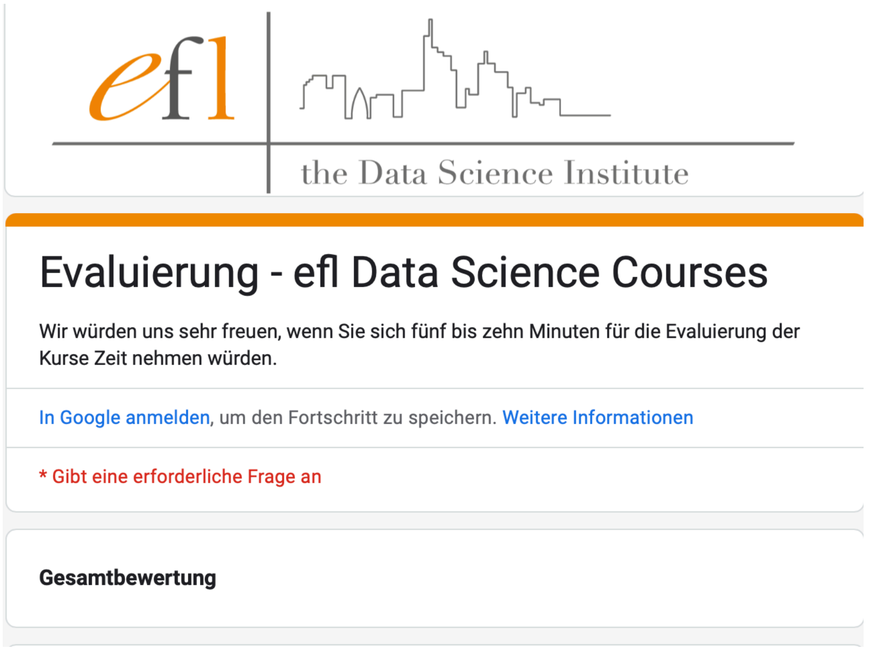
Data Science - Working with Data
Pre-processing, Explorative Data Analysis

Survey



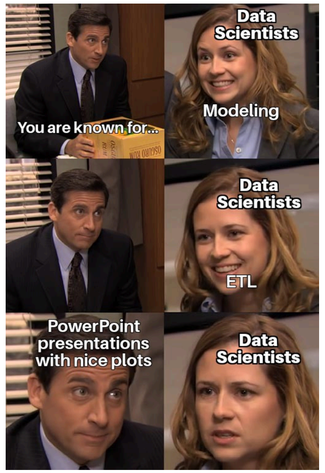
https://docs.google.com/forms/d/e/1FAIpQLSdH4RcSBN20rMX3bQjzfIt1JX_RjTijpiRebA_c062j83RFig/viewform

- Data Science Approach
- Case Study

But why?
Computing Power (Moore‘s Law), e.g., Lundstrom (2003)
Economies of Scale (Wrights’s Law), e.g., Stringham et al. (2015)
Non-linear (exponential) increase in knowledge (Kurzweil‘s Law), e.g., Cyranoski et al. (2011)

Lundstrom, M. (2003). Moore's Law Forever? Science, 299(5604), 210–211.
Cyranoski, D., Gilbert, N., Ledford, H., Nayar, A., & Yahia, M. (2011). Education: The PhD Factory. Nature, 472(1), 276–279.
Stringham, E. P., Miller, J. K., & Clark, J. (2015). Overcoming Barriers to Entry in an Established Industry: Tesla Motors. California Management Review, 57(4), 85–103.

Data Science in Practice



Pre-Processing: What is that?

- Data preprocessing is a major and essential stage whose main goal is to obtain final data sets which can be considered correct and useful for further data mining algorithms [1].

- The idea is to filter the data according to quality criteria, selected according to possible target values, and transformed correspondingly [2] .

[1] García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. Big Data Analytics, 1(1), 1-22.
[2] Abdel-Karim, B. M., Pfeuffer, N., & Hinz, O. (2021). Machine learning in information systems-a bibliographic review and open research issues. Electronic Markets, 31, 643-670.

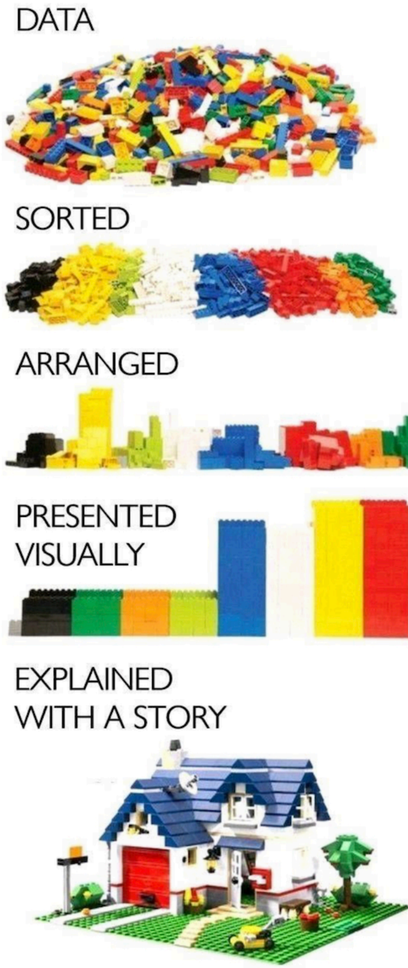
Pre-Processing: But why?

[1]

[2]

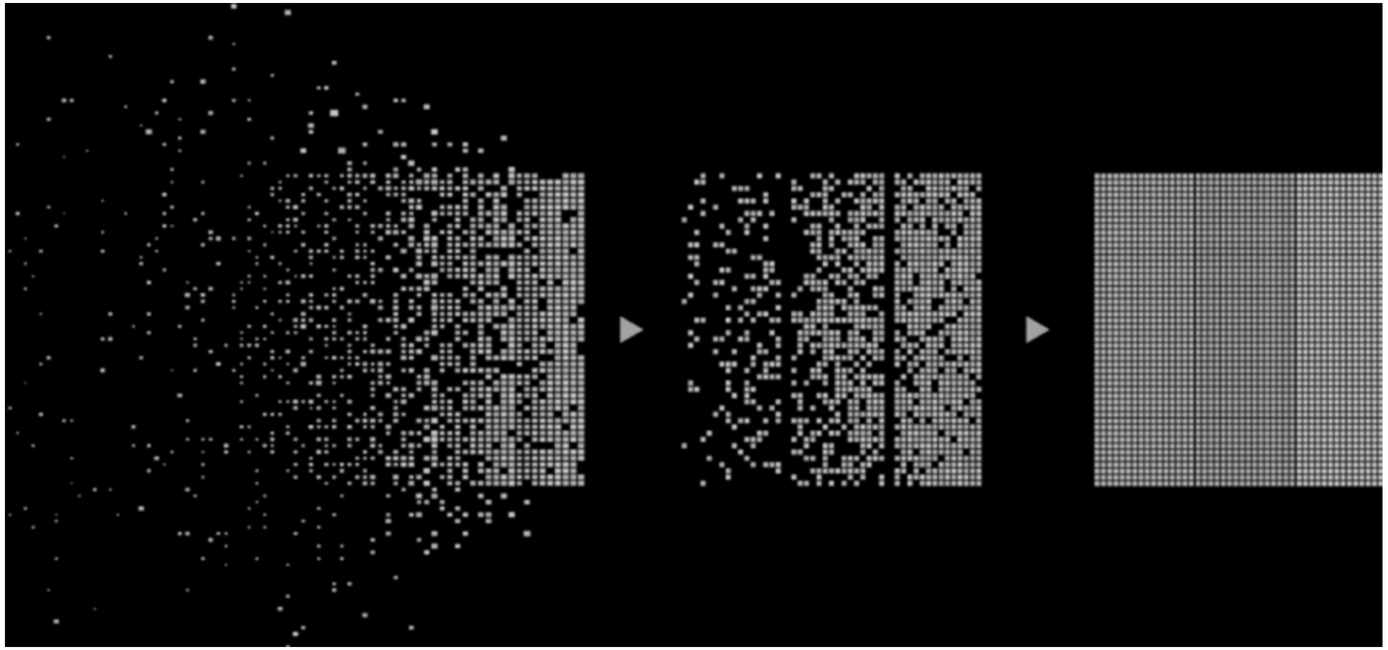
Raw Data

Knowlege



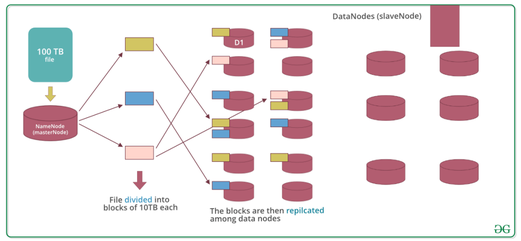
[1] https://www.google.com/url?sa=i&url=https%3A%2F%2Flustich.de%2Fbilder%2Fandere%2Fpapierberg%2F&psig=AOvVaw1shfkEZVN9kWASoXICbwAi&ust=1601110332932000&source=images&cd=vfe&ved=2ahUKEwiMxrL29oPsAhVUNewKHdqPD0MQr4kDegUIARDIAQ
[2] https://www.youtube.com/watch?v=n0ql-yeY9u0

Pre-Processing: How Should it Look?

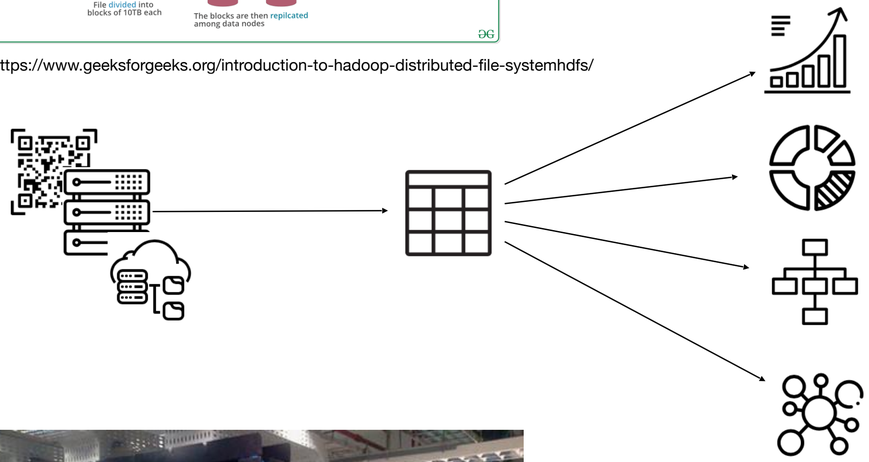


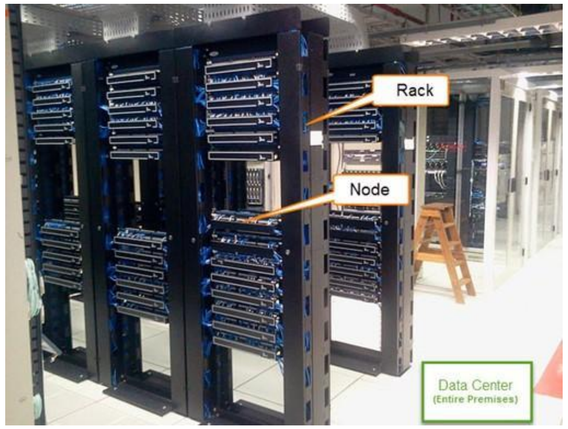
Raw Data Pre-Processing Clean Data

Pre-Processing: Challenge

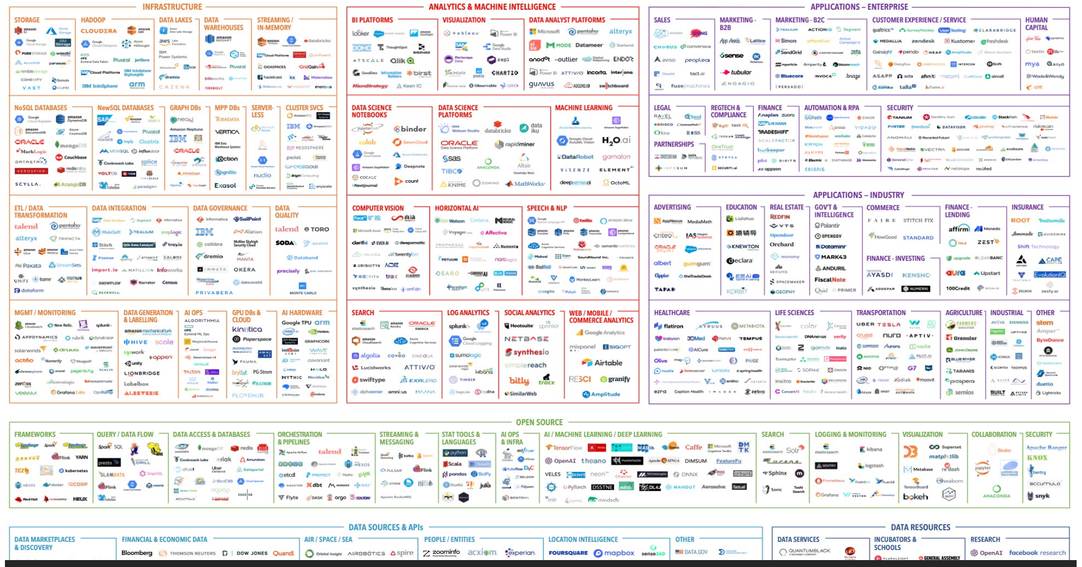


https://www.geeksforgeeks.org/introduction-to-hadoop-distributed-file-systemhdfs/





https://www.guru99.com/learn-hadoop-in-10-minutes.html



| 10

http://46eybw2v1nh52oe80d3bi91u-wpengine.netdna-ssl.com/wp-content/uploads/2020/09/2020-Data-and-AI-Landscape-Matt-Turck-at-FirstMark-v1.pdf

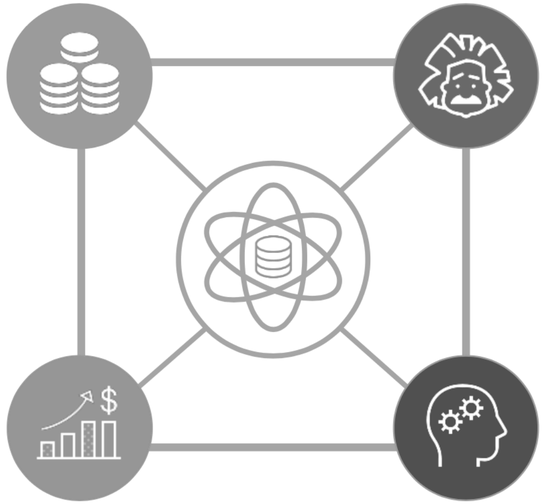
From Data to Machine Learning … Easy?

Database Management	Machine Learning
Database is an active, evolving entity	Database is just a static collection of data
Records may contain missing or erroneous information	Instances are usually complete and noise-free
A typical field is numeric	A typical feature is binary
A database typically contains millions of records	Data sets typically contain several hundred instances
AI should get down to reality	"Databases" is a solved problem and is therefore uninteresting

Table 2. Conflicting Viewpoints between Database Management and Machine Learning. [1]

Needs to shape the data into the right form

Data Engineer
- → Responsible for the acquisition, organization and storage of data
- → Possible skills: Software Engineering, SQL, Hadoop, Python(Pandas)



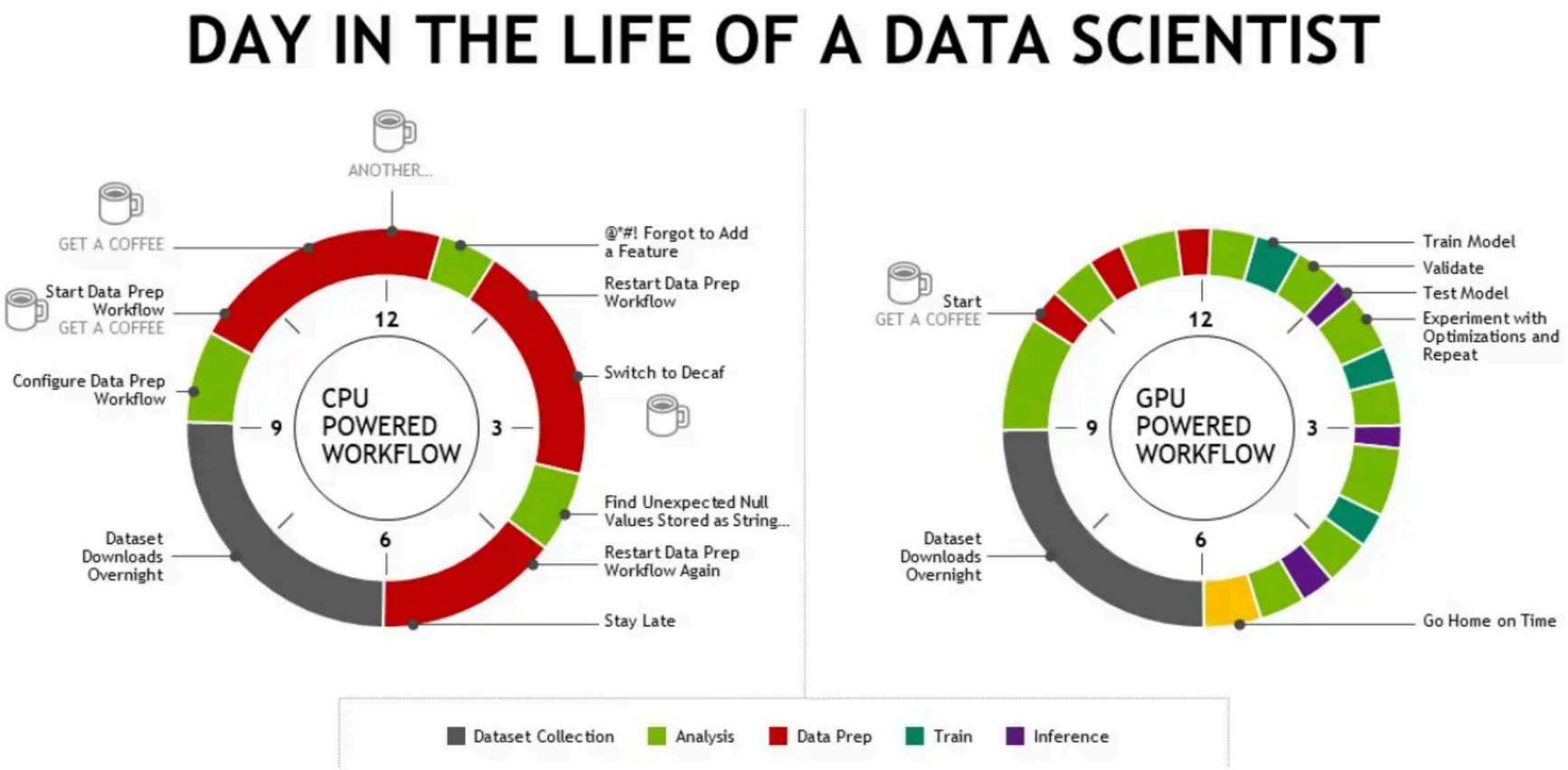
General Data Scientist
- → Coding Generalist who extracts knowledge from data and creates business value with the full data science stack
- → Possible Skills: Domain Knowledge, Statistics, Python/R, Pandas/Numpy, Machine Learning, Data Visualization, etc.

Data Analyst
- → Responsible for the creation of business value with data
- → Possible Skills: Domain Knowledge, Python/R, SQL, Excel, Data Visualization

Machine Learning Engineer
- → Specialist in the field of Machine Learning; Creator of Algorithms
- → Possible Skills: Mathematics/Statistics, Software Engineering, Python, C++, Pandas/Numpy/Pytorch, Machine Learning, Neural Networks, etc.

[1] Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases: An overview. AI magazine, 13(3), 57-57.

Pre-Processing: Critical Reflection



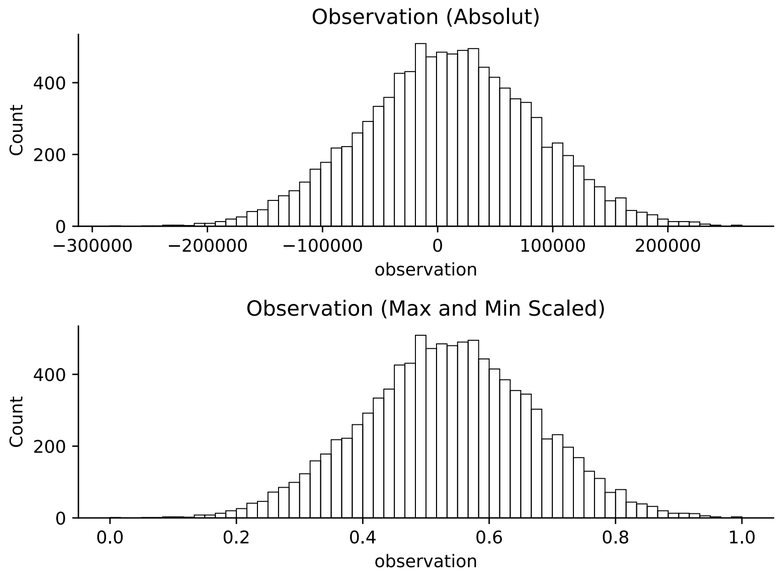
Therefore, it is important to realize that it is common practice to dedicate about 80% of the labor and time within the KDD process to data preparation, as a representative survey on a popular data science portal shows. https://bit.ly/2WwVPho.

In: Abdel-Karim, B. M., Pfeuffer, N., & Hinz, O. (2021). Machine learning in information systems-a bibliographic review and open research issues. Electronic Markets, 31(3), 643-670

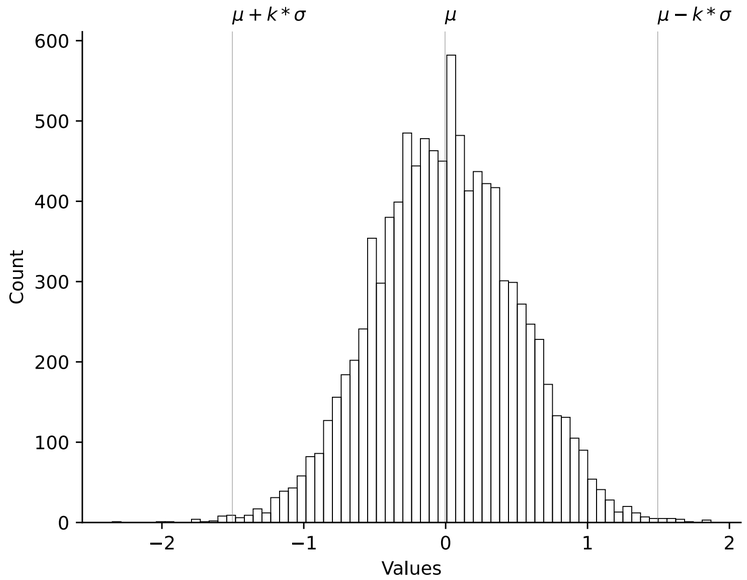
https://blogs.nvidia.com/blog/2018/11/15/accelerated-data-science-hpc/

Pre-Processing: Examples

- Handle Missing Values
  - Delete them
  - Replace them by
  - Constant Value e.g. 99999, 0
  - Calculate Value (mean, median etc.)
  - Interpolation by using a model etc.
  - Using Human Experts
- Min-Max Scaling
- Outlier Detection
- Create Datasets



Min-Max Scaled



Outlier Detection

Index	Quality
0	Strong
1	Medium
2	Low
3	Low
4	Medium

Index	Strong	Medium	Low
0	1	0	0
1	0	1	0
2	0	0	1
3	0	0	1
4	0	1	0

One-Hot-Encoding

- In **scaling**, you're changing the range of your data, while
- In **normalization**, you're changing the shape of the distribution of your data. https://www.kaggle.com/code/alexisbcook/scaling-and-normalization#

Domain Knowledge: Why this is Important?

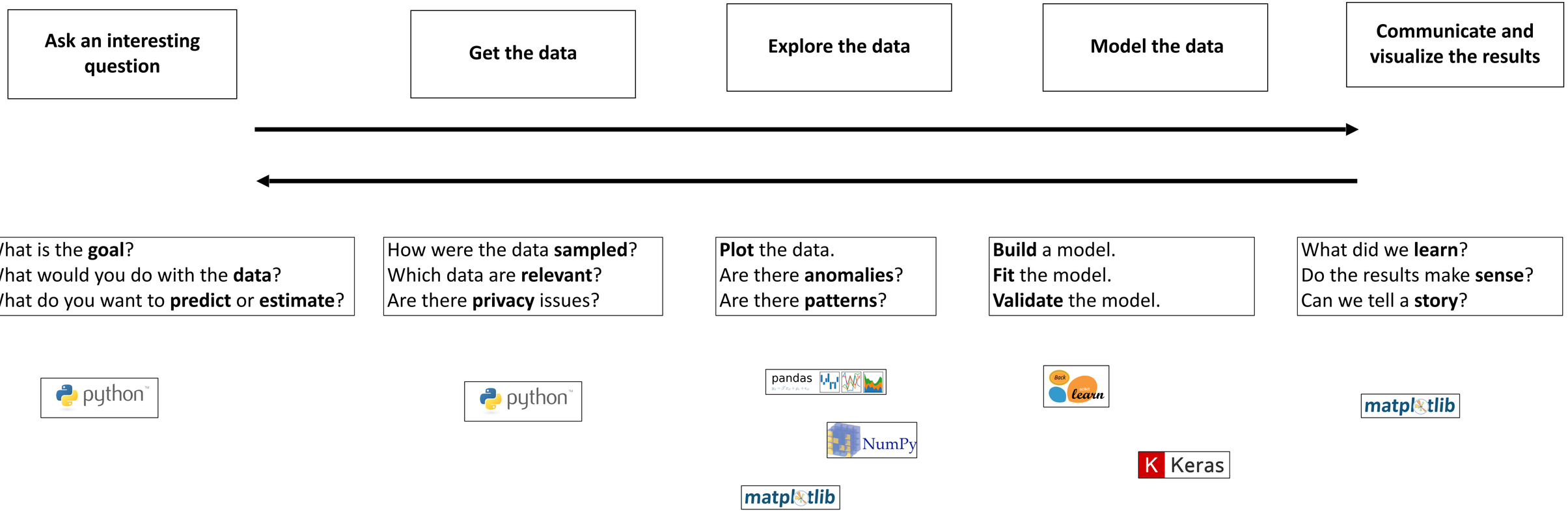
Domain knowledge can be define as knowledge about the environment in which the target implementation operates [1].

https://www.theforage.com/blog/careers/investment-banking

[1] Hjørland, B. & Albrechtsen, H. (1995). Toward A New Horizon in Information Science: Domain Analysis. Journal of the American Society for Information Science, 1995, 46(6), p. 400–425.

KDD in a Nutsehll

The KDD model expects the use of input data derived from a dataset acquisition pro cess (Lara et al. 2014, p. 54).

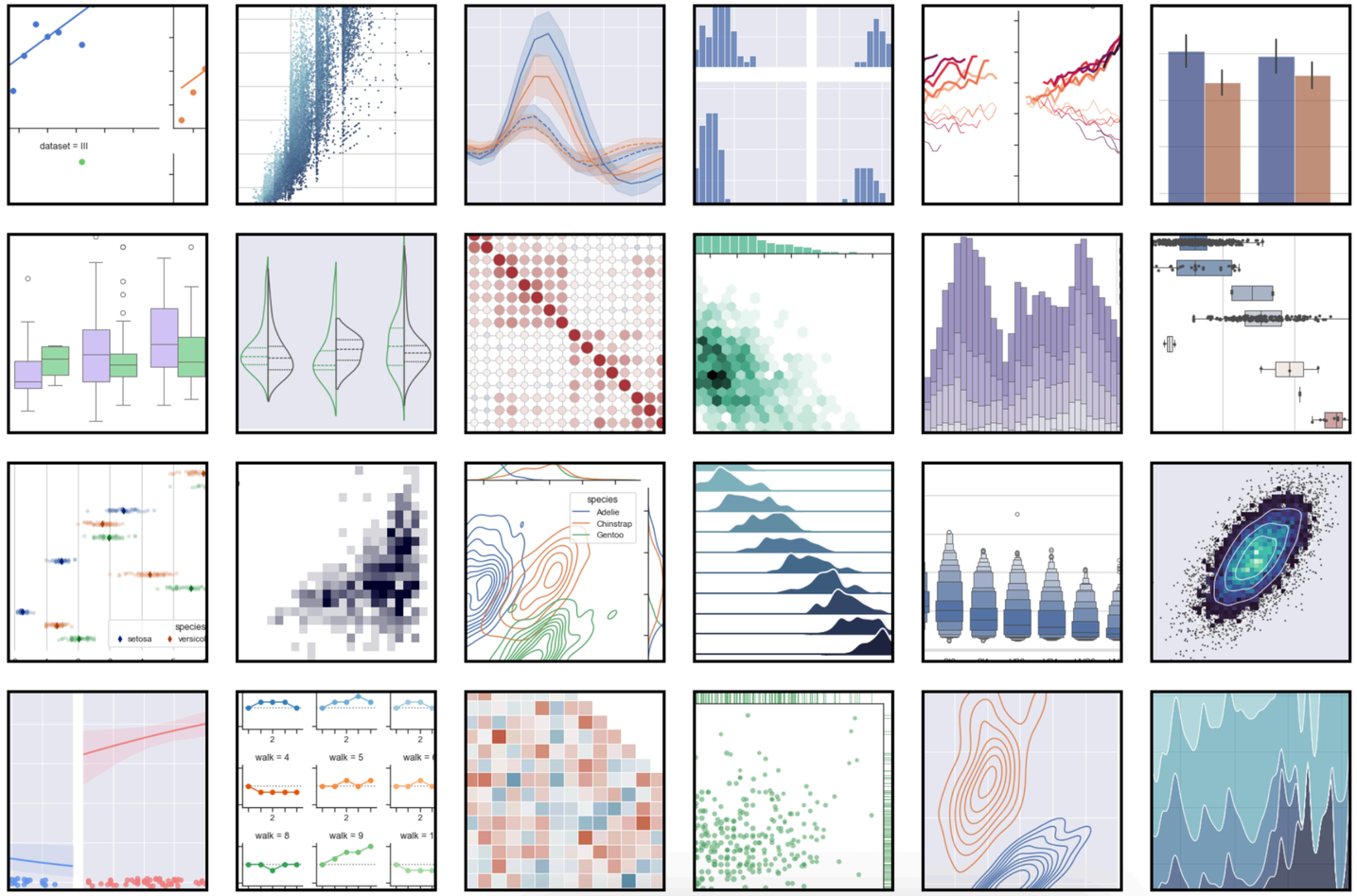


KDD Model: Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37-37.

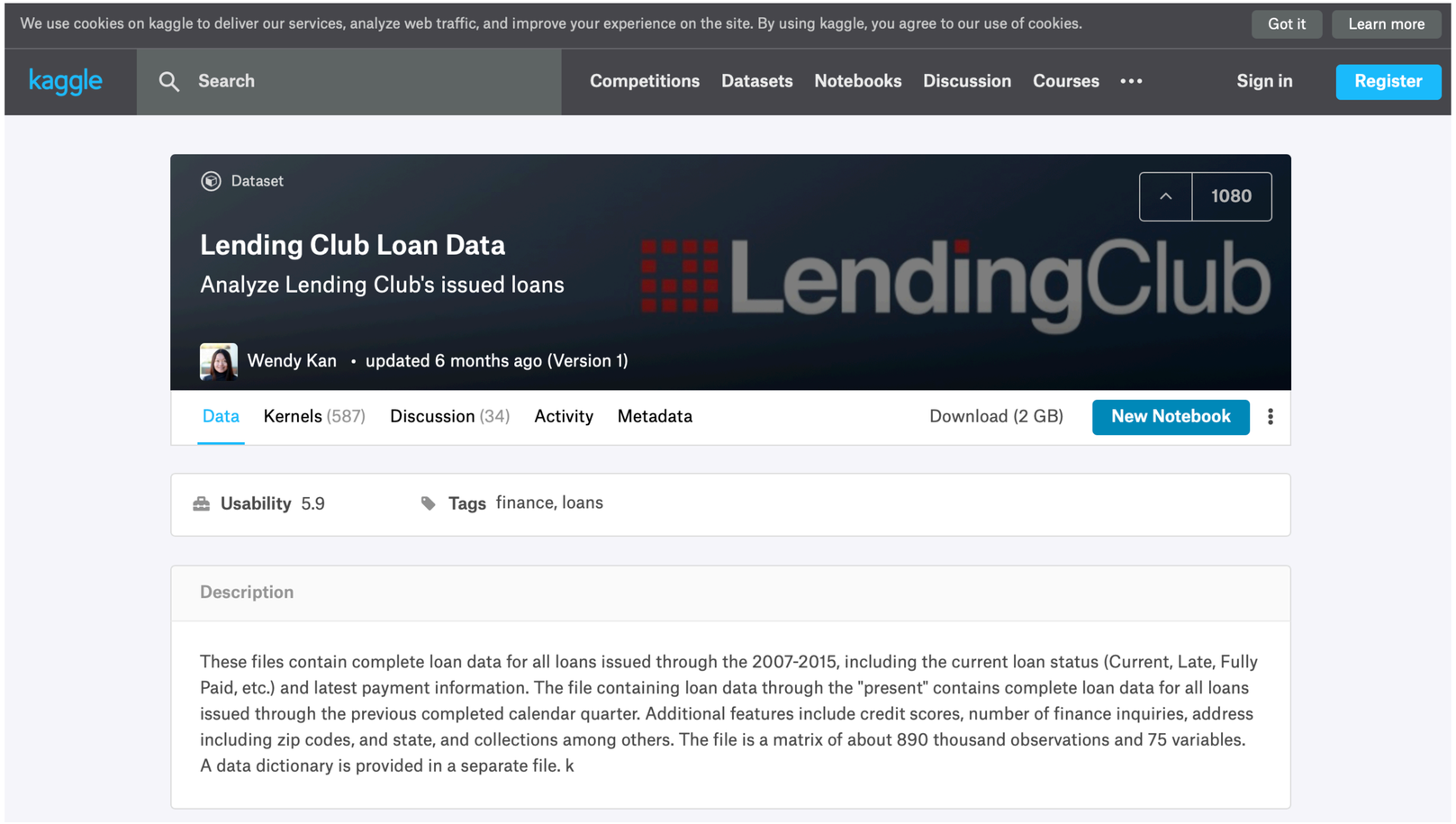
Lara, J.A., Lizcano, D., Mart´ınez, A., Pazos, J. (2014). Data preparation for KDD through automatic reasoning based on description logic. Information Systems, 44(8), 54–72.

Data Visualization: Seaborn

- Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. https://seaborn.pydata.org/
- An extension of the matplotlib is Seaborn.
- This library can operate directly with data frame objects. For data
science, a useful tool!
- It provides a high-level interface for drawing attractive and informative statistical graphics.
- Easy to use with DataFrames



Our Raw Data


We use cookies on kaggle to deliver our services, analyze web traffic, and improve your experience on the site. By using kaggle, you agree to our use of cookies.
Got it
Learn more
kaggle
Search
Competitions
Datasets
Notebooks
Discussion
Courses
Sign in
Register
Dataset
1080
Lending Club Loan Data
Analyze Lending Club's issued loans
LendingClub
Wendy Kan • updated 6 months ago (Version 1)
Data
Kernels (587)
Discussion (34)
Activity
Metadata
Download (2 GB)
New Notebook
Usability 5.9
Tags finance, loans
Description
These files contain complete loan data for all loans issued through the 2007-2015, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file containing loan data through the "present" contains complete loan data for all loans issued through the previous completed calendar quarter. Additional features include credit scores, number of finance inquiries, address including zip codes, and state, and collections among others. The file is a matrix of about 890 thousand observations and 75 variables. A data dictionary is provided in a separate file. k

Case Study

Context: Lending Club

Domain Knowledge: Case Study

A credit rating is an evaluation of the credit risk of a prospective debtor (an individual, a business, company or a government)

The credit rating represents an evaluation of a credit rating agency of the qualitative and quantitative information for the prospective debtor

Usually grouped in groups from A to F
- A = very good
- G = junk

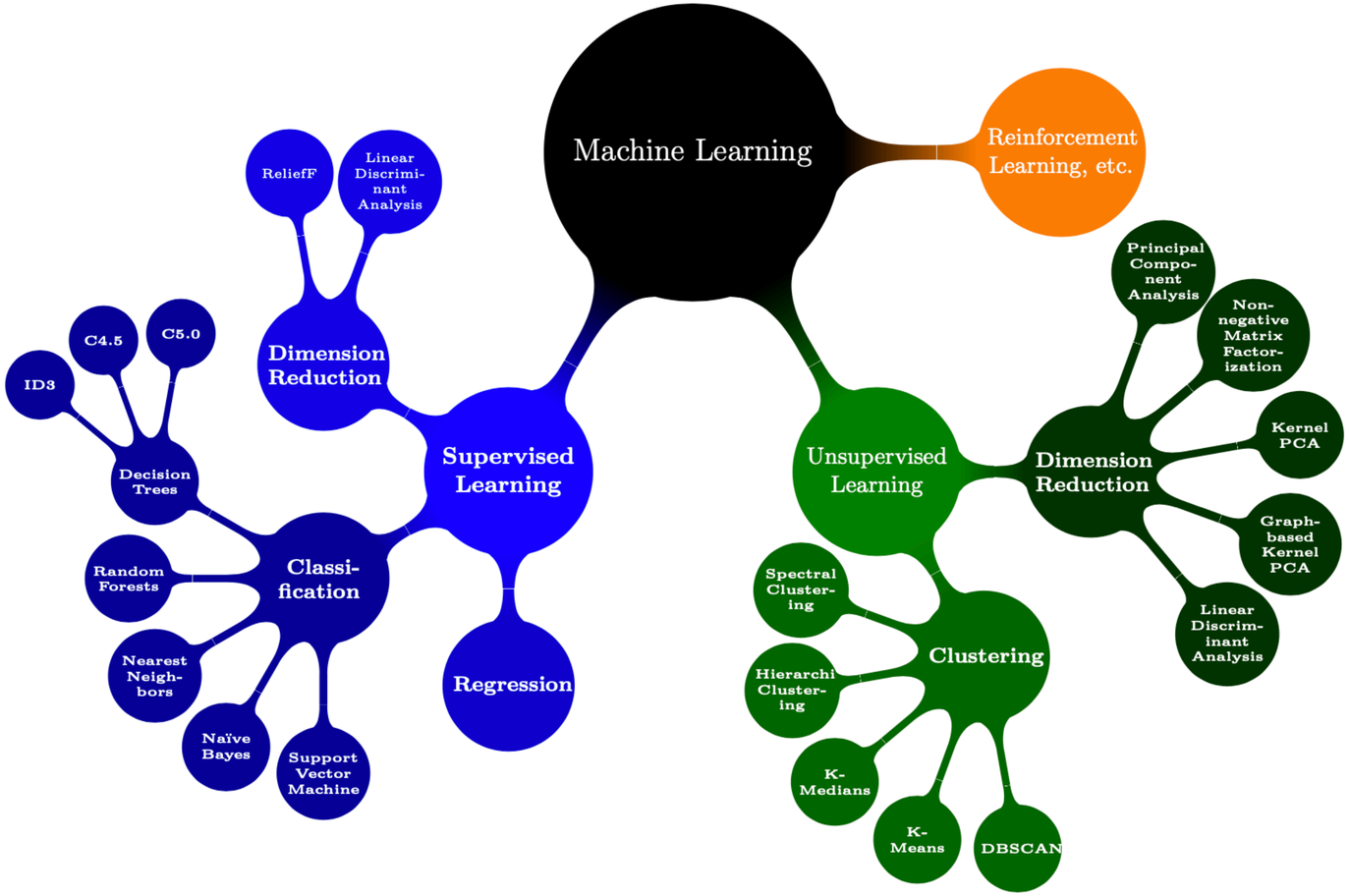
	LoanStatNew	Description
0	loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
1	term	The number of payments on the loan. Values are in months and can be either 36 or 60.
2	int_rate	Interest Rate on the loan
3	installment	The monthly payment owed by the borrower if the loan originates.
4	grade	LC assigned loan grade
5	sub_grade	LC assigned loan subgrade
6	emp_title	The job title supplied by the Borrower when applying for the loan.*
7	emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
8	home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
9	annual_inc	The self-reported annual income provided by the borrower during registration.
10	verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
11	issue_d	The month which the loan was funded
12	loan_status	Current status of the loan
13	purpose	A category provided by the borrower for the loan request.
14	title	The loan title provided by the borrower
15	zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.
16	addr_state	The state provided by the borrower in the loan application
17	dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
18	earliest_cr_line	The month the borrower's earliest reported credit line was opened
19	open_acc	The number of open credit lines in the borrower's credit file.
20	pub_rec	Number of derogatory public records
21	revol_bal	Total credit revolving balance
22	revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
23	total_acc	The total number of credit lines currently in the borrower's credit file
24	initial_list_status	The initial listing status of the loan. Possible values are – W, F
25	application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
26	mort_acc	Number of mortgage accounts.
27	pub_rec_bankruptcies	Number of public record bankruptcies

Lets have a Look in to the Data

- Investor Perspective, first
- Generate some insights
- Understand the challenge in this kind of actives

Out Look

Models



A Schematic Illustration of the Taxonomy and Example Algorithms in Machine Learning in Accordance With the Works of Russell & Norvig (2016) and Bishop (2006)

Roweis, S. T., & Saul, L. K. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science, 290(5500), 2323–2326.

Tip: Work Iterative

Get the Data → **Explore the data** → **Model the data** → **Presentation**

Rebalancing → **Feature Selection** → **Scaling** → **Checking Model with 10 K**

Downsizing

Using Correlation
Using Shape Values

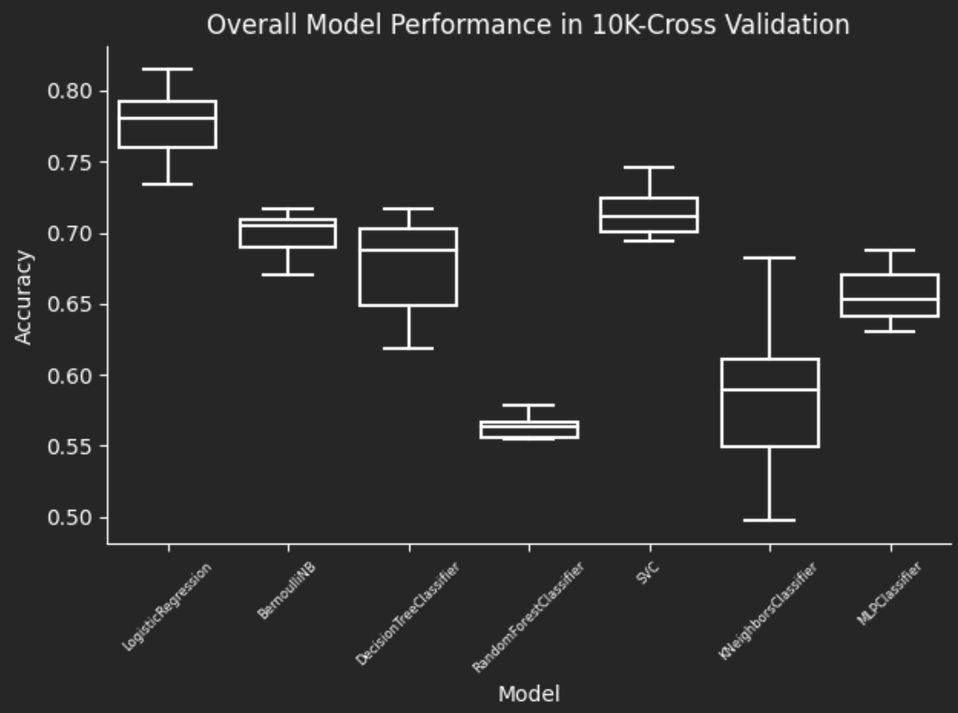
Using Min/Max

	precision
no_bankrupt	0.61
bankrupt	0.64

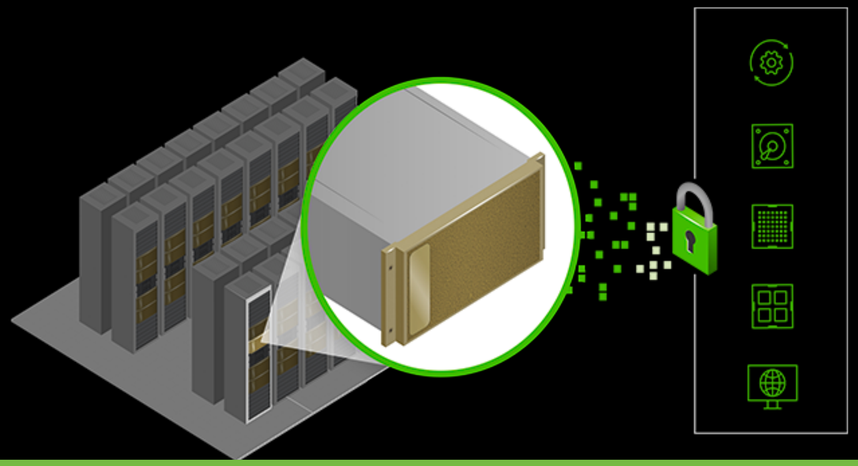
	precision
no_bankrupt	0.60
bankrupt	0.70

	precision
no_bankrupt	0.62
bankrupt	0.81

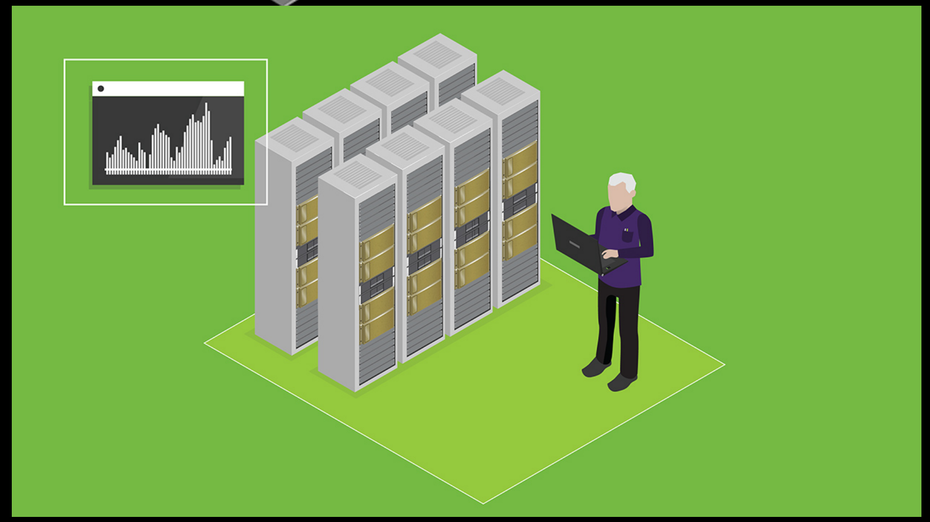
Improvements by using different data pre-processing techniques.



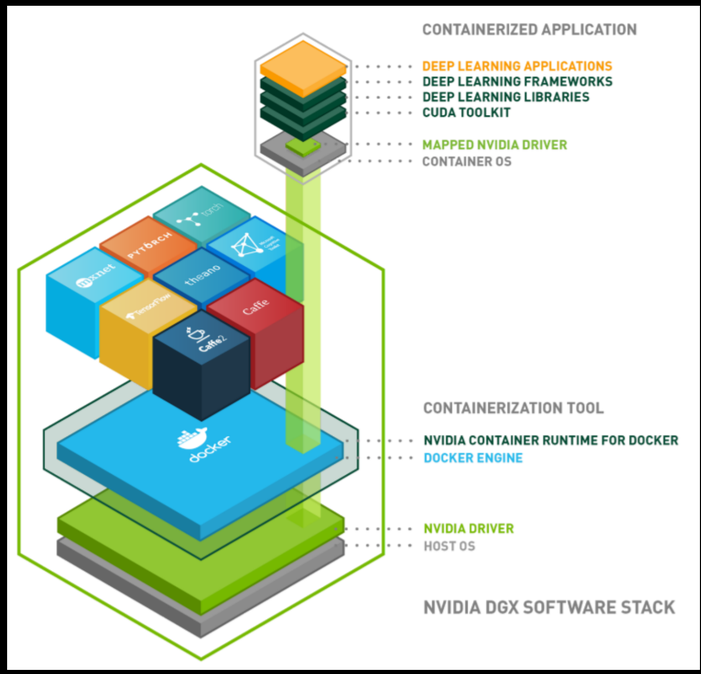
Approach



„Data“ Scientist writing code locally



Executing code on DGX Station via deployed with Container (e.g., **Docker**)!

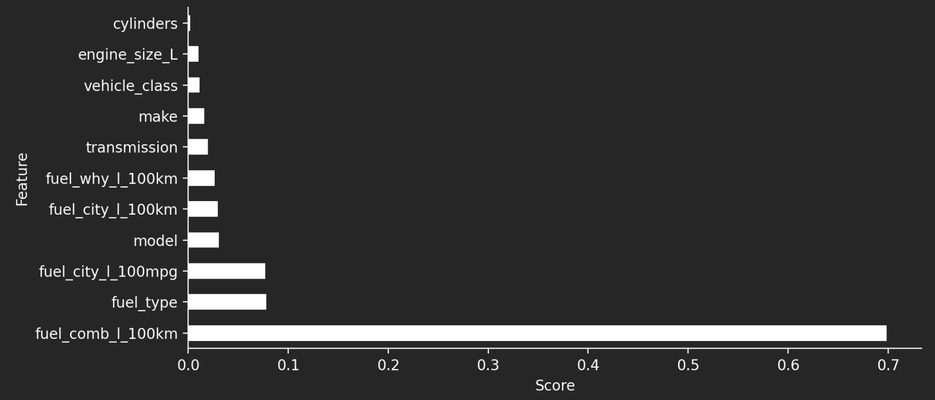


Code and (the rest) is running

Bilder von Nividia https://www.google.com/imgres?imgurl=x-raw-image%3A%2F%2F9c6b15d4b89aab119a5ff1f75089aa8bf7c7bab17d11c8cbed53f57769bd7349&imgrefurl=https%3A%2F%2Fwww.nvidia.cn%2Fcontent%2Fdam%2Fen-zz%2FSolutions%2FData-Center%2Fdgx-station%2Fdgx-station-print-explorer-datasheet-letter-final-web.pdf&tbnid=EauauNQVa5w1mM&vet=10CAkQxiAoCmoXChMIyO6LoprU9AIVAAAAAB0AAAAAEA0..i&docid=gskwpG4XnBhBYM&w=499&h=473&itg=1&q=DGX%20station%20grafikkarten&client=safari&ved=0CAkQxiAoCmoXChMIyO6LoprU9AIVAAAAAB0AAAAAEA0

Data Optimization

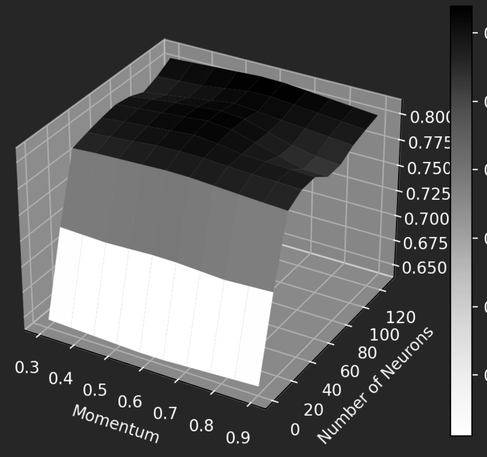
Model Optimization



More Features != Better Performance

Scaling Feature

Starting



1

2

Steps

Research protocol based on Python 3.10 for ML Optimizationunder a MacBook Pro (13", 2018), 2,7 GHz Quad-Core Intel Core i7, 8 GB RAM.

Thank You


YouTube
Suchen
Data Science in a Nutshell
Dr. Benjamin M. Abdel-Karim
Data Science in a Nutshell - Einleitung
Keine Aufrufe · 17.08.2021
TEILEN
SPEICHERN


Benjamin M. Abdel-Karim Hrsg.
Data Science
Best practices mit Python
EBOOK INSIDE
Springer Vieweg