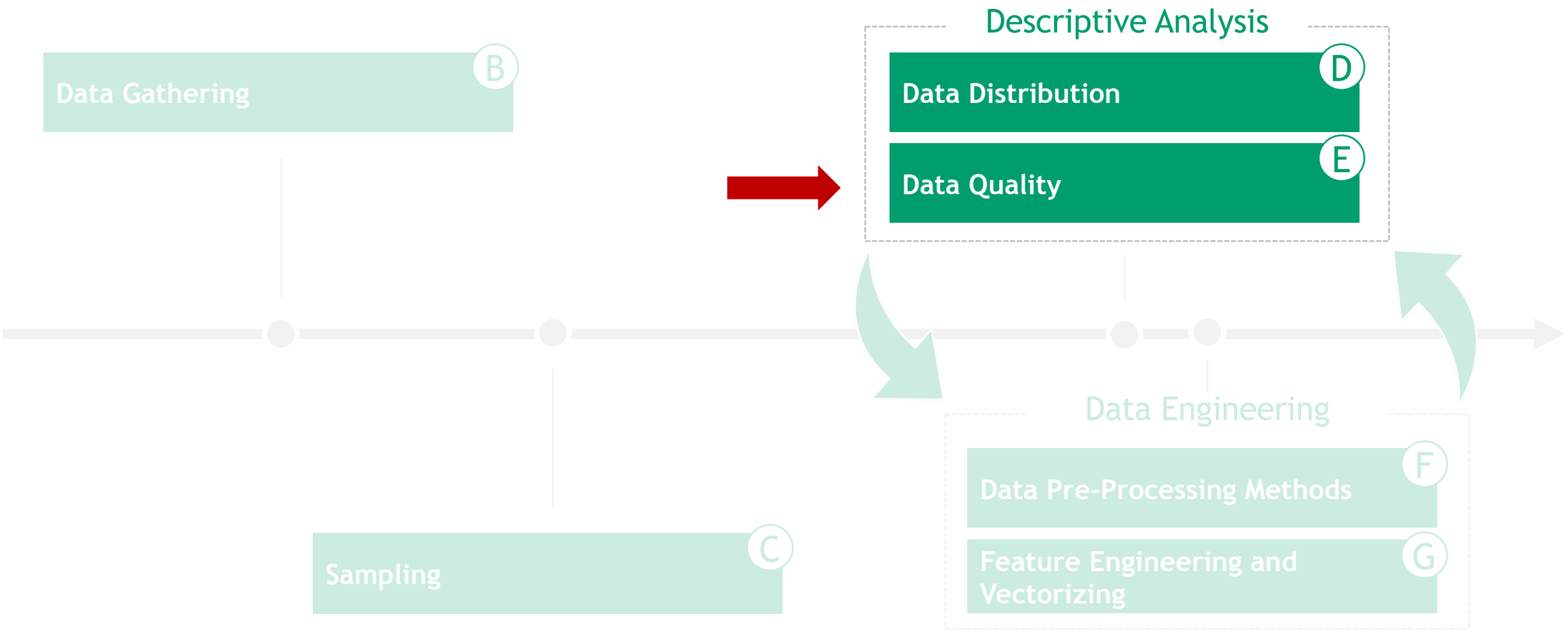
- Whole-genome sequencing (WGS) is a medical technique to determine a DNA sequence completely and at a single time to derive insights about composition.
- ML Model is used to find disease-associated genetic variations.
- Imbalanced Dataset: Positive examples of disease accumulate to several hundred instances, whereas negative examples go into the millions → Next lecture.

Schubach et al. (2017), Imbalance-aware machine learning for predicting rare and common disease-associated non-coding variants [1]
https://www.researchgate.net/publication/333785286/figure/fig2/AS:769750464466944@1560534220202/Figures-from-the-100-000-Genomes-WGS-Report-a-Bar-graph-showing-the-percentage-of-each.ppm [2]



https://images.unsplash.com/photo-1568495341369-1066472c7a27?ixlib=rb-4.0.3&ixid=M3wxMjA3fDB8MHxwaG90by1wYWdlfHx8fGVufDB8fHx8fA%3D%3D&auto=format&fit=crop&w=3024&q=80

Data Quality
Where does data quality come into play?



Data Quality
How can we assess data quality?

[1]

1 **Completeness**
is the proportion of stored data against the potential of a fully

2 **Uniqueness**
measures how many instances are duplicates.

3 **Timeliness**
makes sure that the data is recent enough.

➢ **Motivation:** Avoid "*garbage in, garbage out*" scenario

➢ Regard Data Quality in relation to the initial business problem

checks if the data matches with its predefined syntax.

measures if data is correct or not.

is the absence of difference, when comparing multiple representations of a instance against a definition

Askham et al. (2013), The six primary dimensions for data quality assessment [1]

Data Quality

Data quality can be assessed by six criteria.

[1]

1. **Completeness**
   is the proportion of stored data against the potential of a fully complete dataset.
2. **Uniqueness**
   measures how many instances are duplicates.
3. **Timeliness**
   makes sure that the data is recent enough.
4. **Validity**
   checks if the data matches with its predefined syntax.
5. **Accuracy**
   measures if data is correct or not.
6. **Consistency**
   is the absence of difference, when comparing multiple representations of a instance against a definition

➢ **Typical issues:** Noise, mainly in the area's uniqueness and accuracy

Askham et al. (2013), The six primary dimensions for data quality assessment [1]

Data Quality
Example from Research

Examples from research

Collection period
Mar 15 to May 16
Nov 16 to Feb 17

Keywords

bmw i3; e-tankstelle; eauto; ecar; egolf; electric mobility; electric vehicle; elektroauto; elektrofahrzeug; elektromobil- itaet; elektromobilität; e-mobility; emobility; eup; fortwo electric drive; ladesaeule; ladesäule; miev; nissan leaf; opel ampera; peugeot ion; renault zoe; tesla model s

# collected tweets	107,441
# collected tweets after ad and spam reduction	6,996

Kühl, N., Scheurenbrand, J., & Satzger, G. (2016). Needmining: Identifying micro blog data containing customer needs.

Data Quality
Example from Research

- IT incident tickets provide a documentation of an IT problem description by a costumer as well as the solution progress (and structured data such as time, id, priority etc.)
- Dataset of incident tickets is combined with customer satisfaction survey conducted after completion of the incident process
- Model is built to predict customer satisfaction using the incident ticket data as input
- **Expectation:** Model performance improves when textual problem description is included
- **Result:** Model performance does not improve or even slightly decreases
  - Nearly no emotions detected in text
  - Very technical focus of description
  - Large amount of features compared to data instances → curse of dimensionality

Problem Description [1]
„I checked the error log and there was a dump taken. Please send me the commands to retrieve the dump for you so that we can find out if there is anything that we need to take action on.“
„Local help server has empty content, need further assistance.“
„Need assistance in migrating from ver 4.5 to ver 4.7. Current configuration is ver 4.5 running on a windows 2003 std SP2 machine and I need to migrate over to a windows 2012 R2“

Baier, Kühl, Schüritz, Satzger (2020), Will the customers be happy? Identifying unsatisfied customers from service encounter data [1]


https://images.unsplash.com/photo-1606337321936-02d1b1a4d5ef?ixlib=rb-4.0.3&ixid=M3wxMjA3fDB8MHxwaG90by1wYWdlfHx8fGVufDB8fHx8fA%3D%3D&auto=format&fit=crop&w=3000&q=80

0 Introduction

A Problem Statement

B Data Gathering

C Sampling

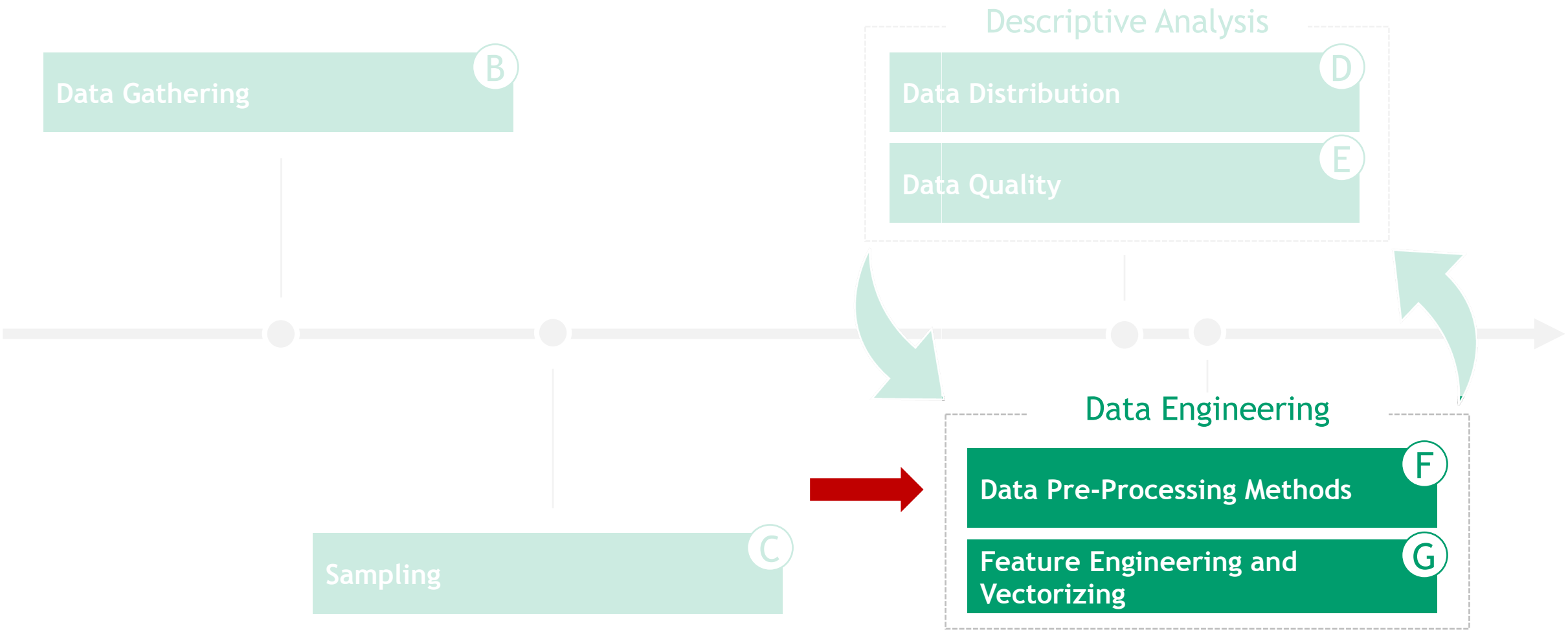
D Data Distribution

E Data Quality

F Data Pre-processing

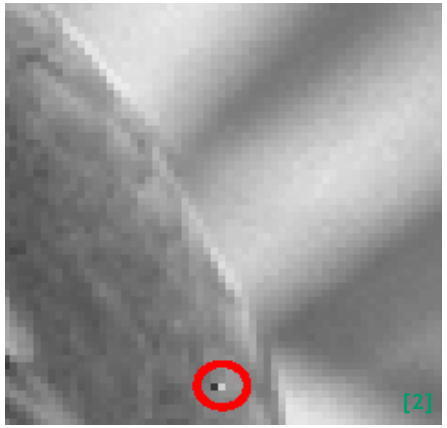
G Feature Engineering

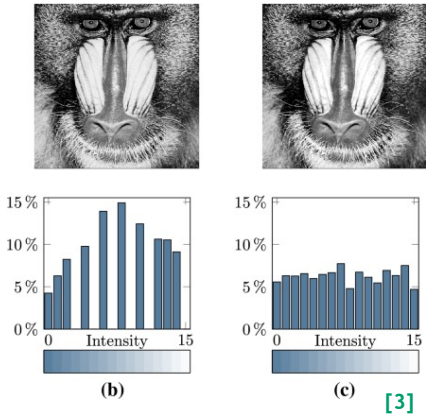
Data Pre-Processing
Where we are on the map.



Data Preprocessing | Images

Clean image data before it is ready to be used in a ML model.


[2]



[3]


[4]


[5]


[6]

Sensor Correction	Lightning Correction	Noise Correction	Geometric Correction	Colour Correction
Include dead pixel correction, geometric lens distortion, and vignetting.	*include rank filtering, histogram equalization, and LUT remap.*	*Include noise removal, e.g., using smoothing.*	*Include rotating, flipping and change of perspective.*	*Include redistribution of color saturation or correction for illumination artifacts.*

[1]

https://www.embedded-vision.com/sites/default/files/apress/computervisionmetrics/chapter2/9781430259299_Ch02.pdf [1]
https://d3i71xaburhd42.cloudfront.net/635c7d86290e2c208c1632fed0e223d79f37e0af/6-Figure11-1.png [2]
https://media.springernature.com/lw685/springer-static/image/art%3A10.1007%2Fs00371-022-02723-8/MediaObjects/371_2022_2723_Fig1_HTML.png [3]
https://assets.website-files.com/6005fac27a49a9cd477afb63/609c398641edcb28c15bce0e_Brian_Matiash_Wildlife01_Settings.jpg [4]
https://img.fotocommunity.com/schief-4f984272-d4a7-4c52-8839-70f70d69e16a.jpg?height=1080 [5]
https://phlearn.com/wp-content/uploads/2012/08/color-landscapes-1200px.jpg [6]

This started as a text, meeting all requirements.

Data Preprocessing | Text

How to structure unstructured text data.

{This, started, as, a, text, meeting, all, requirements}

{determiner, verb, preposition, determiner, noun, verb, determiner, noun}

“Thi start as a text , meet all requir ”

*“This started as a text, **meeting** all requirement”*

e.g., “assembly“ vs “consultation”

Tokenization	Part-of-speech (POS)	Stemming	Lemmatization
process of separating a string of text into its component words.	*(POS) tagging is the process of identifying grammatical word classes.*	*algorithmic process of reducing words to their stem, base or root form. The result does not necessarily include real words.*	*looks up each word in a dictionary and analyzes the POS to select the correct form in case of disambiguation.*

Deep dive in Lecture 6: *Large Language Models*

Data Preprocessing | Text
How to structure unstructured text data.

"started text, meeting all requirement"

has (among other) the bigram (n=2) element *"all requirements"* or the trigram (n=3) element *"meeting all requirements"*

For "meeting"
Nouns:
{meeting, group meeting}, {merging, meeting, coming together}
Verbs:
{meet, encounter, receive}, {suffer, meet}

For {meet, encounter, receive} the hypernym is *{have, experience}*, which itself has again the hypernym *{change}*

Substitution	N-grams	Synsets	Hypernyms
Symbol/word/character removal allows to remove unimportant features. Common examples are emoticon substitution or stop word removal.	*sets of co-occurring words or concepts with the given window n.*	*Synonym sets are groups of expressions that comprise the same concept.*	*depict a superordinate relationship from a semantic perspective.*
		Semantic Pre-processing	

Data Preprocessing
How to structure unstructured text data.

Data Preprocessing
How can we transform text into trainable numbers?

@regsprecher: We need more charging Stations and it has to happen soon!

we, need, more, charge, station, have, happen, soon

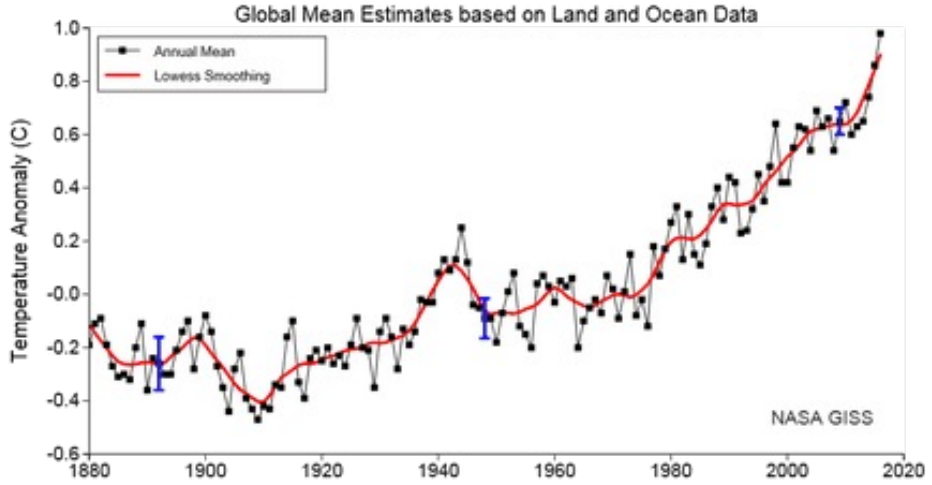
tweetid	user	(...)	car	like	we	need	more	charg	station	(...)
123	a_1		0	0	1	1	1	1	1	
126	a_2		1	1	0	0	0	1	1	
...	...		...	...	...	...	...	...	...	

- Pre-processing
  1. Username removal
  2. Downcasing
  3. Tokenization
  4. Stemming
  5. Stop word removal
- Feature generation
  1. Conversation to vector features
  2. Represent Boolean token presence

Bag-of-Words transformation

Data Preprocessing | Numbers

Remove outliers in data and adapt to model requirements.

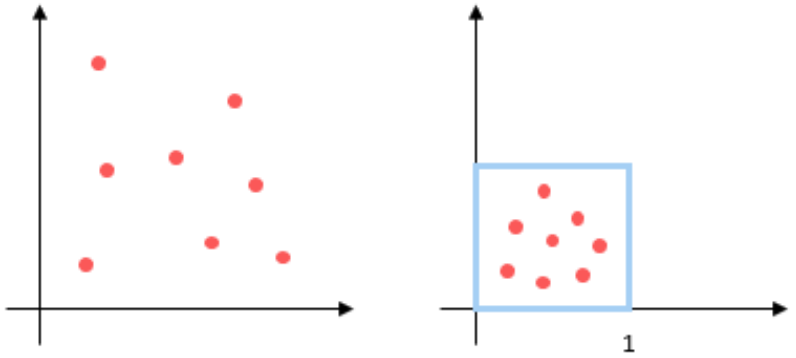


[2]

Smoothing

[1]

*is a technique to **reduce the effect of outliers** towards the bulk of data points. Outliers will be weighted with a lower weight compared to other, more common data points. This technique should remove noise and reveal the underlying data.*



[3]

Normalizing

[1]

*is a technique to **transform the ranges of different attributes** into one. Some machine learning algorithms will favor attributes with a larger range than others.*

Schütze et al. (2008), Introduction to information retrieval ; Salkind (2010), Encyclopaedia of research design [1]
https://i.guim.co.uk/img/media/aa7d42bf15cbd44c74eac8f23f73c7d61b782e6e/0_0_904_480/master/904.png?width=620&dpr=2&s=none [2]
https://miro.medium.com/v2/resize:fit:1400/format:webp/1*HW7-kYjj6RKwrO-5WTLkDA.png [3]

Data Preprocessing | Numbers
Remove outliers in data and adapt to model requirements.

[2]

Example: A dataset contains 10 datapoints. Every datapoint with two previous datapoints is not used with its original value but with the mean of its own value and the value of its two predecessors. This smoothing method is called simple moving average.

[3]

Example: A dataset contains human weight data from 50 to 100kg and the length of their thumbs from 5 to 9cm. A min-max normalization reduces both ranges to [0, 1] by subtracting the corresponding minimum value of every instance in each dataset and dividing the result by the span of the corresponding range.

Smoothing [1]

*is a technique to **reduce the effect of outliers** towards the bulk of data points. Outliers will be weighted with a lower weight compared to other, more common data points. This technique should remove noise and reveal the underlying data.*

Normalizing [1]

*is a technique to **transform the ranges of different attributes** into one. Some machine learning algorithms will favor attributes with a larger range than others.*

Schütze et al. (2008), Introduction to information retrieval ; Salkind (2010), Encyclopaedia of research design [1]
https://i.guim.co.uk/img/media/aa7d42bf15cbd44c74eac8f23f73c7d61b782e6e/0_0_904_480/master/904.png?width=620&dpr=2&s=none [2]
https://miro.medium.com/v2/resize:fit:1400/format:webp/1*HW7-kYjj6RKwrO-5WTLkDA.png [3]

Data Preprocessing | Numbers
How to deal with time series data.

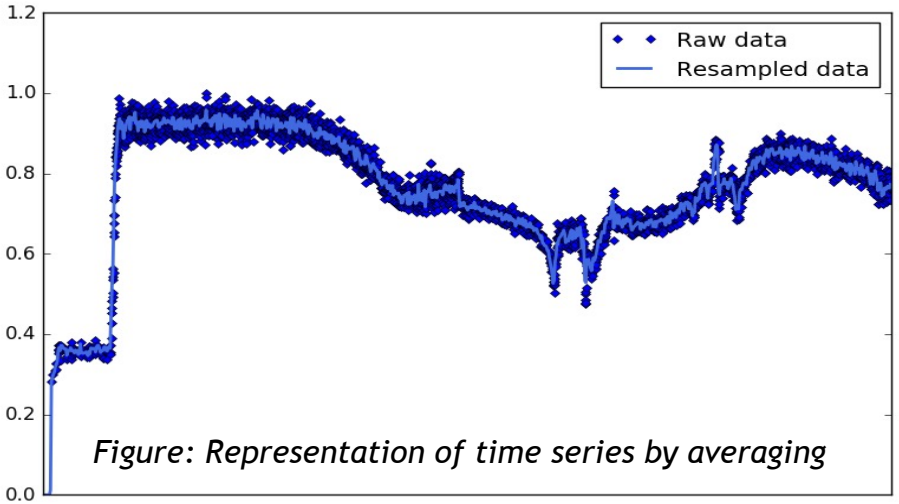


Figure: Representation of time series by averaging

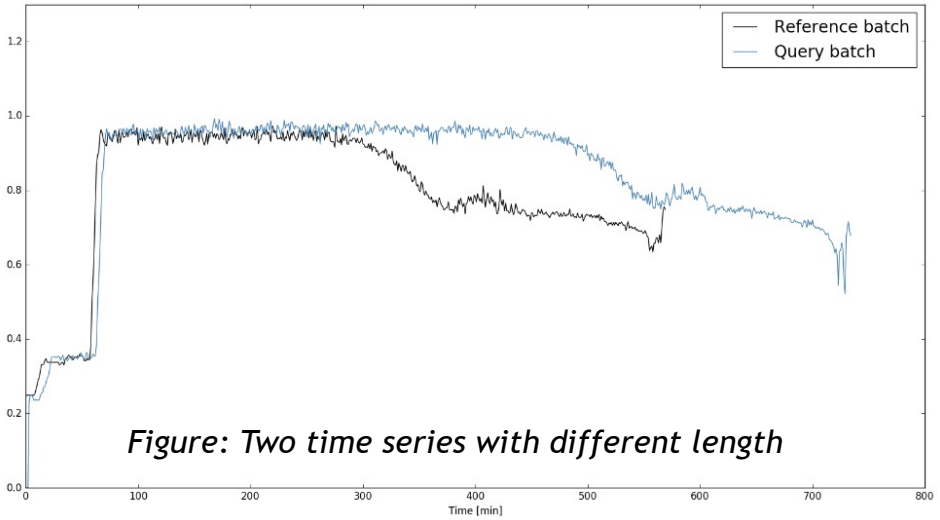


Figure: Two time series with different length

Replacement of missing values	Representation	Alignment
Value previous or after missing value, mean, median, interop-lation	*Sampling, Average, Linear interpolation, only important points such as minima and maxima, Symbolic representation, PCA, t-SNE ...*	*Simple alignment methods, indicator variable, Correlation, Optimized Warping, Dynamic Time Warping…*

Data Preprocessing | Numbers
Time series example from research.

Examples from research

Use case:	Prediction of faults in a chemical batch production process.
Challenge:	Every batch run has a different duration until the final chemical product is retrieved.

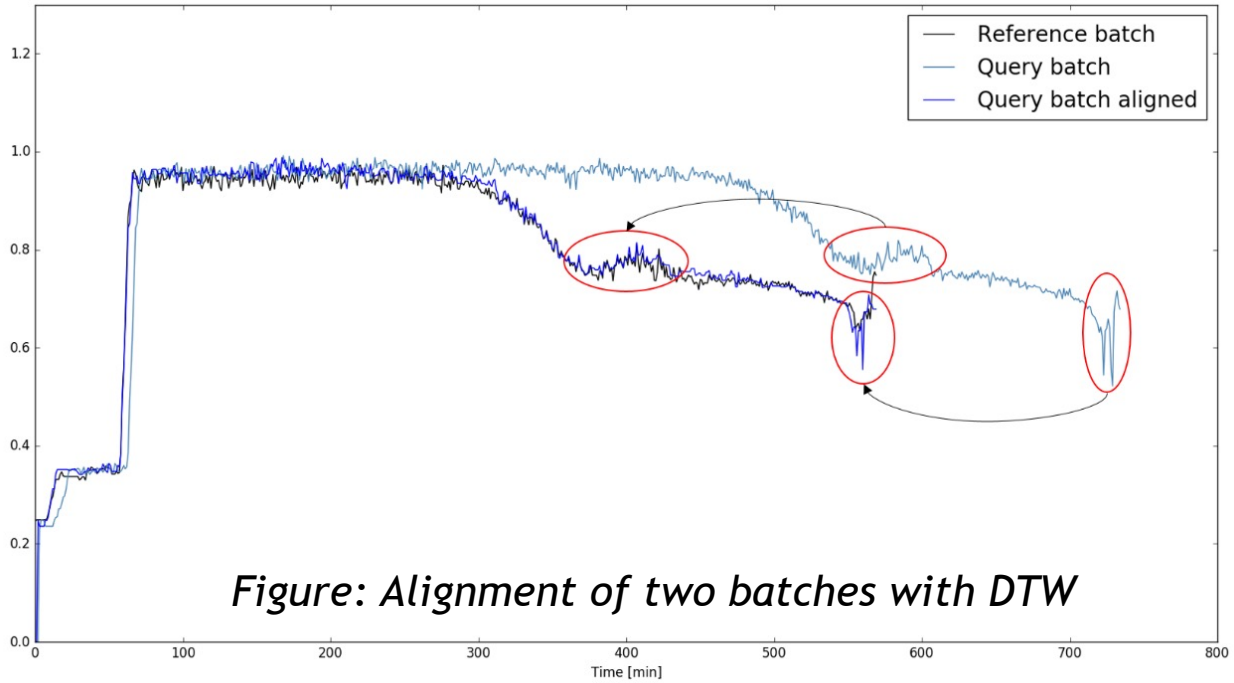
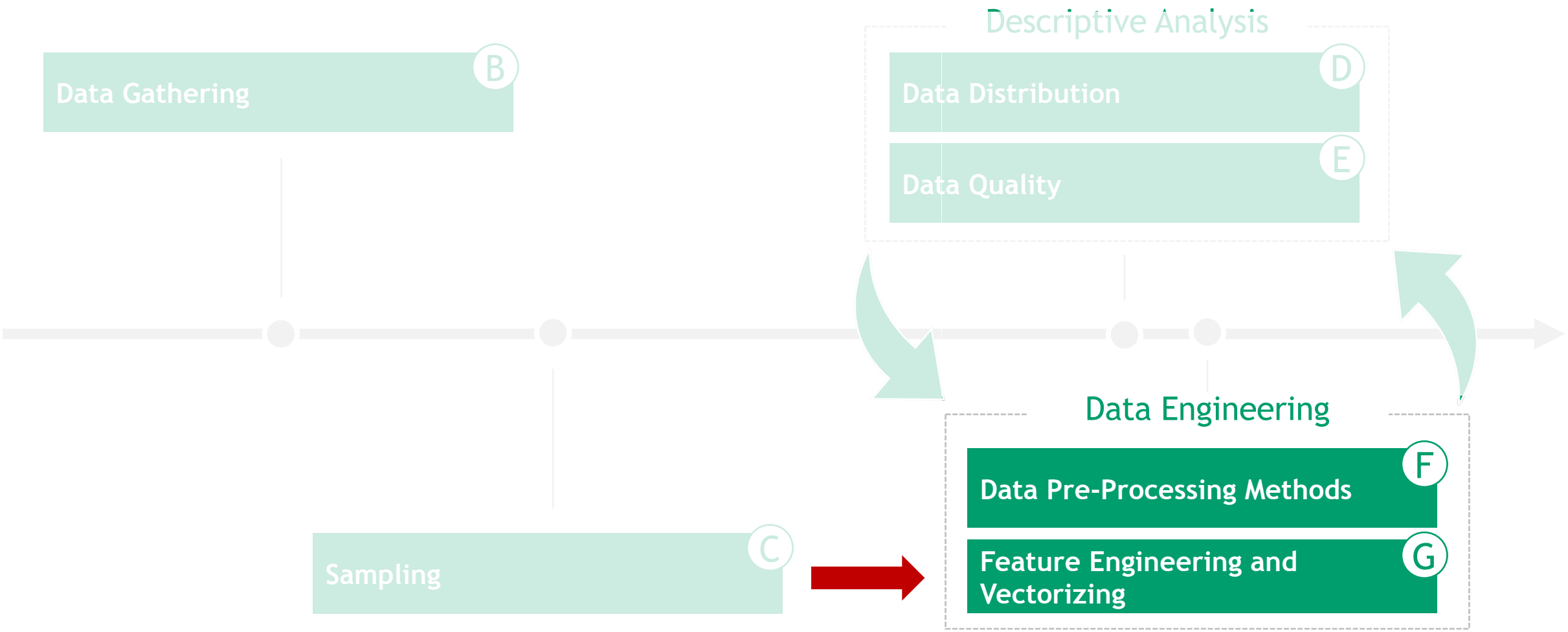


Figure: Alignment of two batches with DTW

- Use Dynamic Time Warping (DTW) for the alignment of batch data
  - Similar patterns of time series are matched
  - Also allows for an online alignment

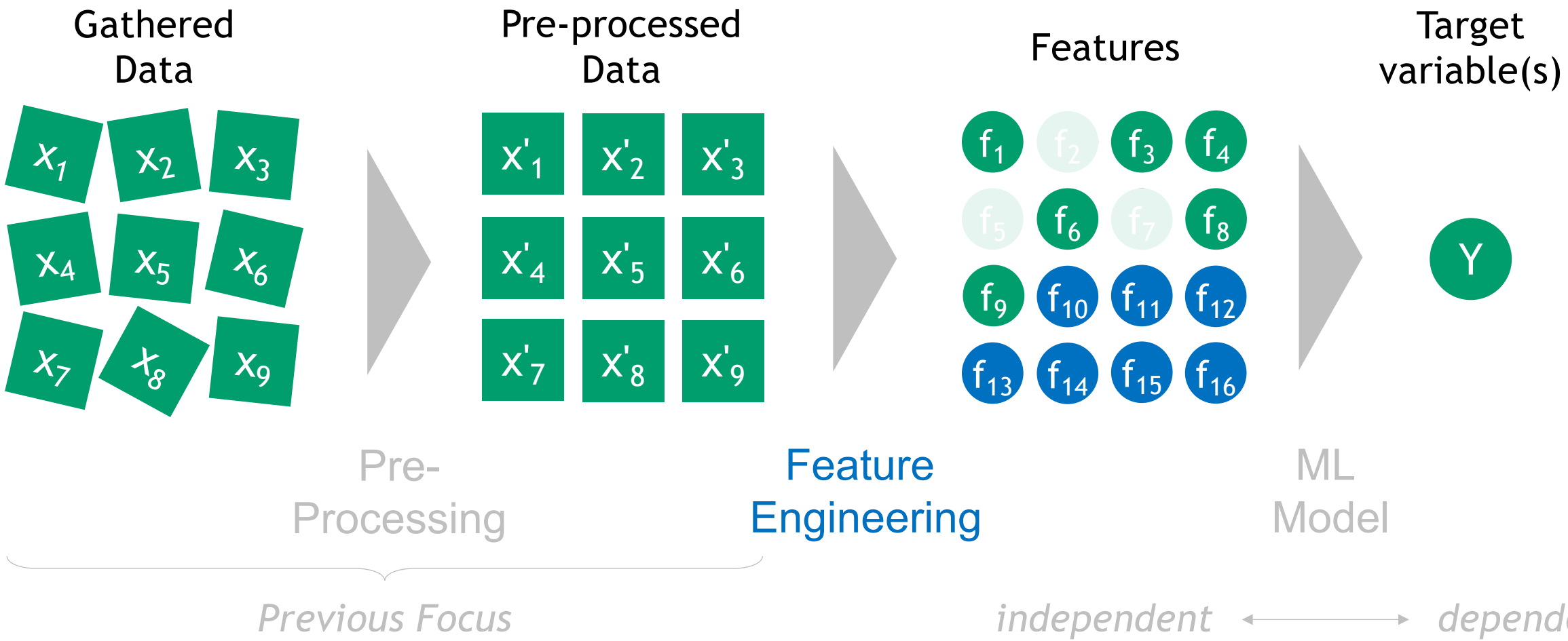
- 0 Introduction
- A Problem Statement
- B Data Gathering
- C Sampling
- D Data Distribution
- E Data Quality
- F Data Pre-processing
- **G Feature Engineering**

Feature Engineering
An essential step of data engineering.



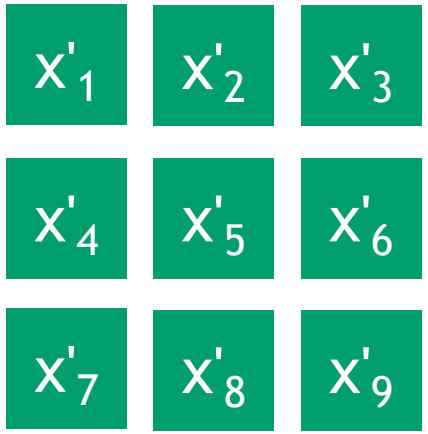
Feature Engineering

We can incorporate domain knowledge.

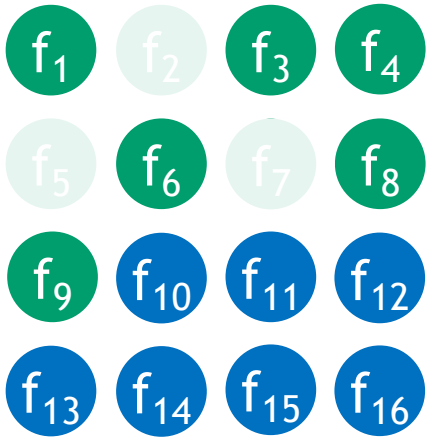


Feature Engineering
We can incorporate domain knowledge.

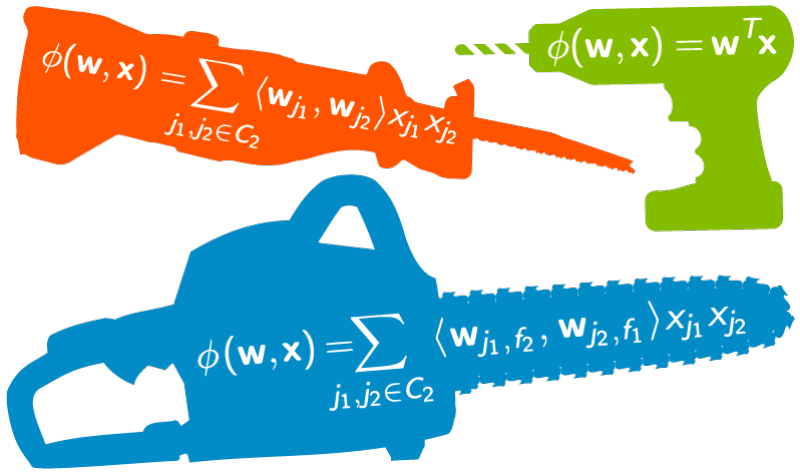
Pre-processed Data



Features



Use features to create new features. It can either be necessary because the feature is not processible by machine algorithms (e.g., text) or useful — and result in a better performance because it reveals underlying information of the data.



Feature engineering makes use of a variety of methods, most of which involve mathematical procedures.

Feature Engineering
Exemplary methods to work with numeric data.

Raw Data	numeric data can usually be consumed by machine learning algorithms without any further adaptations.

[1]

Statistical Measures	Binarization	Rounding	Binning	Monotonic Transformation
min, max, mean, median, var, count, quantiles …	reduces the feature to either one or zero. This is used when the frequency is not important.	*reduces the precision of the data points by converting the values into numeric integers.*	*assigns data points to discrete feature characteristics (bins).* *- **Fixed-width binning:** pre-fixed range of bins* *- **Adaptive binning:** pre-fixed number of bins*	*log-transformation or the box-cox transformation stabilize variance and make the data more normal distribution like.*

Sarkar (2018), Understanding Feature Engineering ; towardsdatascience.com [1]

Feature Engineering
Categorical Data

Transforming:
of categorical values into numeric labels [1]

- **Transforming nominal attributes:** with *no sense of order* amongst them into numeric labels.
- **Transforming ordinal attributes:** with *a sense of order* amongst them into numeric labels.

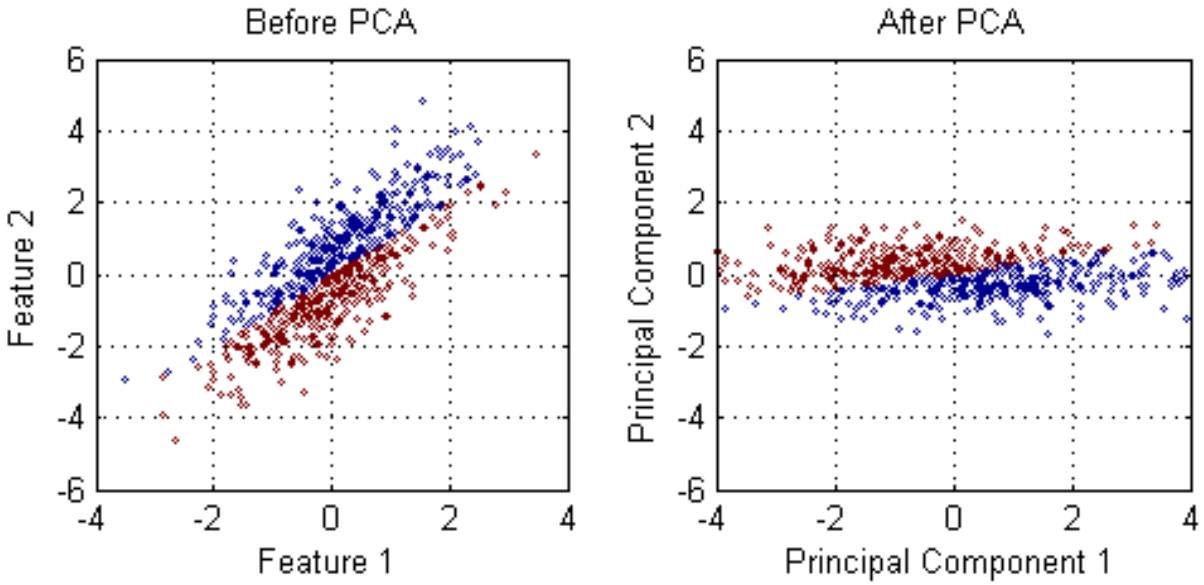
Encoding:
of categorical values is performed after transformation. The numeric labels are not a continuous numeric feature because they cannot be compared directly. Therefore, encoding adds dummy features for each unique numeric label. [1]

- **One-hot Encoding Scheme:** transforms n distinct numeric labels into n binary features, write 1 when the instance contains numeric label and 0 if not.
- **Dummy Coding Scheme:** transforms *n* distinct numeric labels into *n-1* binary features. One numeric label is represented by a vector of all zeros for all *n-1* binary features.

Sarkar (2018), Understanding Feature Engineering ; towardsdatascience.com [1]

Feature Engineering
Dimension reduction helps to concentrate information.

Besides manual feature engineering, automated feature engineering methods are statistical, computerized techniques to support feature engineering, e.g., PCA, t-SNE [1]



Principal component analysis (PCA) : is needed when features are correlated to each other. It calculates a (smaller) set of artificial features which are linearly uncorrelated to each other.

- This example with two features illustrates the effect of a PCA. Before PCA, both features are highly correlated whereas the new, artificial features are uncorrelated to each other.

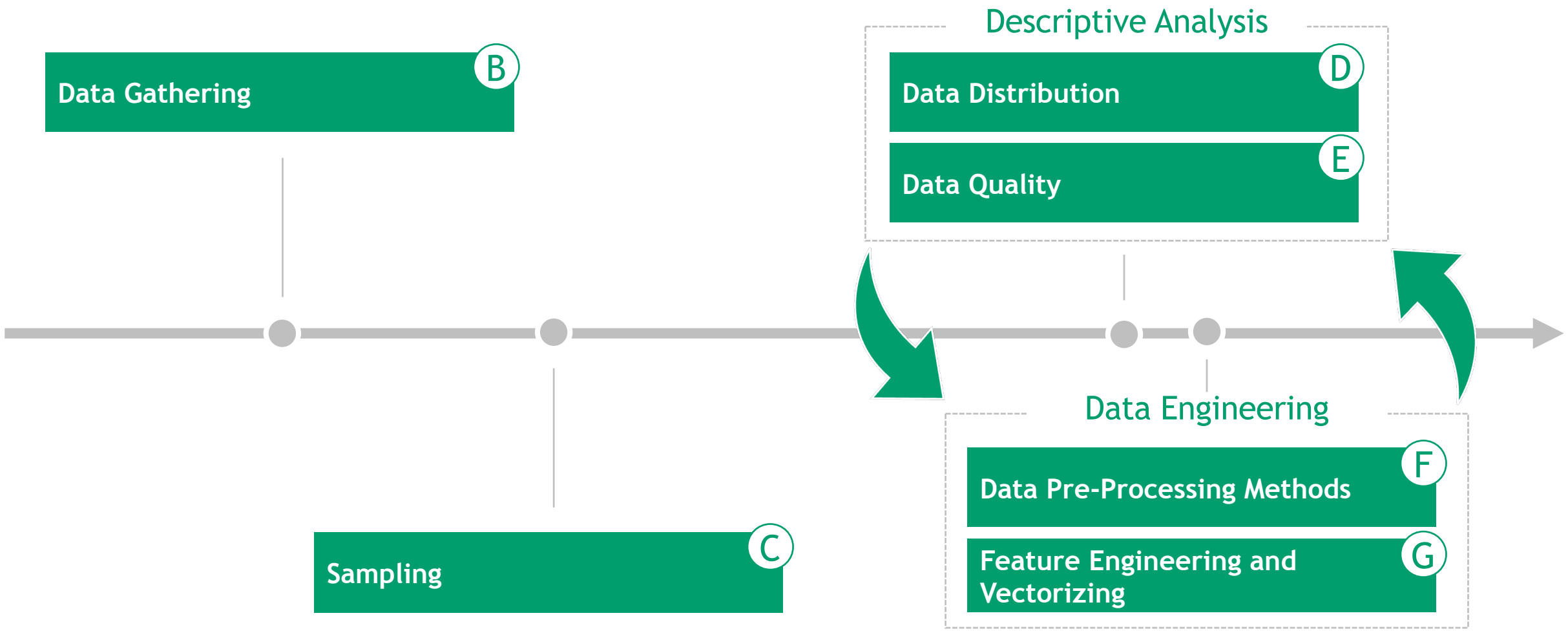
Jolliffe (2011), Principal component analysis ; Li et al. (2015), A robust deep model for improved classification of AD/MCI patients [1]

Feature Engineering
Include automated feature engineering in process.

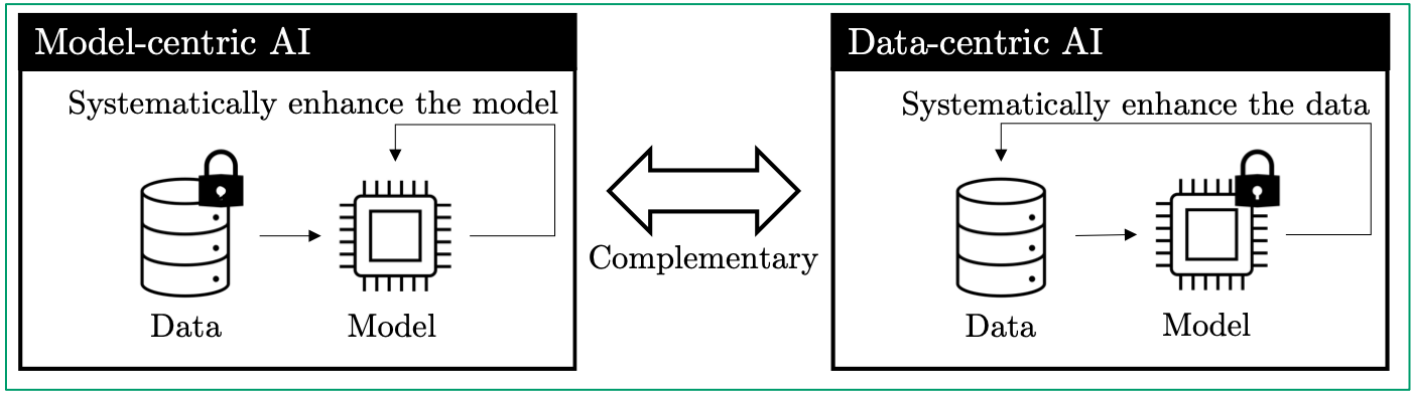
Automated Feature Engineering	Automates the feature engineering process and extracts useful and meaningful features of the given dataset, e.g., brute force, with deep feature synthesis or by neural networks.	[1]
Deep feature synthesis	is one of the most promising steps to perform automated feature engineering. It stacks primitive calculations to generate more complex features ➢ **Aggregation primitives:** use related instances as input and output a single value, e.g., sum, count, average. ➢ **Transform primitives:** are applied to a single instance. They take one or more features as input and output a new feature, e.g., convert to absolute, round, logical AND.	[2]

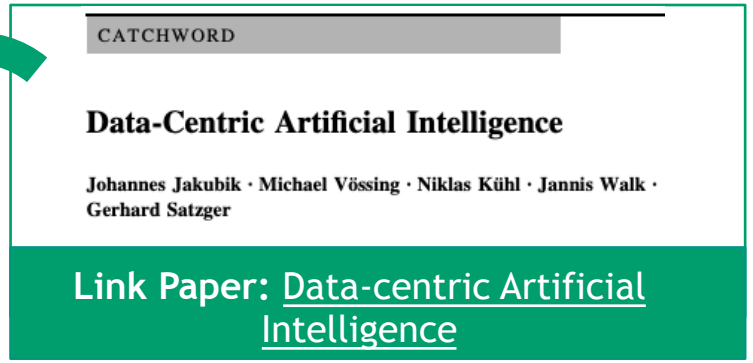
Khurana et al. (2017), Feature Engineering for Predictive Modeling using Reinforcement Learning ; Featuretools.com [1]
Kanter and Veeramachaneni (2015) Deep feature synthesis: Towards automating data science endeavors [2]

The AI Lifecycle
Coming back to the initiation process.



The AI Lifecycle
New phenomena: Data-Centric AI (1/3)





Link Paper: Data-centric Artificial Intelligence

Definition: Model-centric Artificial Intelligence [1]

Model-centric artificial intelligence is the paradigm emphasizing that the choice of the suitable model type, architecture, and hyperparameters from a wide range of possibilities is essential for building effective and efficient AI-based systems.

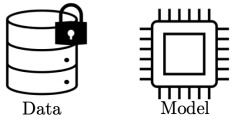
Definition: Data-centric Artificial Intelligence [1]

Data-centric artificial intelligence is the paradigm emphasizing that the systematic design and engineering of data are essential for building effective and efficient AI-based systems.

Jakubik, J., Vössing, M., Kühl, N., Walk, J., & Satzger, G. (2024). Data-centric artificial intelligence. Business & Information Systems Engineering, 1-9 [1]

The AI Lifecycle
New phenomena: Data-Centric AI (2/3)

Model-centric paradigm: Systematically enhance the model



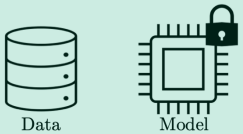
Improve quality on an instance level

- Improve model selection, architecture, and hyperparameters

Data-centric paradigm: Systematically enhance the data

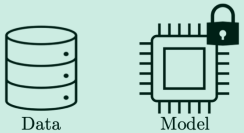
[1]

Refine (better data)



Improve data quality on an instance level

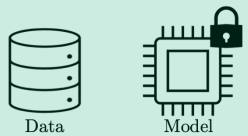
- R1 Improve feature quality (e.g., semi-automated identification of corrupted pixels in images)
- R2 Improve label quality (e.g., semi-automated identification of label errors)



Improve data quality on a dataset level

- R3 Increase volume of high-relevance instances (e.g., semi-automated augmentation of important edge cases)
- R4 Reduce volume of low-quality instances (e.g., semi-automated tools to group the dataset by data quality levels)
- R5 Increase volume of relevant features (e.g., semi-automated tools for synthetic feature generation)
- R6 Reduce volume of unmeaningful features (e.g., semi-automated identification of irrelevant features)

Extend (more data)



Acquire new data for blind spots on a dataset level

- E1 Acquire new instances (e.g., semi-automated estimation of the required type of additional images)
- E2 Acquire new features (e.g., semi-automated assessment of the benefit from additional sensors)
- E3 Acquire new labels (e.g., semi-automated tools to determine the sequence of labeling images)

Jakubik, J., Vössing, M., Kühl, N., Walk, J., & Satzger, G. (2024). Data-centric artificial intelligence. Business & Information Systems Engineering, 1-9 [1]

The AI Lifecycle
New phenomena: Data-Centric AI (3/3)

Tabular Data

	f_1	f_2	f_3	f_{new}	
x_1	10	1.76	A		
x_2	20	NaN	B		
x_3	15	9.01	C		
x_{new}					

Change the volume of instances as part of R3, R4, and E1 by, e.g., utilizing synthetic data.

Improve feature or target quality as part of R1 and R2, and E3 by, e.g., utilizing imputations.

Change the volume of features as part of R5, R6, and E2 by, e.g., engineering features.

Image Data

Instance

Change the volume of instances as part of R3, R4, and E1 by, e.g., adding augmented images.

Improve feature or target quality as part of R1, R2, and E3 by, e.g., correcting target label.

~~"Rotweiler"~~ "Labrador retriever"

Change the volume of features as part of R5, R6, and E2 by, for example, cropping an image.

[1]

Jakubik, J., Vössing, M., Kühl, N., Walk, J., & Satzger, G. (2024). Data-centric artificial intelligence. Business & Information Systems Engineering, 1-9 [1]

Summary

Problem Statement	Data	Analysis	Preprocessing
A well-defined problem statement with objectives and business success criteria is key for a successful AI / machine learning application.	*Data can be separated based on different criteria, e.g., its structure, origin and type.*	*An analysis of the data distribution and its quality gives first insights for further steps of supervised machine learning.*	*Data preprocessing and feature engineering prepare a given dataset to be used by machine learning algorithms.*