Artificial Intelligence and Data Analytics for Engineers (AIDAE)

Lecture 3
May, 15th

Anas Abdelrazeq
Andrés Posada
Marco Kemmerling Today's Lecturer
Vladimir Samsonov

Learning Objectives

Learning Objective w.r.t. Knowledge/Understanding. After successfully completing this lecture, the students will have achieved the following learning outcomes:

- Have an understanding of why data preparation is an important step in the analysis process.
- Know about the different methods and tools in data preparation.
- Know about difference in data preparation with regard to various modalities.

Recap Pandas/Matplotlib/Scikits Learn

Pandas

- Data manipulation (mostly tables)
- Data analysis

How to install?

- conda install pandas

How does it look like?

```
import pandas as pd
df = pd.read_csv('data.csv')
```

Pandas: reading/saving

```python
# read write
import pandas as pd
df = pd.read_csv('data.csv')
df.to_csv("data_file.csv")

df_2 = pd.read_excel("file.xlsx")
df_2.to_excel("dir/file.xlsx", sheet_name="sheet")

from sqlalchemy import create_engine
engine = create_engine('sqlite:///foo.db')
df_3 = pd.read_sql_table("tableName", engine)
df_3.to_sql("tableName",engine)
```

Pandas: information and filtering

```
data = [["tom",10],["pete",15],["jean",30],["puff",35],["pete",5]]
df = pd.DataFrame(data=data, columns=["name","age"])

# info
df.columns
df.shape
df.info()

# filters
df[df.name == "tom"]
df[df.age > 15]
df[df.name == "tom"]
df[(df.age > 10) & (df.name == "pete")]

df.iloc[0] # by position
```

Pandas: operations

```
# operations
df["age"].sum()
df["age"].cumsum()
df["age"].min()
df["age"].max()
df["age"].mean()
df["age"].median()

sum_one = lambda x: x + 1
df["new age"] = df["age"].apply(sum_one)

upper = lambda s: s[0].upper() + s[1:]
df["name"] = df["name"].apply(upper)
```

Resources and Libraries

Matplotlib

- Plots
- More plots

How to install?

- conda install matplotlib

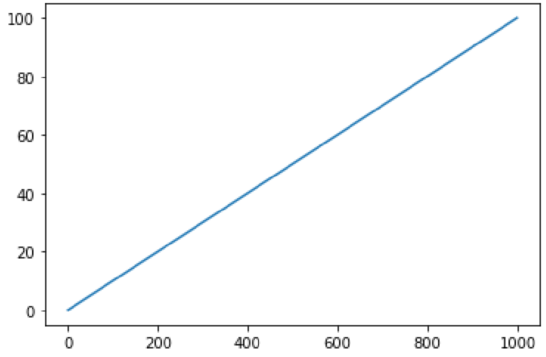
How does it look like?

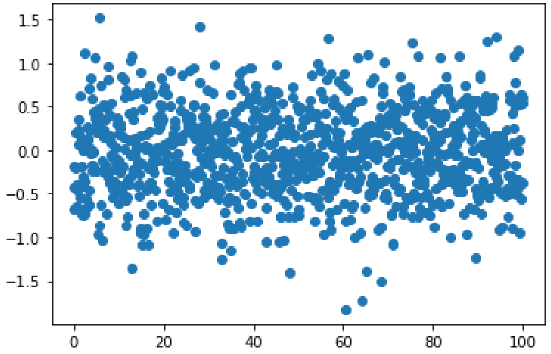
```
import matplotlib.pyplot as plt
plt.plot([1,5,4,2,5,1,4,5])
plt.show()
```

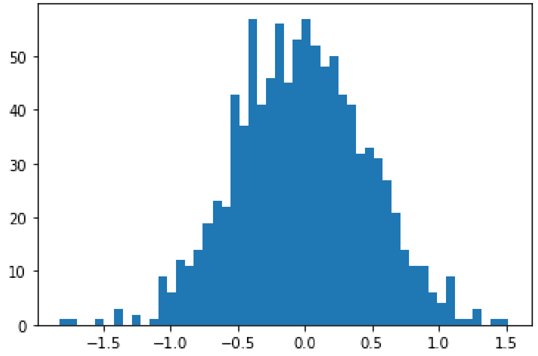
Resources and Libraries

Matplotlib: plots

```
# plots
import matplotlib.pyplot as plt
x = np.linspace(0, 100, 1000)
y = np.random.normal(0, 0.5, 1000)


plt.plot(x)
plt.show()

plt.scatter(x, y)
plt.show()

plt.hist(y, bins = 50)
plt.show()
```

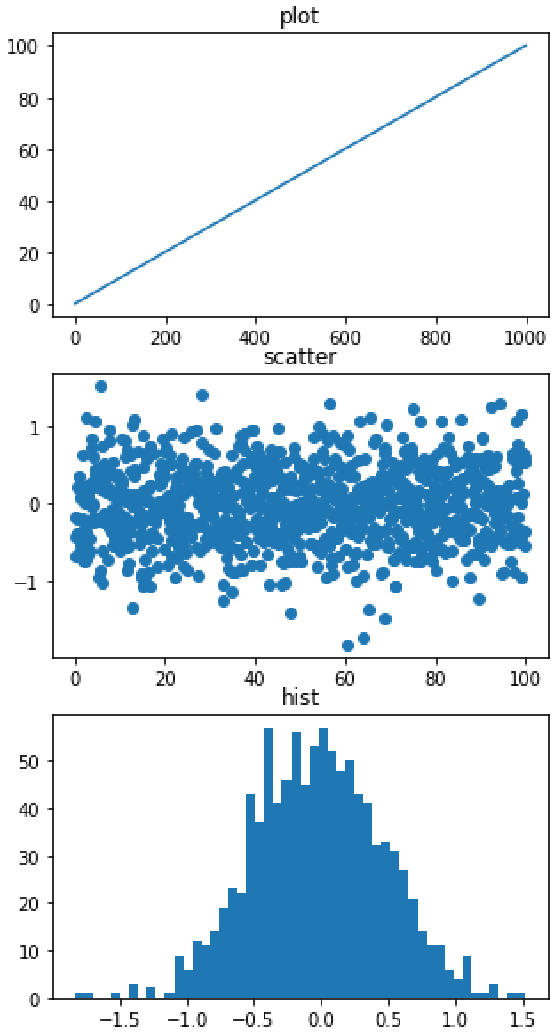






Matplotlib: subplots

```
# sub-plots
plt.figure(figsize=(5,10))

plt.subplot(3,1,1)
plt.plot(x)
plt.title("plot")

plt.subplot(3,1,2)
plt.scatter(x,z)
plt.title("scatter")

plt.subplot(3,1,3)
plt.hist(z, bins = 50)
plt.title("hist")

plt.show()
```



Resources and Libraries

Scikit Learn

- data mining
- data analysis

How to install?

- conda install scikit-learn

How does it look like?

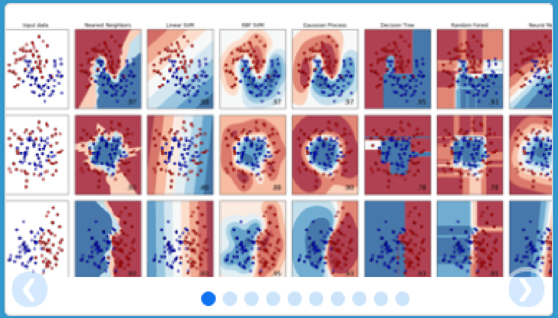
```python
from sklearn import tree
X = [[0, 0], [1, 1]]
Y = [0, 1]
# model
clf = tree.DecisionTreeClassifier()
clf = clf.fit(X, Y)
clf.predict([[2., 2.]])
```

Resources and Libraries

scikit learn

Home | Installation | Documentation | Examples

Google Custom Search



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.
Algorithms: SVM, nearest neighbors, random forest, …

— Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, ridge regression, Lasso, …

— Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, …

— Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: PCA, feature selection, non-negative matrix factorization.

— Examples

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning
Modules: grid search, cross validation, metrics.

— Examples

Preprocessing

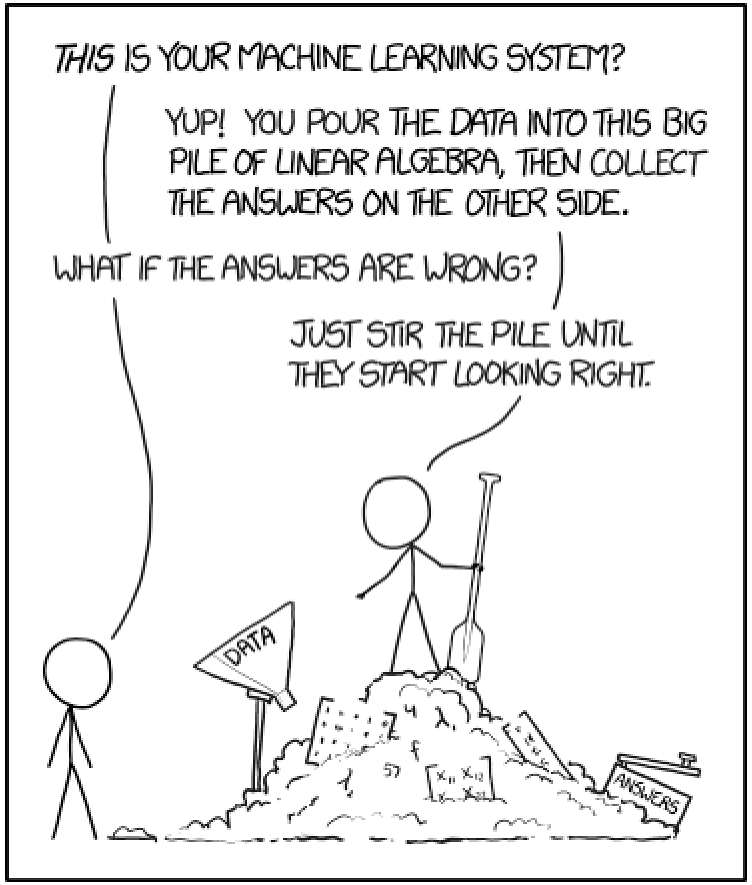
Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.
Modules: preprocessing, feature extraction.

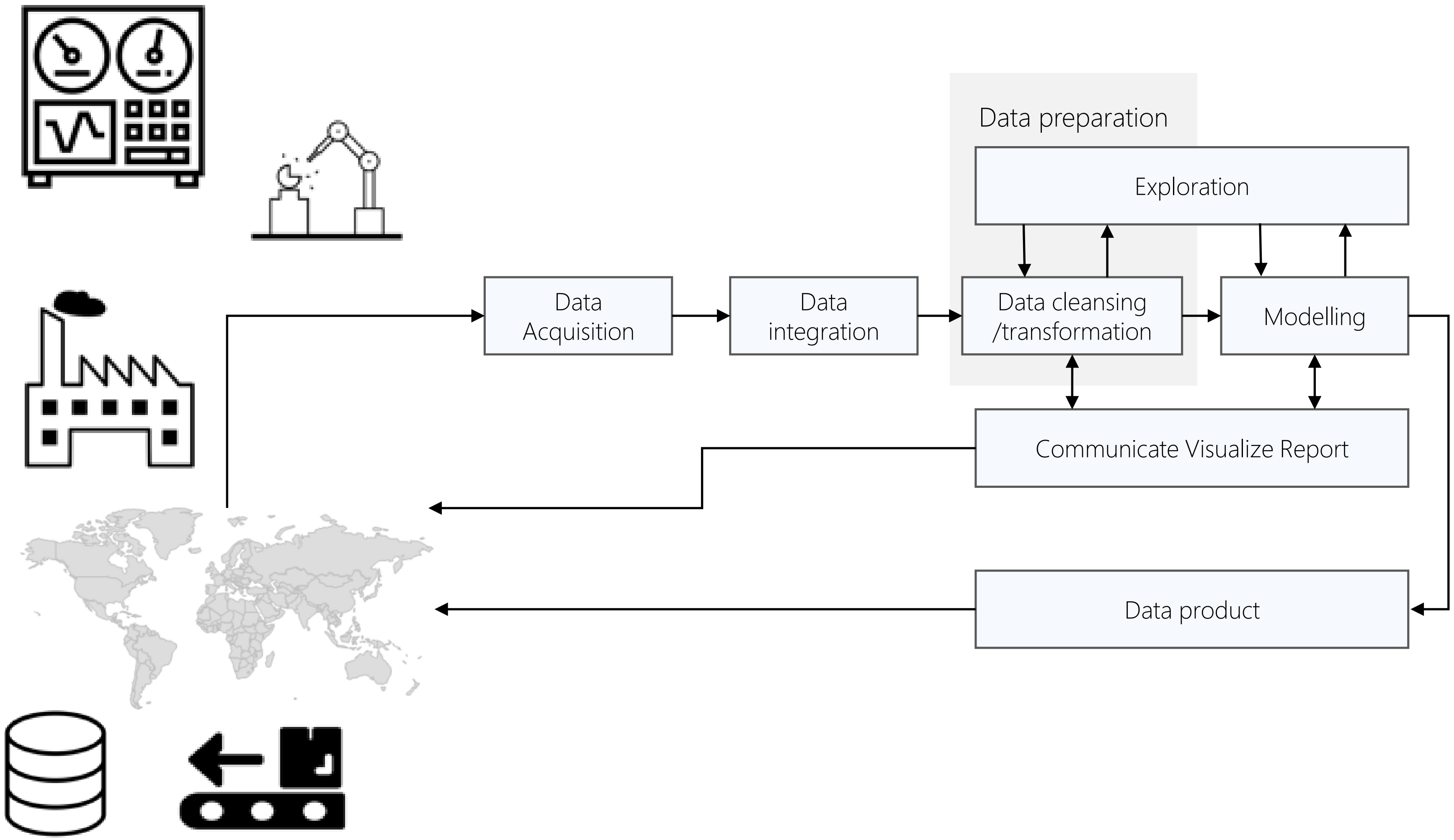
— Examples

Data Preparation: Introduction

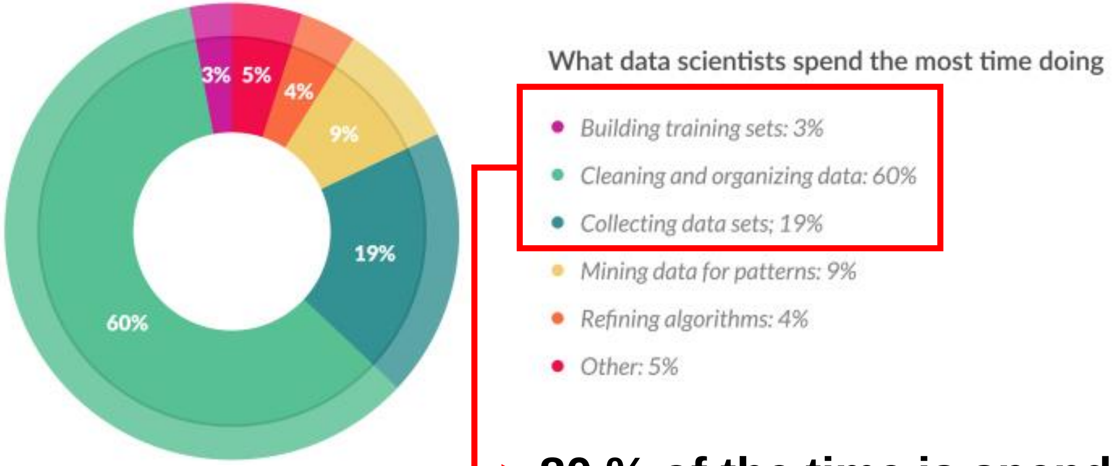
Overview: Randomly changing things until they work – or is there a better approach?



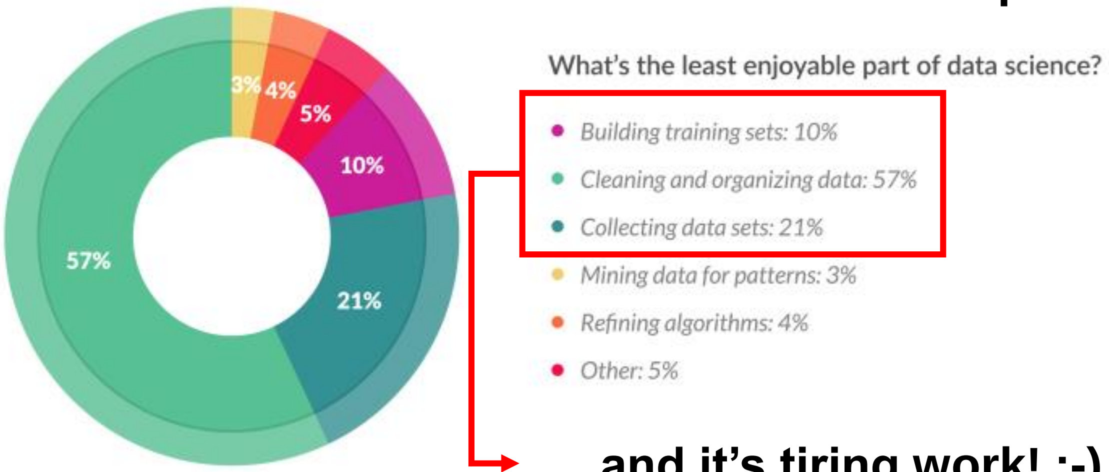
Overview: From Data Acquisition to the Data Product



Data Analytics Tasks



80 % of the time is spend on data preparation!



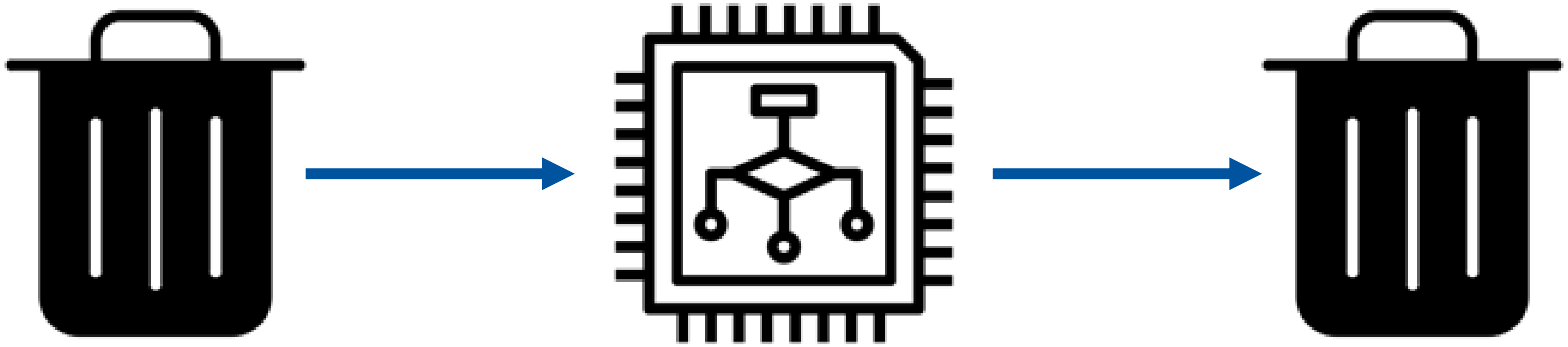
… and it's tiring work! ;-)

Source: https://whatsthebigdata.com/2016/05/01/data-scientists-spend-most-of-their-time-cleaning-data/

What is Data Preparation and why is it important?

Working Definition

Data preparation (data preprocessing) is the process of modifying raw data into a state suitable for analysis (e.g. by removing outliers).



“Garbage in, garbage out” **– real world data is messy. Sometimes values are missing, sometimes it contains errors, sometimes it’s inconsistent etc.**

What tasks does Data Preparation involve?

Raw Data

Data Preparation

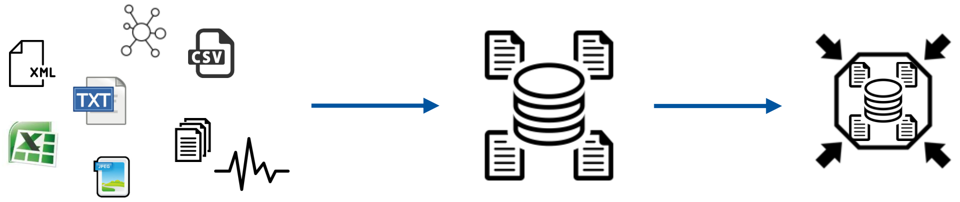
Analysis

Data Cleaning

Scope of this Lecture!

Data Transformation

Data Integration and Data Reduction



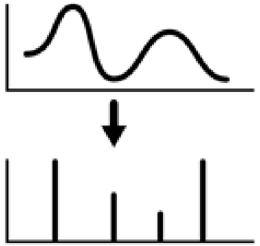
What tasks does Data Preparation involve?

Data Cleaning

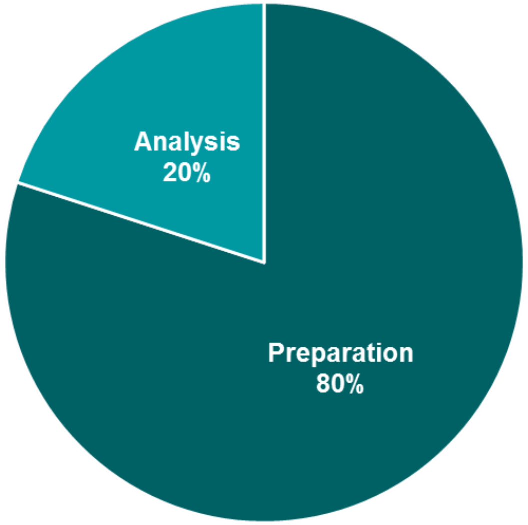
- Deal with missing values
- Identify/remove outliers
- Resolve inconsistencies
- Deal with noisy data

Data Transformation

- Normalization
- Aggregation
- Discretization



Data Integration and Data Reduction



Example for Real World Industry Data (1/2)

BOSCH kaggle

Id, L0_S0_F0, L0_S0_F2, L0_S0_F4, L0_S0_F6, L0_S0_F8, L0_S0_F10, L0_S0_F12, L0_S0_F14, L0_S0_F16, …, …, …, L3_S50_F4241, L3_S50_F4243, L3_S50_F4245, L3_S50_F4247, L3_S50_F4249, L3_S50_F4251, L3_S50_F4253, L3_S51_F4256, L3_S51_F4258, L3_S51_F4260, L3_S51_F4262, Response

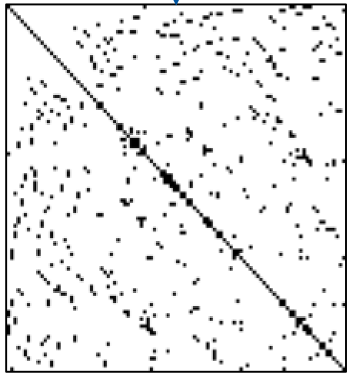
15 3.5 6.9 1.2 17.23 … 1

⋮ Dimensions: 1.183.748 x 970 ⋮

23 9.2 8.4 3.8 99.33

99 3.8 3.2 6.3 24.19 … 0

Essentially, the Bosch Data is one big matrix … and it's super sparse!



Missing values are a common problem in real world data!

IfU IMA RWTH AACHEN UNIVERSITY

Example for Real World Industry Data (2/2)

Wheel damage data from inspection and maintenance reports:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	axleLoadT	axleNo	axles	base_	base	base_	base	bumpL	bumpR	comp.	defe	defec	dxM	dynBwL	dynBwR	endAxle	flatL	flatR	gpsLength	gpsWidth
2	22.2	112	2					null	null	0			9.96	1.13	1.11	112	null	null	12.0509	52.5907
3	22.1	111	2					null	null	0			3.94	1.09	1.14	112	null	null	12.0509	52.5907
4	21.9	110	2					null	null	0			9.96	1.06	1.07	110	null	null	12.0509	52.5907
5	22.3	109	2					null	null	0			4.56	1.16	1.08	110	null	null	12.0509	52.5907
6	22.1	108	2					null	null	0			9.96	1.06	1.08	108	null	null	12.0509	52.5907
7	21.8	107	2					null	null	0			3.94	1.08	1.09	108	null	null	12.0509	52.5907
8	21.5	106	2					null	null	0			9.96	1.03	1.11	106	null	null	12.0509	52.5907
9	22.2	105	2					null	null	0			4.56	1.11	1.12	106	null	null	12.0509	52.5907
10	21.8	104	2					null	null	0			9.96	1.07	1.15	104	null	null	12.0509	52.5907
11	21.8	103	2					null	null	0			3.94	1.06	1.07	104	null	null	12.0509	52.5907
12	21.7	102	2					null	null	0			9.97	1.04	1.14	102	null	null	12.0509	52.5907
13	22.4	101	2					null	null	0			3.91	1.1	1.13	102	null	null	12.0509	52.5907
14	19	100	4					null	null	0			1.81	1.13	1.12	100	null	null	12.0509	52.5907
15	19.3	99	4					null	null	0			12.98	1.16	1.18	100	null	null	12.0509	52.5907
16	18.8	98	4					null	null	0			1.81	1.16	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
17	18.6	97	4					null	null	0	FS	LFS	3.23	1.39						
18	18.6	96	4					null	null	0			1.8	1.16						
19	19.2	95	4					null	null	0			12.99	1.16						

Datasets generated from data aggregation systems can have a significant amount of empty data. Said datasets can also have wrongly formatted fields or mixed units in a single column.

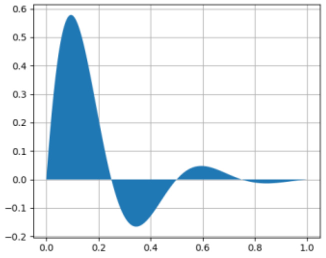
Data Exploration

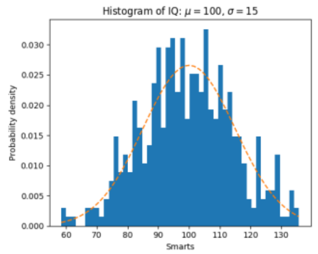
What is Data Exploration?

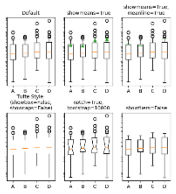
Working Definition

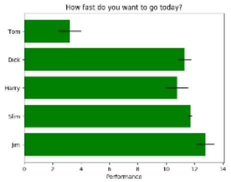
Data exploration is the process of creating an initial understanding of the properties (e.g. distribution or characteristics) of the data at hand.

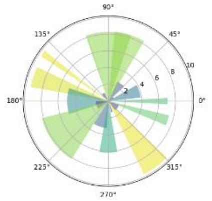
- Size
- Types
- Modalities
- Formatting
- Completeness
- Relationships
- Ownership
- …











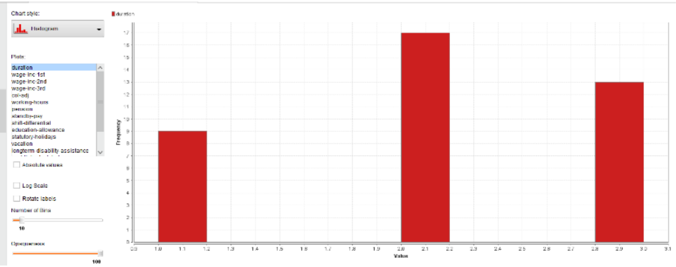
Points of Interest | **Visualization/Statistics**

Initial Data → **Data Exploration**

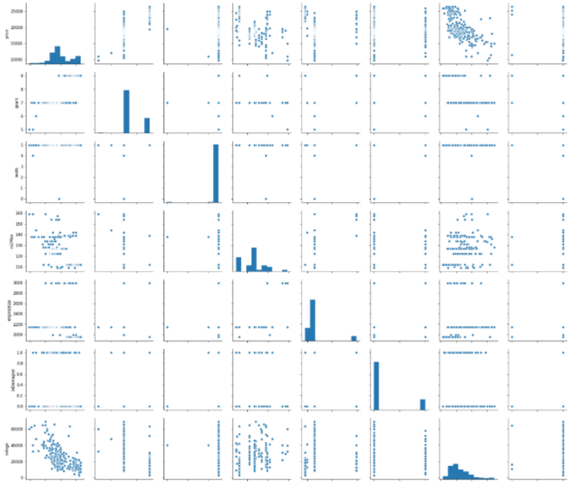
Tools and Methods for Data Exploration

Approaches to Data Exploration

There are plenty of commercial (e.g. Tableau, Rapidminer, data iku) or open source tools (e.g. DIVE) for data exploration. You can use these tools or build your own stack/process for the data at hand.

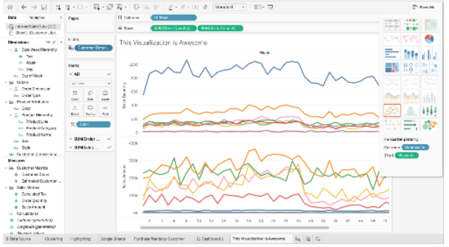


rapidminer

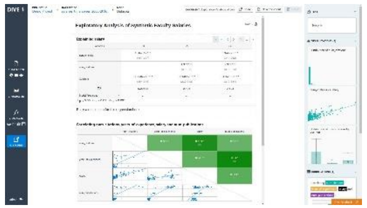


SQL

R



tableau



MIT DIVE

Python: Matplotlib/Pandas/Numpy

Tool-driven Methods | **Scripting Methods**

Methods can be automatic (e.g. identifying outliers) or manual or both

Types of “Data Challenges” and (Preparation) Tools

Cleaning: What is Noisy Data?

Working Definition

Data is noisy, if it contains attributes or values which can potentially harm the understanding or the analysis of it. That is, noisy data has to be removed before the analysis task.

Real world data is (always) noisy!

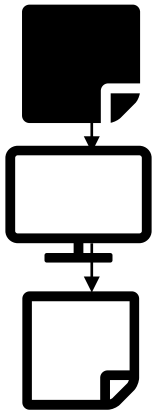
Causes

- Defect sensors
- Improper placement of sensors
- Systematic errors in data collection
- Manual errors
- Data from different sources
- Programming errors
- Incorrect measurements

Cleaning deals with noisy data!

Cleaning: Automatic Vs Manual Processes

100% automatic



Doesn't exist….
Data cleaning is highly dependant on context and problem

Assistance tools

Tools that provide interfaces, common transforms and algorithms for assisted data cleaning.
Has issues with domain or use case specific dirt.

Programming

There is more control over the data cleaning process.
Multiple libraries exist to enable more specialized cleaning. Ex:
Dedupe (de duplicate),
fuzzywuzzy (phonetics),
arrow (dates),
scrubadub (privacy).

100% Manual

Worst case scenarios.
Ex: format has been compromised and data can't be read by other tools.

Cleaning: Missing Values

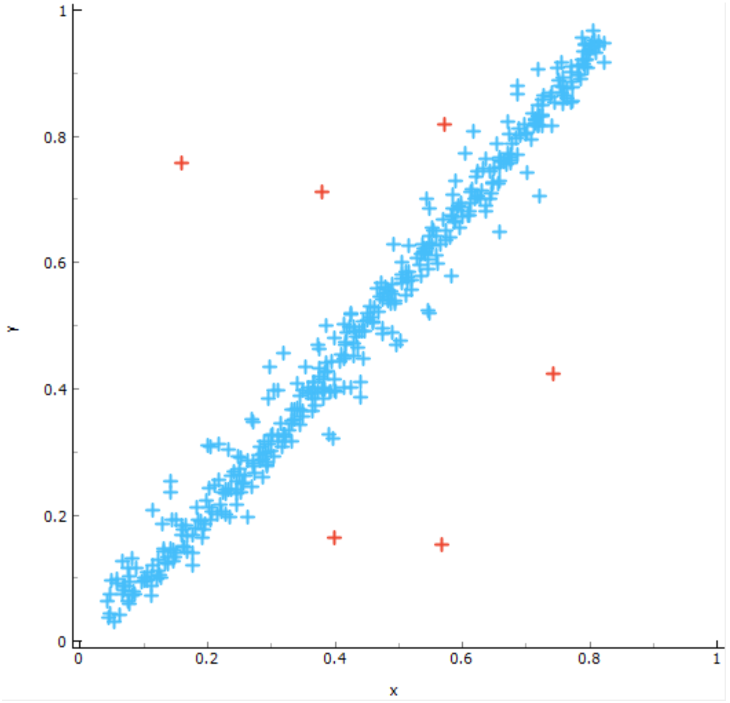
The nature of missing data can be divided in:
Missing Completely at Random (MCAR): not related to the missing value or the other values.
Missing at Random (MAR): not related to the missing data, it is related to some of the observed data.
Missing not at Random (MNAR): missing because of the hypothetical value or dependent on some other variable.

- **Missing data handling**
  - **Imputation**
    - Time-Series Problem
      - No trend or Seasonality: Mean, Median, Mode, Random, Sample
      - Trend and Seasonality: Interpolation + Seasonal adjustment
    - General Problem
      - Categorical: Make NA a level, Multiple imputation, regression.
      - Continuous: Mean, Median, Mode, Multiple Imputation, regression.
  - **Deletion**
    - Deleting rows (listwise)
    - Pairwise Deletion
    - Deleting Columns

Cleaning: Outliers

Working Definition

An **outlier** is an observation that lies an abnormal distance from other values in a random sample from a population.



Outlier modelling

- Visual exploration.
- Statistical tests.
- Modelling (linear model, isolation forest, Robust Covariance, One Class SVM, Local Outlier Factor).
- Projection exploration.

Cleaning: Inconsistencies

Working Definition

Data is inconsistent, if the data attributes don't match their values (and vice versa) or if the data values change "midway".

Semantic of data attribute and value don't match. Hard for tools to automatically detect! Manual approach necessary.

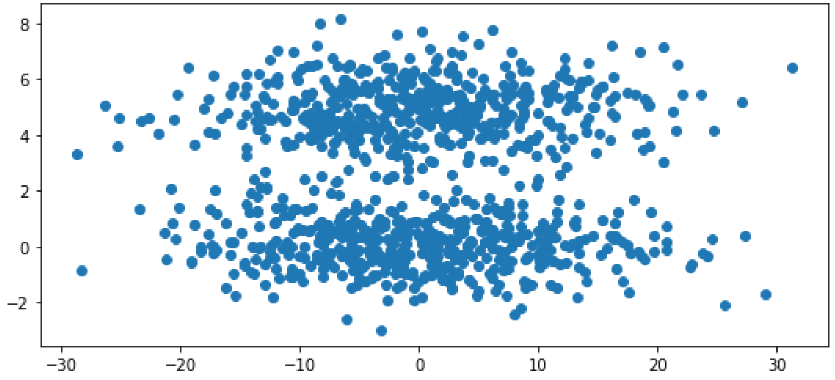
Colour	Quality
ABB	Good
Fanuc	Poor
Kuka	2
ABB	5
Denso	Good

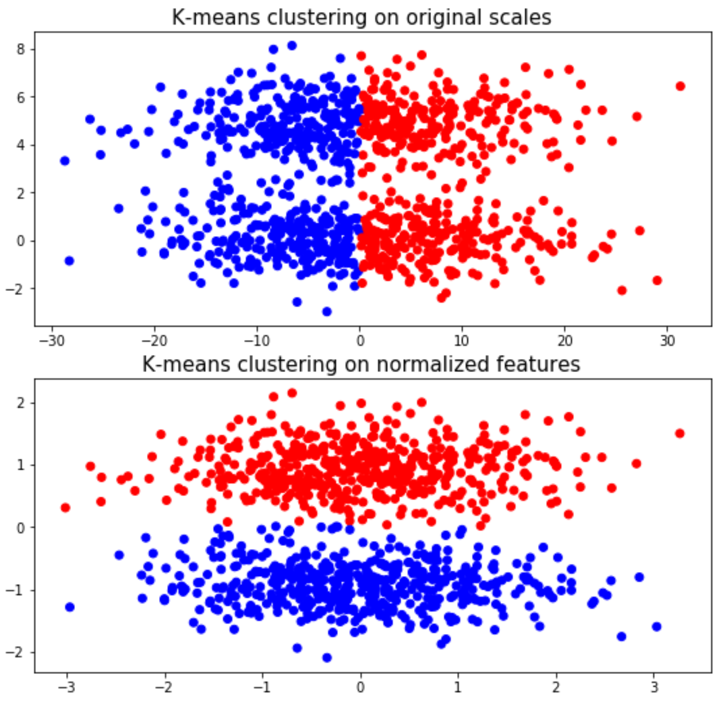
Data values are inconsistent (e.g. Low versus 5). Can be detected automatically, but matching has to be derived manually (e.g. is 2 good or poor?)

Transformation: Normalization

Working Definition

Normalization is the task of changing the values of numeric columns to a common scale, without distorting differences in the ranges of values.



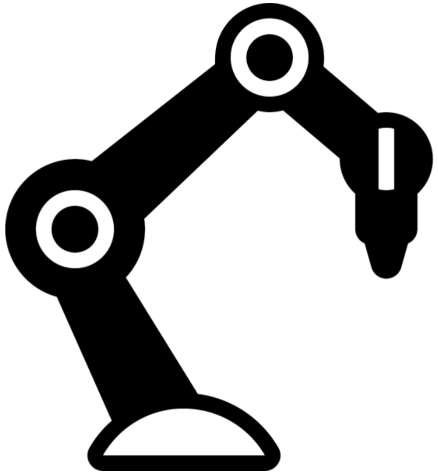


Normalization reduces Knock-on effects on the learning ability of algorithms (depending on the algorithm). Ensuring standardized features, implicitly weights all features equally in their representation.

Transformation: Aggregation

Working Definition

Data Aggregation is the process of aggregating a minimum of two attributes into one (e.g. two data columns into one). It can either be done automatically (e.g. correlation detection) or manual.



Bought	Defects
04/2019	3
01/2010	5
03/1998	9
08/2018	4
07/2005	3

Reliability
Low
High
High
Low
High

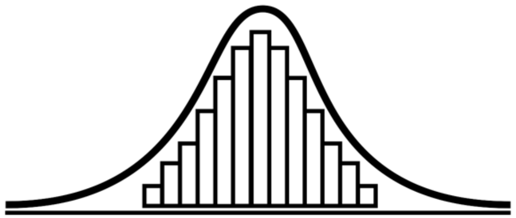
Data Aggregation reduces the variability of your data. It operates on attributes, not values (as opposed to Discretization, see next slide).

Transformation: Discretization

Working Definition

Data Discretization is defined as a process of converting continuous data attribute values into a finite set of intervals and associating with each interval some specific data value. [Jin, Breitbart et al., 2009]

Machine Age Attribute



[1,2,3,4,5,6,10,12,18,20,23,25]

"New" "Mid" "Old"

Example

Number of discrete states

↓

Map values to states

General Process

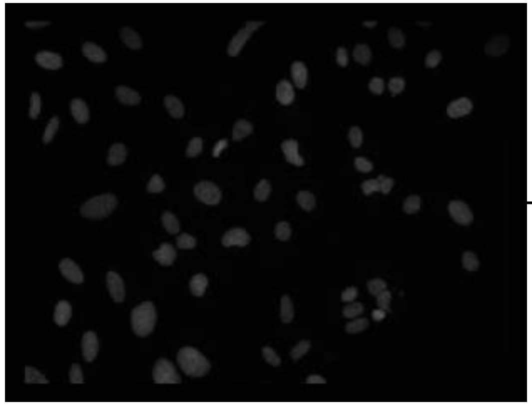
- Can be supervised or unsupervised
- Binning
- Histogram Analysis
- Clustering Analysis
- Decision-tree Analysis
- Correlation

Methods

Data Preparation w.r.t. to various Modalities

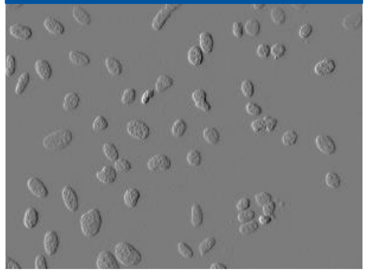
Data Preparation w.r.t. Image Data

Acquiring

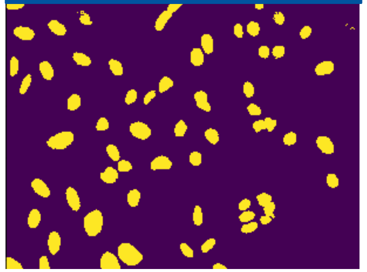


Prepare

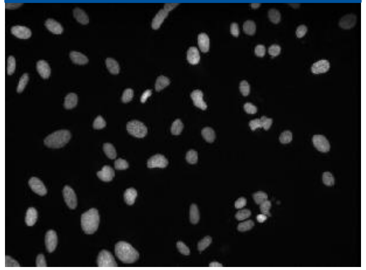
Transform



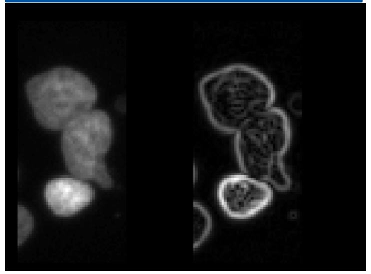
Segment



Equalize



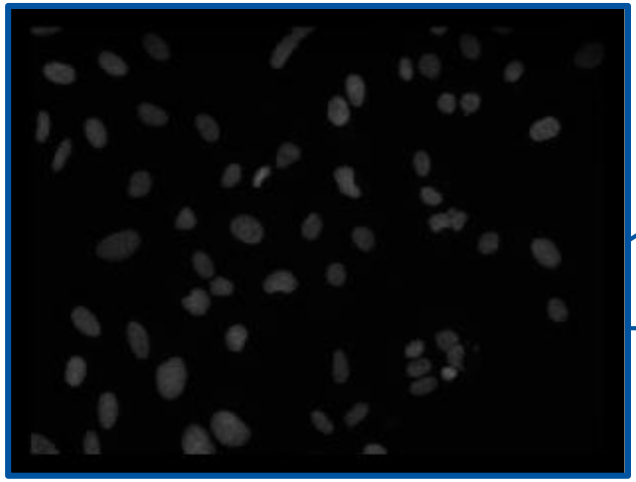
Augment

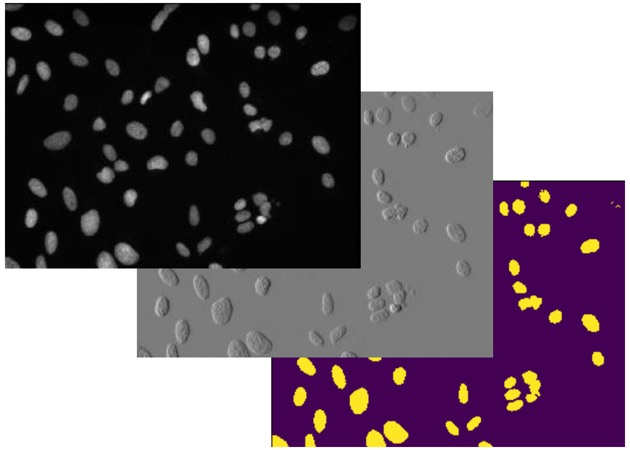


Rule of thumb: If you need to acquire image data, take into account light, noise, background and scale before compromising.

Next steps, e.g. Analysis

Data Preparation w.r.t. Image Data





Problem specific

Each transform that is done can generate loss of information relevant to the problem.

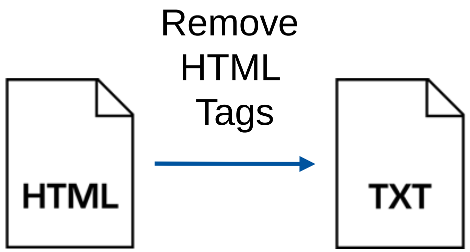
Most common operations

- Resize.
- Denoise.
- Thresholding.
- Light correction.
- Segmentation.
- Morphology.
- Perspective correction.

Data Preparation w.r.t. Textual Data (Example)

Acquiring data from webpages, e.g. Wikipedia

```
<ul><li><i><a href="/wiki/Cross-
validation_(statistics)" title="Cross-validation
(statistics)">Cross-validation</a></i>. By splitting
the data into multiple parts, we can check if an
analysis (like a fitted model) based on one part of
the data generalizes to another part of the data as
well. Cross-validation is generally inappropriate,
though, if there are correlations within the data, e.g.
with <a href="/wiki/Panel_data" title="Panel
data">panel data</a>. Hence other methods of
validation sometimes need to be used. For more on
this topic, see <a
href="/wiki/Statistical_model_validation"
title="Statistical model validation">statistical model
validation</a>.</li>
<li><i><a href="/wiki/Sensitivity_analysis"
title="Sensitivity analysis">Sensitivity
analysis</a></i>. A procedure to study the behavior
of a system or model when global parameters are
(systematically) varied. One way to do that is via <a
href="/wiki/Bootstrapping_(statistics)"
title="Bootstrapping
(statistics)">bootstrapping</a>.</li></ul>
```



Cross-validation. By splitting the data into multiple parts, we can check if an analysis (like a fitted model) based on one part of the data generalizes to another part of the data as well. Cross-validation is generally inappropriate, though, if there are correlations within the data, e.g. with panel data. Hence other methods of validation sometimes need to be used. For more on this topic, see statistical model validation. Sensitivity analysis. A procedure to study the behavior of a system or model when global parameters are (systematically) varied. One way to do that is via bootstrapping.

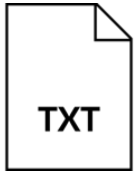
Next steps, e.g. Analysis

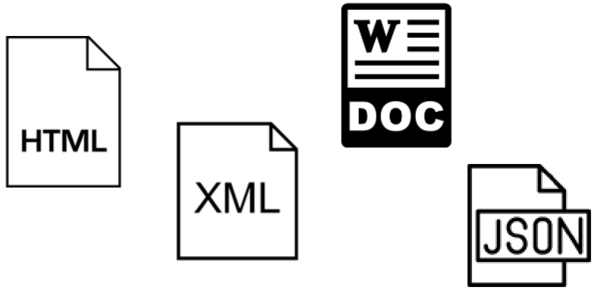
API

Rule of thumb: If you need to acquire data from the web don't try your own crawler. Better use a services' API for cleaner data (e.g. already structured data, w/o tags etc .)

Data Preparation w.r.t. Textual Data (Different Tasks)

Cross-validation. By splitting the data into multiple parts, we can check if an analysis (like a fitted model) based on one part of the data generalizes to another part of the data as well. Cross-validation is generally inappropriate, though, if there are correlations within the data, e.g. with panel data. Hence other methods of validation sometimes need to be used. For more on this topic, see statistical model validation. Sensitivity analysis.





Tokenization (Segmentation)

Task of splitting text (as one large string) into sentences, words etc., e.g. [“By”, “splitting”, “the”, “data”, “into”, “multiple”]

Normalization

Task of converting text to same case (upper/lower) remove punctuation, convert words (“one”) to their number representations (“1”) etc.

Noise Removal

Task of removing headers, footers, tags, various metadata etc.

Data Preparation w.r.t. Textual Data (Sensitive Data)

Given your task is to prepare textual data from a customer relationship management system and to remove all sensitive information. How to proceed?

*"Hi, my name is **Julius Caesar**. I'm living in **Park Street 204, New York** and I want to change my credit card number from **1432 4004 2391 2341** to **6372 9932 2834 1834**. For verification my birth date is **July 23, 1956**. Can you help me?"*

TXT

Identification

Task of identification of sensitive information within the text, e.g. using regular expressions, blacklist, whitelists etc.

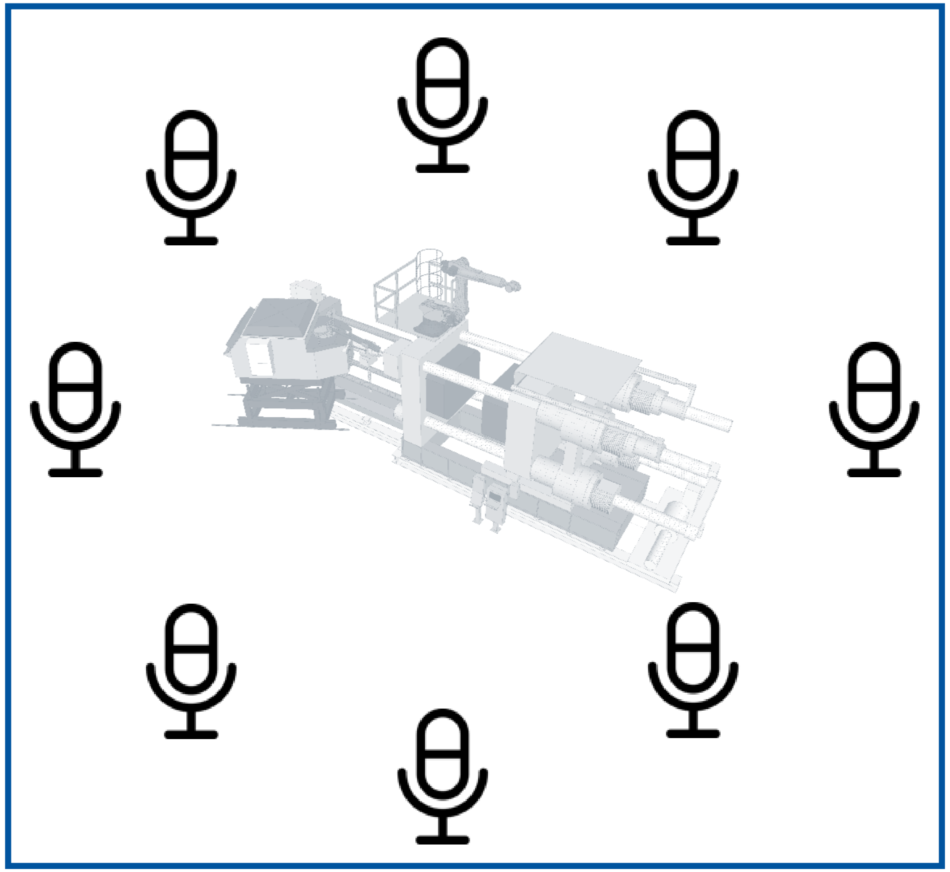
Anonymization/Pseudonymization

Task of deleting or transforming all sensitive information into insensitive pieces of information ("Julius Caesar" to "John Smith").

Relationship to Engineering is Important!

Machine data has sometimes to be anonymized for analytics. For example the Bosch Kaggle data was pseudonymized w.r.t. to machine labels to prevent competitors from gaining insights into Bosch production (PS: Bosch failed).

Data Preparation w.r.t. Audio Data



Example: High-pressure die casting process with audio sensors.

Normalization

Normalize different sample rates, quantization levels, sound amplitudes etc.

Cleaning

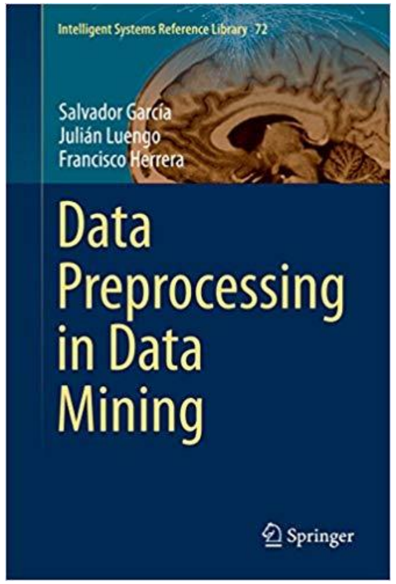
Remove background noises, remove silence intervals, inference from mobile phone usages etc.

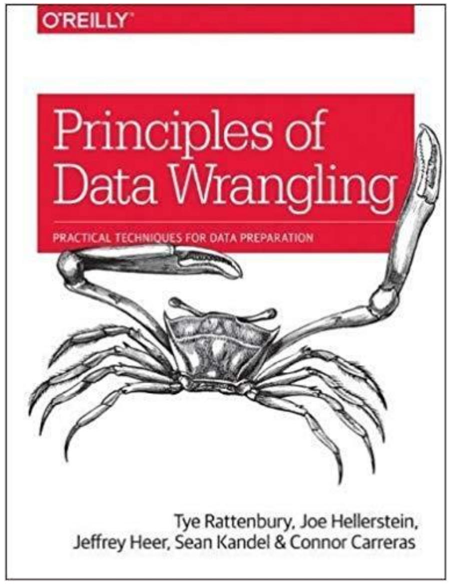
pyAudio Analysis

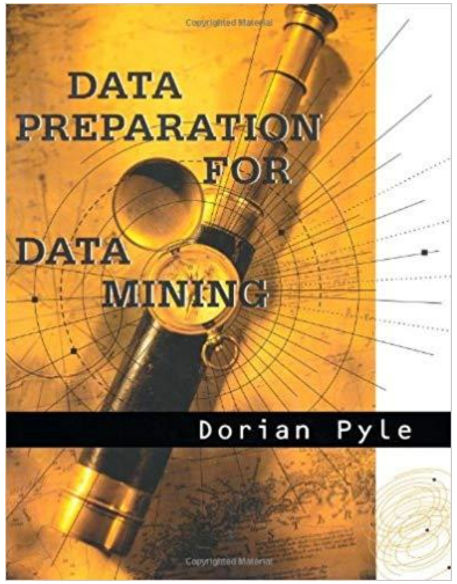
Great Python tool for audio data preparation (and analysis)!

Further Reading Material

- https://scikit-learn.org/stable/modules/preprocessing.html
- https://www.coursera.org/lecture/big-data-machine-learning/data-preparation-XMoi8
- https://www.kdnuggets.com/2018/12/six-steps-master-machine-learning-data-preparation.html
- http://www.jstatsoft.org/article/view/v059i10/v59i10.pdf (Tidy Data)
- https://www.fosteropenscience.eu/sites/default/files/pdf/2933.pdf (Data Exploration)







Thank you for your attention!

Lecture Team AIDAE
aidae@ima-ifu.rwth-aachen.de