

Introduction to Data Science

Part 1: About Data Science

Nicolas Pfeuffer

Table of Contents

- 1. What is Data Science?**
- 2. Data Science Roles and Skills**
- 3. Data Science Process**
- 4. The first step: Asking the right Questions**
- 5. Step 2 and 2^{1/2} : A first glance at our dataset**

The sexiest job of the 21st century

Data scientists' **most basic, universal skill is the ability to write code**. This may be less true in five years' time, when many more people will have the title "data scientist" on their business cards. More enduring will be the need for data scientists to **communicate in language that all their stakeholders understand**—and to demonstrate the special skills involved in **storytelling with data, whether verbally, visually, or—ideally—both**.

<https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>

"The next sexy job"

"The ability to **take data**—to be able to **understand it**, to **process it**, to **extract value** from it, to **visualize it**, to **communicate it**—that's going to be a hugely important skill."

Hal Varian, Google's Chief Economist, NYT, 2009

"Data Scientist (n.): Person who is **better at statistics than any software engineer** and **better at software engineering than any statistician**."

Josh Wills, 2012

Data Science Headlines

Lessons learned from 2012 U.S. presidential campaign sure to play more important role in future elections.



Analytics, especially microtargeting, played a critical role in both campaigns, and Obama won in significant part because his analytics performance was better. Photo courtesy of "Obama for America."

2,185 views | Aug 27, 2019, 06:11pm

Data Science Will Drive Auto Industry In Future, Tata Consultancy Services Says

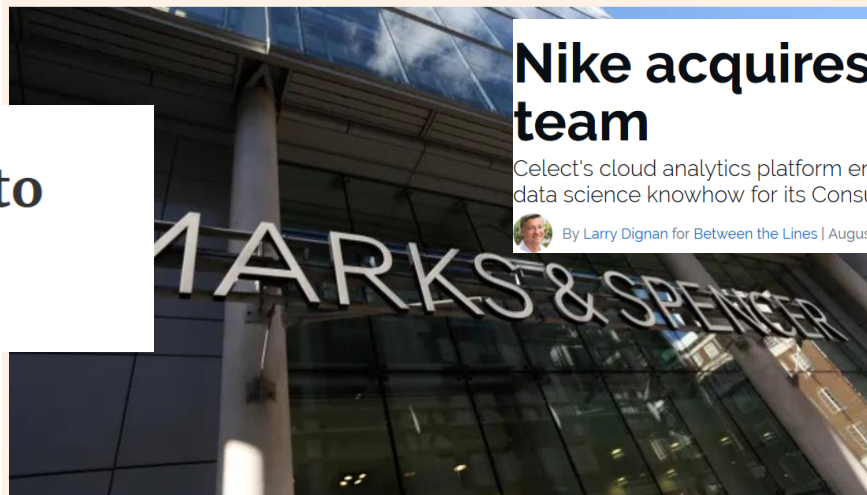
Cargill's iQShrimp helps farmers manage risk, make better decisions

Digital software provides shrimp producers with data clarity and predictive insights

MINNEAPOLIS – March 5, 2018 – Shrimp farmers can tap into the first cloud-based solution in the aquaculture industry thanks to Cargill's iQShrimp. The predictive software uses machine learning and sensors to give them real-time visibility into their farm operations. iQShrimp is a first-generation offering driven by iQuatic™, Cargill's digital platform for aquaculture.

M&S aims to turn staff into data scientists

Retailer wants workers to become digitally savvy in market rocked by likes of Amazon



M&S is first publicly listed retailer in the UK to go down the path of creating a data skills academy © Bloomberg

Nike acquires Celect, adds to data science team

Celect's cloud analytics platform enables retailers to optimize inventory across channels. Nike plans to use that data science knowhow for its Consumer Direct Offense strategy.

By Larry Dignan for [Between the Lines](#) | August 6, 2019 -- 21:06 GMT (22:06 BST) | Topic: [Big Data Analytics](#)

Data Science for Social Good



Using a range of data both held by Ofsted and publicly available, the goal of this project is to build a risk model that can help Ofsted to prioritise its inspections of IFAs.



The goal of the project is to reduce the time it takes to find a new rough sleeper and to improve access to services for particularly vulnerable individuals.



The goal was a resource that can provide real-time local labor market data that helps employers, job-seekers, and educators collaborate to target skills training and match potential employees to jobs.

Definitions of Data Science

Data Science aims to **gain insights into data** through **computation, statistics, and visualization**.

by Pfister, Blitzstein and Kaynig for their course „CS109 Data Science“ at Harvard John A. Paulson School of Engineering

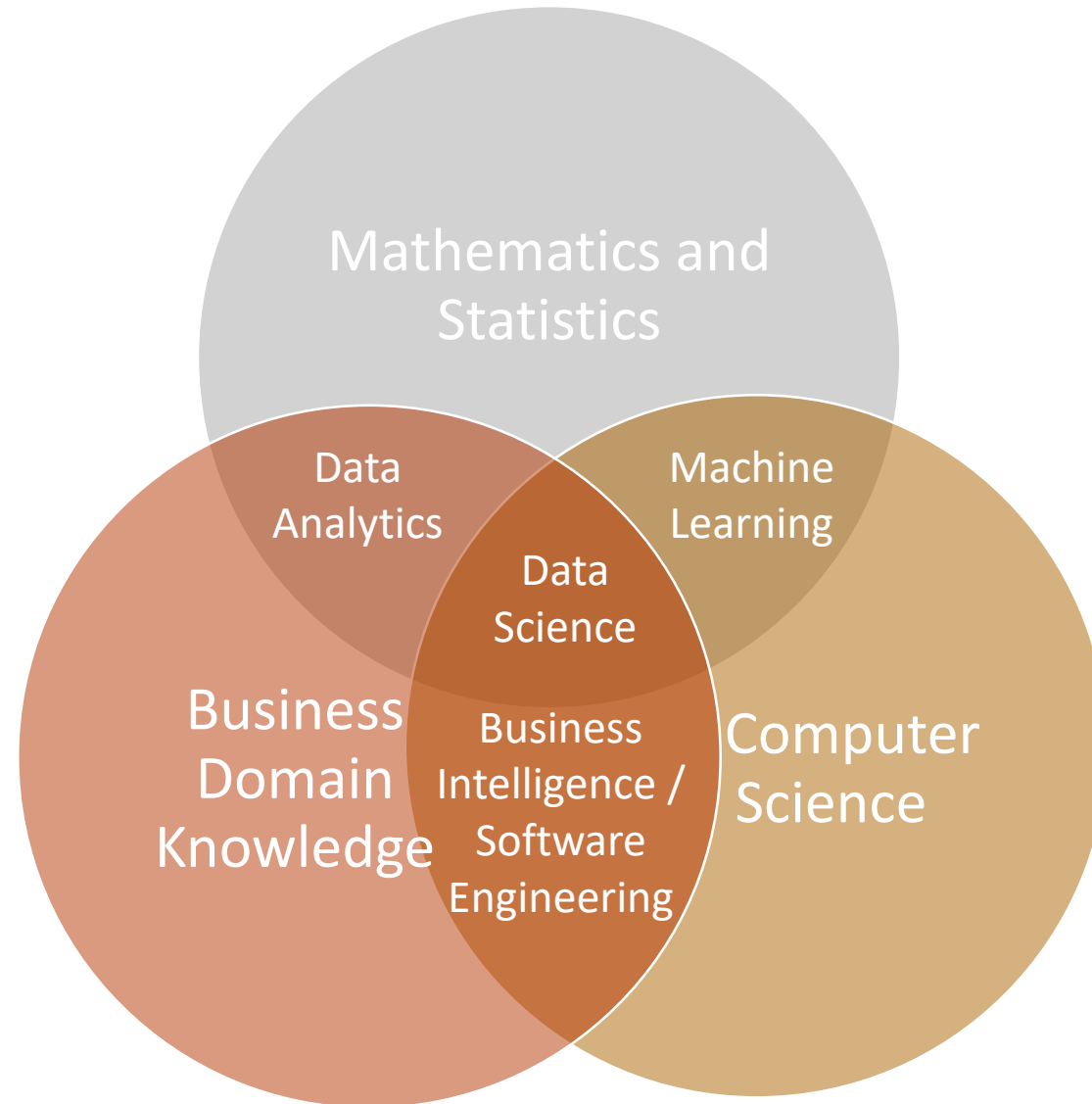
Data science is the **study of the generalizable extraction of knowledge from data**.

Vasant Dhar. Data Science and Prediction. Communications of the ACM, December 2013, Vol. 56 No. 12, Pages 64-73

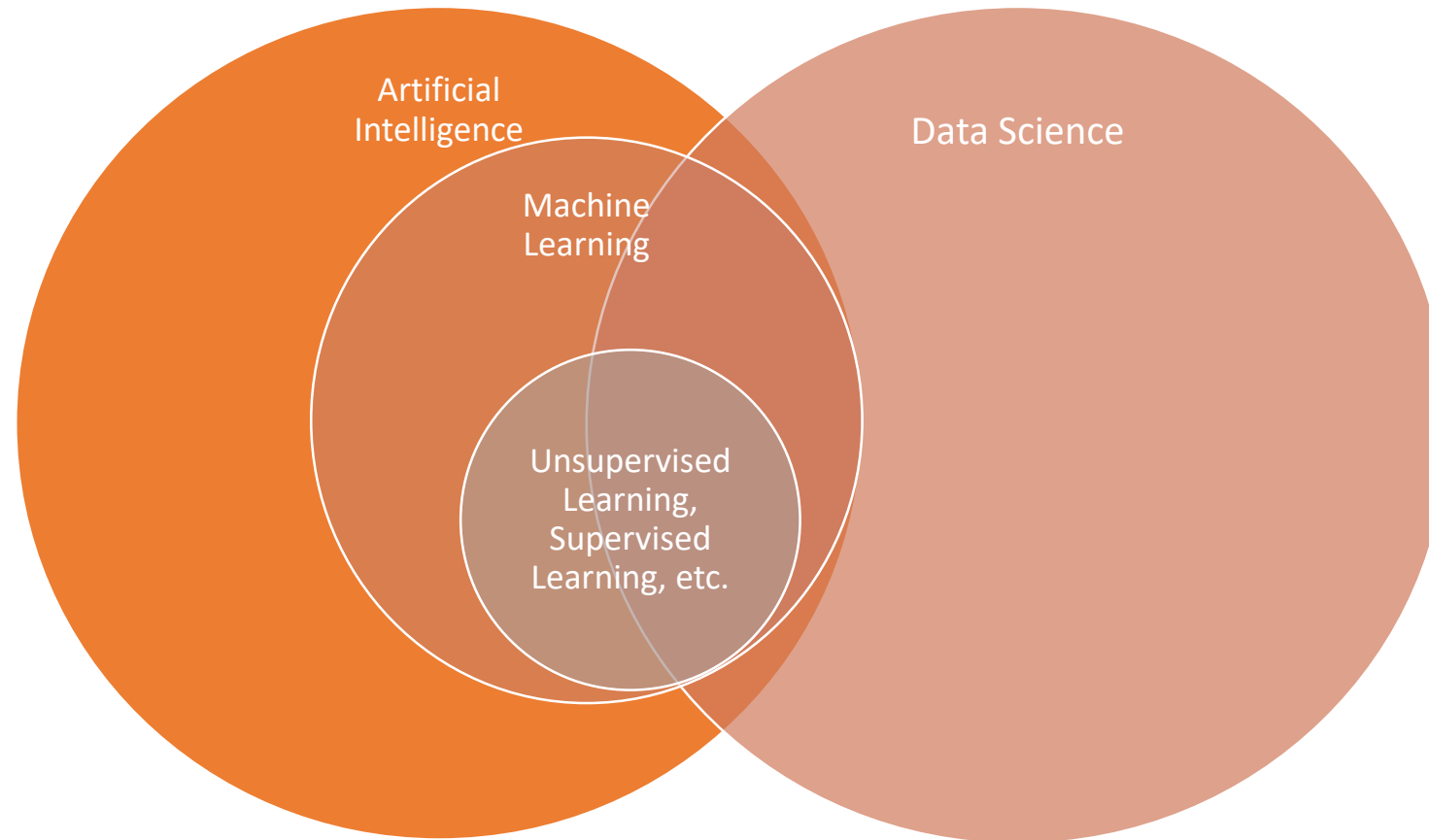
[...]Data Science is the **science of learning from data**; it studies the **methods involved in the analysis and processing of data** and **proposes technology to improve methods in an evidence-based manner**. The scope and impact of this science will expand enormously in coming decades as scientific data and data about science itself become ubiquitously available.

David Donoho. 50 Years of Data Science. 2015, p. 38

How is Data Science Distinguishable?



How is Data Science Distinguishable?



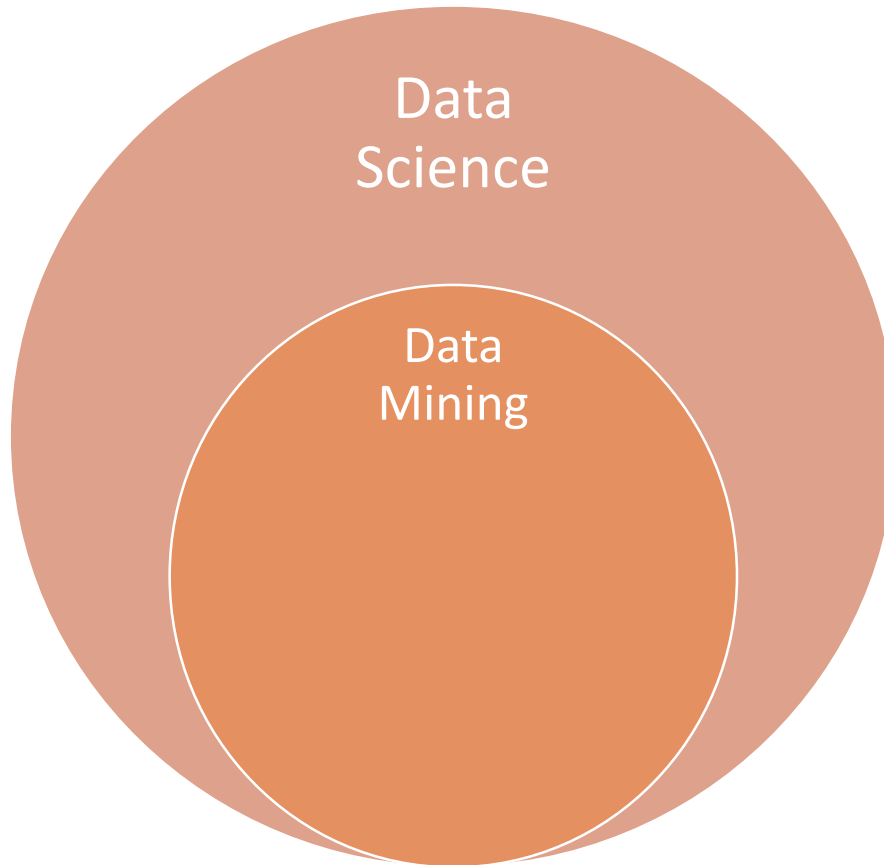
Data Mining

A process that uses statistical, mathematical, artificial intelligence and machine-learning techniques to extract and identify useful information and subsequent knowledge from large databases.



Data mining tools find patterns in data and may even infer rules/models from them.

How is Data Science Distinguishable?



- Although at first glance quite similar, but different
 - Data Mining:
 - is part of a broader Knowledge Discovery Process
 - Goal → Get Data, Find useful patterns in data
 - Data Science:
 - Follow (Research/Business) Questions /Hypotheses
 - Goals → Analyse Data, Generate Predictions, Find answers in an evidence-based manner
- Data Mining can be seen as a sub-task in a data science process

Table of Contents

1. What is Data Science?

2. Data Science Roles and Skills

3. Data Science Process

4. The first step: Asking the right Questions

5. Step 2 and 2^{1/2} : A first glance at our dataset

Typical Roles in Data Science

“I worry that the Data Scientist role is like the mythical “webmaster” of the 90s: master of all trades.”

Aaron Kimball, CTO Wibidata

“No matter how you define data science, you’ll find practitioners for whom the definition is totally, absolutely wrong.”

Joel Grus, in his book “Data Science from Scratch”. O’Reilly, 2015.

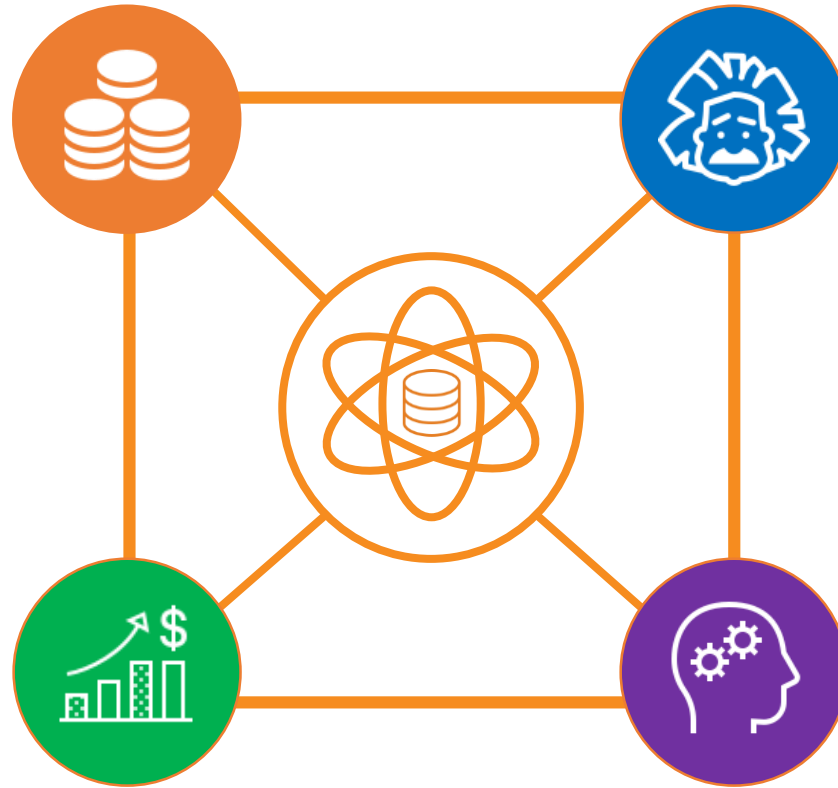
Typical Roles in Data Science

Data Engineer

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

Data Analyst

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



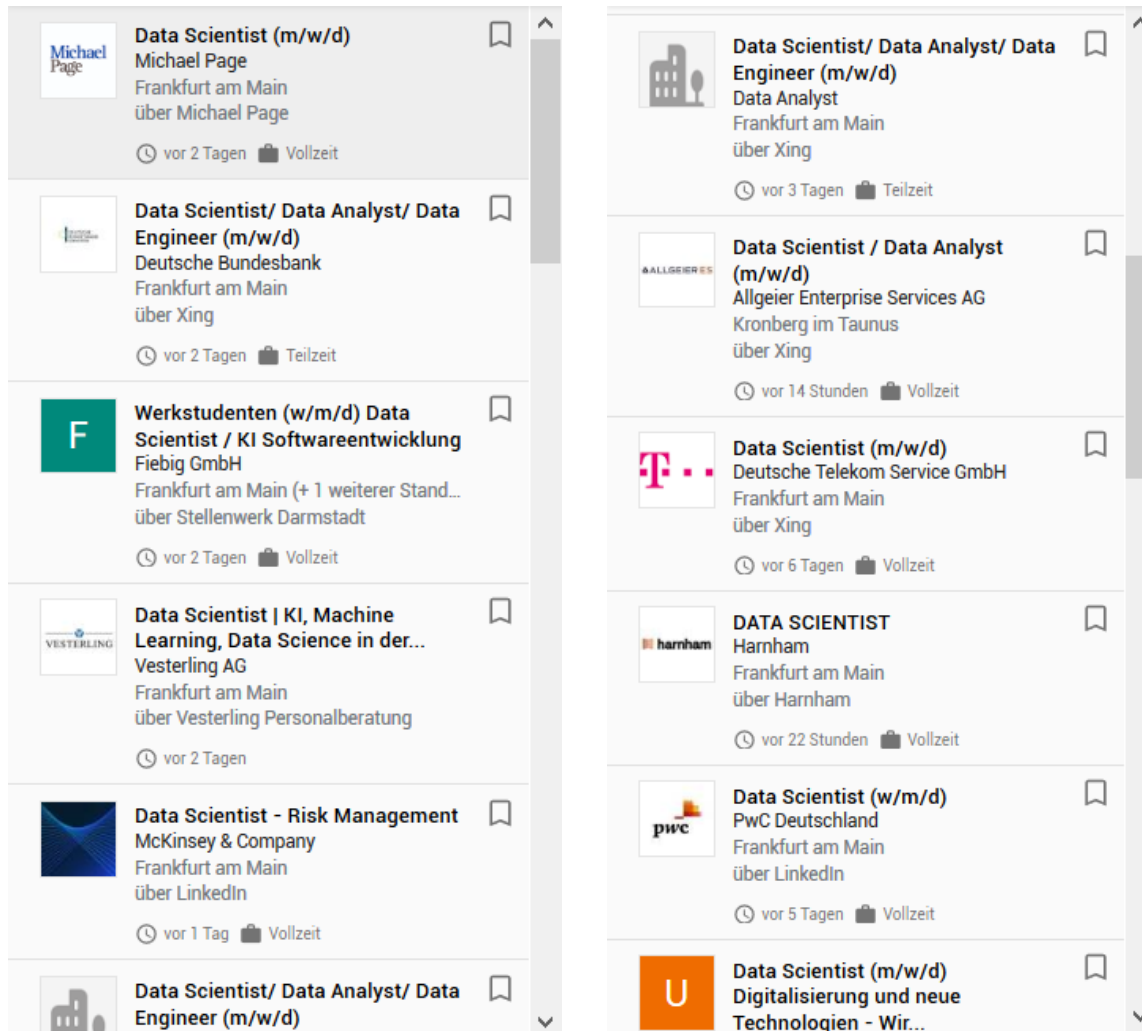
General Data Scientist

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

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.

Job Offers in Data Science –Data Scientist ?



The image shows two side-by-side screenshots of job offer listings. The left column contains listings from Xing, and the right column contains listings from LinkedIn. Each listing includes a company logo, the job title, the company name, the location, and the source of the listing. Some listings also indicate the time since they were posted and the type of employment (full-time or part-time).

Company	Job Title	Location	Source	Time Posted	Employment Type
Michael Page	Data Scientist (m/w/d)	Frankfurt am Main	über Michael Page	vor 2 Tagen	Vollzeit
Deutsche Bundesbank	Data Scientist/ Data Analyst/ Data Engineer (m/w/d)	Frankfurt am Main	über Xing	vor 2 Tagen	Teilzeit
Fiebig GmbH	Werkstudenten (w/m/d) Data Scientist / KI Softwareentwicklung	Frankfurt am Main (+ 1 weiterer Standort)	über Stellenwerk Darmstadt	vor 2 Tagen	Vollzeit
Vesterling AG	Data Scientist KI, Machine Learning, Data Science in der...	Frankfurt am Main	über Vesterling Personalberatung	vor 2 Tagen	
McKinsey & Company	Data Scientist - Risk Management	Frankfurt am Main	über LinkedIn	vor 1 Tag	Vollzeit
harnham	DATA SCIENTIST	Frankfurt am Main	über Harnham	vor 22 Stunden	Vollzeit
PwC Deutschland	Data Scientist (w/m/d)	Frankfurt am Main	über LinkedIn	vor 5 Tagen	Vollzeit
U	Data Scientist (m/w/d)	Digitalisierung und neue Technologien - Wir...			

Your profile

- Accredited Master's degree or equivalent in (business) informatics, (business) mathematics or economics with above-average results
- Experience as Data Scientist, Data Analyst or Data Engineer
- Good knowledge of big data systems, data management, data organization, data analysis and visualization
- Practical experience in dealing with data management systems, data organisation and administration as well as the creation and evaluation of data sets
- Good knowledge of current technologies as well as IT trends and market developments in the area of KI/ML
- Programming experience (e.g. Python, Java, R)
- Knowledge and experience in innovation management are an advantage
- Strong analytical and conceptual skills
- Commitment, flexibility and ability to work in a team as well as service and customer orientation
- Good communication, presentation and moderation skills
- Very good knowledge of German and English

Job Offers in Data Science – General DS

The image shows two side-by-side screenshots of job offer listings. The left column contains five listings, and the right column contains five listings. Each listing includes a company logo, the job title, the company name, the location, the source, and the time posted. The listings are as follows:

Company	Job Title	Location	Source	Time Posted
Michael Page	Data Scientist (m/w/d)	Frankfurt am Main	über Michael Page	vor 2 Tagen
Deutsche Bundesbank	Data Scientist/ Data Analyst/ Data Engineer (m/w/d)	Frankfurt am Main	über Xing	vor 2 Tagen
Fiebig GmbH	Werkstudenten (w/m/d) Data Scientist / KI Softwareentwicklung	Frankfurt am Main (+ 1 weiterer Standort)	über Stellenwerk Darmstadt	vor 2 Tagen
Vesterling AG	Data Scientist KI, Machine Learning, Data Science in der...	Frankfurt am Main	über Vesterling Personalberatung	vor 2 Tagen
McKinsey & Company	Data Scientist - Risk Management	Frankfurt am Main	über LinkedIn	vor 1 Tag
Allgeier Enterprise Services AG	Data Scientist / Data Analyst (m/w/d)	Kronberg im Taunus	über Xing	vor 14 Stunden
Deutsche Telekom Service GmbH	Data Scientist (m/w/d)	Frankfurt am Main	über Xing	vor 6 Tagen
Harnham	DATA SCIENTIST	Frankfurt am Main	über Harnham	vor 22 Stunden
PwC Deutschland	Data Scientist (w/m/d)	Frankfurt am Main	über LinkedIn	vor 5 Tagen
U	Data Scientist (m/w/d)	Digitalisierung und neue Technologien - Wir...		

Your profile

- Accredited Master's degree or equivalent in (business) informatics, (business) mathematics or economics with above-average results
- Experience as **Data Scientist, Data Analyst or Data Engineer**
- Good **knowledge of big data systems, data management, data organization, data analysis and visualization**
- Practical experience in dealing with data management systems, data organisation and administration as well as the creation and **evaluation of data sets**
- **Good knowledge of current technologies** as well as IT trends and market developments **in the area of KI/ML**
- Programming experience (e.g. **Python, Java, R**)
- Knowledge and experience in innovation management are an **advantage**
- **Strong analytical and conceptual skills**
- Commitment, flexibility and ability to work in a team as well as service and customer orientation
- **Good communication, presentation and moderation skills**
- Very good knowledge of German and English

Job Offers in Data Science – Data Scientist ?

The image shows two side-by-side screenshots of job offer listings. The left screenshot is from Xing and the right is from LinkedIn. Both list various Data Scientist and Data Analyst positions in Frankfurt am Main.

Left Column (Xing):

- Data Scientist (m/w/d)**
Michael Page
Frankfurt am Main
über Michael Page
vor 2 Tagen | Vollzeit
- Data Scientist/ Data Analyst/ Data Engineer (m/w/d)**
Deutsche Bundesbank
Frankfurt am Main
über Xing
vor 2 Tagen | Teilzeit
- Werkstudenten (w/m/d) Data Scientist / KI Softwareentwicklung**
Fiebig GmbH
Frankfurt am Main (+ 1 weiterer Stand...)
über Stellenwerk Darmstadt
vor 2 Tagen | Vollzeit
- Data Scientist | KI, Machine Learning, Data Science in der...**
Vesterling AG
Frankfurt am Main
über Vesterling Personalberatung
vor 2 Tagen
- Data Scientist - Risk Management**
McKinsey & Company
Frankfurt am Main
über LinkedIn
vor 1 Tag | Vollzeit
- Data Scientist/ Data Analyst/ Data Engineer (m/w/d)**


Right Column (LinkedIn):


- Data Scientist/ Data Analyst/ Data Engineer (m/w/d)**
Data Analyst
Frankfurt am Main
über Xing
vor 3 Tagen | Teilzeit
- Data Scientist / Data Analyst (m/w/d)**
Allgeier Enterprise Services AG
Kronberg im Taunus
über Xing
vor 14 Stunden | Vollzeit
- Data Scientist (m/w/d)**
Deutsche Telekom Service GmbH
Frankfurt am Main
über Xing
vor 6 Tagen | Vollzeit
- DATA SCIENTIST**
Harnham
Frankfurt am Main
über Harnham
vor 22 Stunden | Vollzeit
- Data Scientist (w/m/d)**
PwC Deutschland
Frankfurt am Main
über LinkedIn
vor 5 Tagen | Vollzeit
- Data Scientist (m/w/d)**
Digitalisierung und neue Technologien - Wir...


Your Profile


- Masters degree in computer science
- Several years of experience with machine learning models including their operation
- Fit in Python & Libraries such as scikit-learn, pandas, NumPy, TensorFlow
- Experienced in working with Big Data platforms and tools like Hadoop, Spark und Flink
- Profound know-how in software development incl. software testing, Continuous Integration, code reviews, etc.
- Constructive in discussions in English and willing to learn German
- Ready to travel up to max. 1/4 of your working time


Job Offers in Data Science – ML Engineer


**Data Scientist (m/w/d)**
Michael Page
Frankfurt am Main
über Michael Page
vor 2 Tagen Vollzeit


**Data Scientist/ Data Analyst/ Data Engineer (m/w/d)**
Deutsche Bundesbank
Frankfurt am Main
über Xing
vor 2 Tagen Teilzeit


**Werkstudenten (w/m/d) Data Scientist / KI Softwareentwicklung**
Fiebig GmbH
Frankfurt am Main (+ 1 weiterer Stand...)
über Stellenwerk Darmstadt
vor 2 Tagen Vollzeit


**Data Scientist | KI, Machine Learning, Data Science in der...**
Vesterling AG
Frankfurt am Main
über Vesterling Personalberatung
vor 2 Tagen


**Data Scientist - Risk Management**
McKinsey & Company
Frankfurt am Main
über LinkedIn
vor 1 Tag Vollzeit


**Data Scientist/ Data Analyst/ Data Engineer (m/w/d)**
S...
...


**Data Scientist/ Data Analyst/ Data Engineer (m/w/d)**
Data Analyst
Frankfurt am Main
über Xing
vor 3 Tagen Teilzeit

**Data Scientist / Data Analyst (m/w/d)**
Allgeier Enterprise Services AG
Kronberg im Taunus
über Xing
vor 14 Stunden Vollzeit

**Data Scientist (m/w/d)**
Deutsche Telekom Service GmbH
Frankfurt am Main
über Xing
vor 6 Tagen Vollzeit

**DATA SCIENTIST**
Harnham
Frankfurt am Main
über Harnham
vor 22 Stunden Vollzeit

**Data Scientist (w/m/d)**
PwC Deutschland
Frankfurt am Main
über LinkedIn
vor 5 Tagen Vollzeit

**Data Scientist (m/w/d)**
Digitalisierung und neue Technologien - Wir...
...

Your Profile

- **Masters degree in computer science**
- Several years of **experience with machine learning models** including their operation
- Fit in Python & Libraries such as **scikit-learn, pandas, NumPy, TensorFlow**
- Experienced in working with **Big Data platforms** and tools like **Hadoop, Spark und Flink**
- **Profound know-how in software development** incl. software testing, Continuous Integration, code reviews, etc.
- Constructive in discussions in English and willing to learn German
- Ready to travel up to max. 1/4 of your working time

Job Offers in Data Science – Data Scientist ?

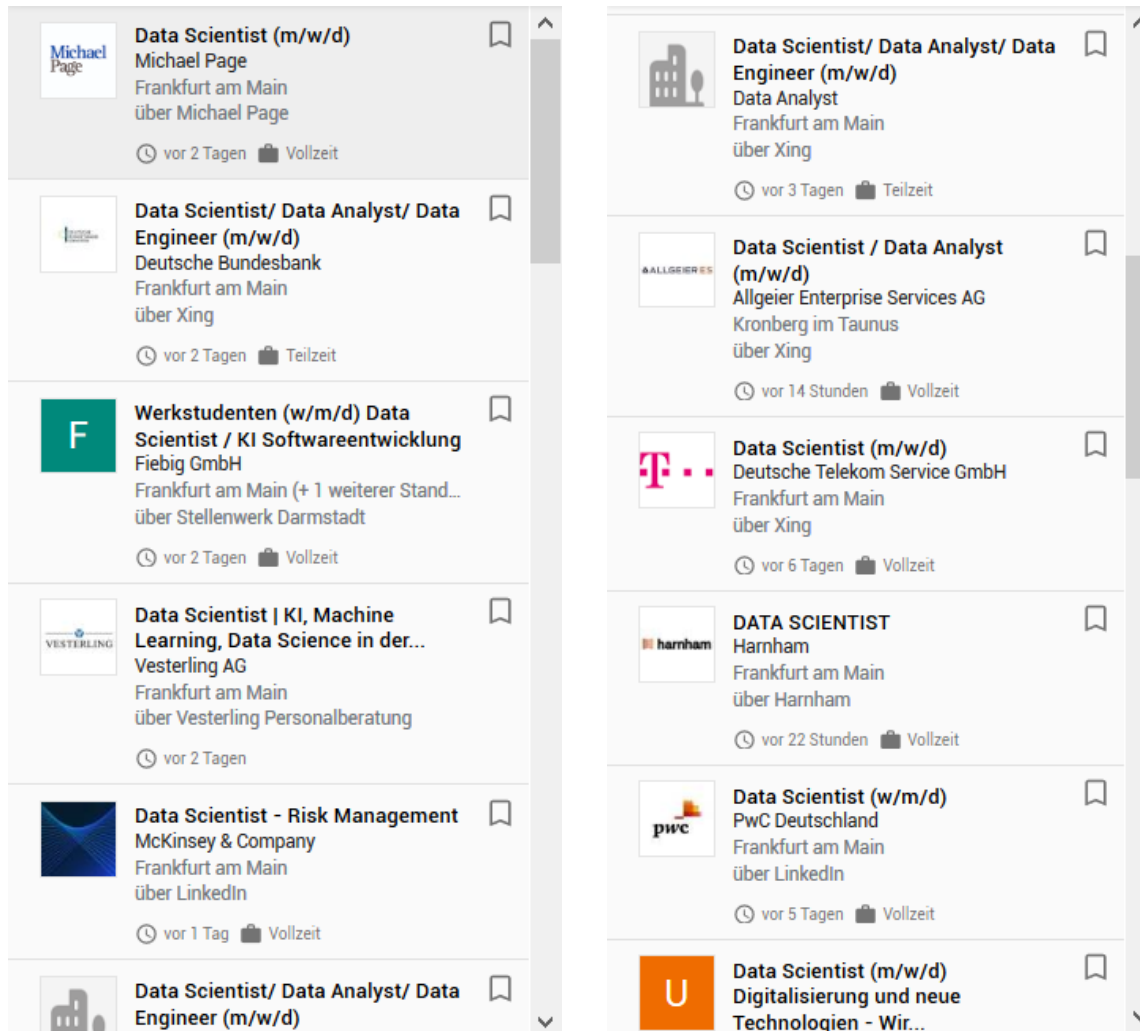
The image shows two side-by-side screenshots of job offer listings. The left column contains five listings, and the right column contains five listings. Each listing includes a company logo, the job title, the company name, the location, the source, and the time posted. The listings are as follows:

Company	Job Title	Location	Source	Time Posted
Michael Page	Data Scientist (m/w/d)	Frankfurt am Main	über Michael Page	vor 2 Tagen
Deutsche Bundesbank	Data Scientist/ Data Analyst/ Data Engineer (m/w/d)	Frankfurt am Main	über Xing	vor 2 Tagen
Fiebig GmbH	Werkstudenten (w/m/d) Data Scientist / KI Softwareentwicklung	Frankfurt am Main (+ 1 weiterer Standort)	über Stellenwerk Darmstadt	vor 2 Tagen
Vesterling AG	Data Scientist KI, Machine Learning, Data Science in der...	Frankfurt am Main	über Vesterling Personalberatung	vor 2 Tagen
McKinsey & Company	Data Scientist - Risk Management	Frankfurt am Main	über LinkedIn	vor 1 Tag
harnham	DATA SCIENTIST	Frankfurt am Main	über Harnham	vor 22 Stunden
PwC	Data Scientist (w/m/d)	Frankfurt am Main	über LinkedIn	vor 5 Tagen
U	Data Scientist (m/w/d)	Digitalisierung und neue Technologien - Wir...		

Your Profile

- Carrying out qualitative and quantitative analyses
- Evaluation, interpretation and visualization of the results
- Derivation of action measures for the optimization of business processes
- Coordination and close cooperation with internal departments
- Project work in a team with subsequent presentation

Job Offers in Data Science – Data Analyst



The screenshot displays two columns of job listings from a portal. Each listing includes a company logo, the job title, company name, location, source, and posting details.

Company	Job Title	Location	Source	Posting Time	Contract Type
Michael Page	Data Scientist (m/w/d)	Frankfurt am Main	über Michael Page	vor 2 Tagen	Vollzeit
Deutsche Bundesbank	Data Scientist/ Data Analyst/ Data Engineer (m/w/d)	Frankfurt am Main	über Xing	vor 2 Tagen	Teilzeit
Fiebig GmbH	Werkstudenten (w/m/d) Data Scientist / KI Softwareentwicklung	Frankfurt am Main (+ 1 weiterer Standort)	über Stellenwerk Darmstadt	vor 2 Tagen	Vollzeit
Vesterling AG	Data Scientist KI, Machine Learning, Data Science in der...	Frankfurt am Main	über Vesterling Personalberatung	vor 2 Tagen	
McKinsey & Company	Data Scientist - Risk Management	Frankfurt am Main	über LinkedIn	vor 1 Tag	Vollzeit
Allgeier Enterprise Services AG	Data Scientist / Data Analyst (m/w/d)	Kronberg im Taunus	über Xing	vor 14 Stunden	Vollzeit
Deutsche Telekom Service GmbH	Data Scientist (m/w/d)	Frankfurt am Main	über Xing	vor 6 Tagen	Vollzeit
Harnham	DATA SCIENTIST	Frankfurt am Main	über Harnham	vor 22 Stunden	Vollzeit
PwC	Data Scientist (w/m/d)	Frankfurt am Main	über LinkedIn	vor 5 Tagen	Vollzeit
U	Data Scientist (m/w/d) Digitalisierung und neue Technologien - Wir...				

Your Profile

- Carrying out **qualitative and quantitative analyses**
- **Evaluation, interpretation and visualization** of the results
- **Derivation of action measures** for the optimization of business processes
- Coordination and close cooperation with internal departments
- Project work in a team with subsequent presentation

Skills of the modern data scientist:



Basic Skills

```
# Variables

iVariableA = 1
iVariableB = 2
iVariableC = iVariableA/iVariableB

# Data Structures

LNames = ['Gary','Marc','Paula']
LNames_reversed = LNames.reverse()
sNameGary = LNames[0]

DLanguages ={'England':'English',0101:'Python'}
sDataScienceLanguage = DLanguages['0101']

# Control Structures

iNumber = 5
if iNumber < 10:
    print('Number is less than 10')
else:
    print('Number is greater or equal to 10')

# Loops

LNumbers = [1,2,3,4,5,6,7,8,9,10]

for iNum in LNumbers:
    print(iNum)
```

Functions

```
# Functions

LNumbers = [5,23,37,49,50,46,30,46,70]

def count_list (list):
    # 1. while loop
    iCount = 0
    iSizeofList = len(LNumbers)
    i = 0

    while i < iSizeofList:
        if LNumbers[i] > 30:
            iCount = iCount + 1
            i = i + 1

    return iCount

iFinalCounts = count_list(LNumbers)

# Objects and Classes (not part of these lectures)

class DSClass:
    '''This is an example Data Scientist class'''
    def __init__(self,title = 'Data Scientist'):
        self.title = title

    def work(self):
        print('Doing Data Science Magic')
```

Basic Libraries

```
# os

import os

for i in range(10):
    os.mkdir("folder_"+str(i))

# csv and writing files
import csv

with open ("test.csv", "w") as testFile:
    writer = csv.writer(testFile, delimiter=",")
    for row in data:
        writer.writerow(row)

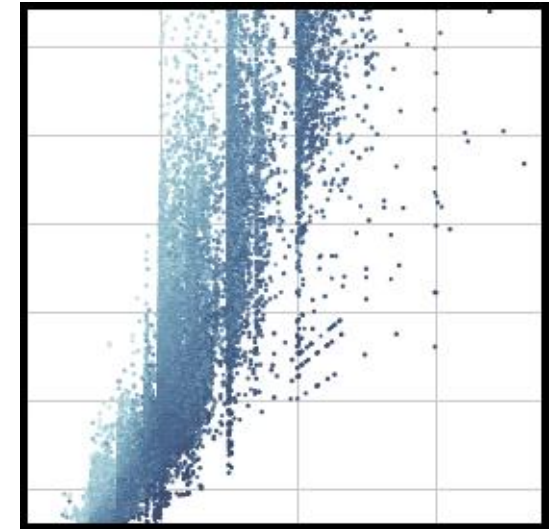
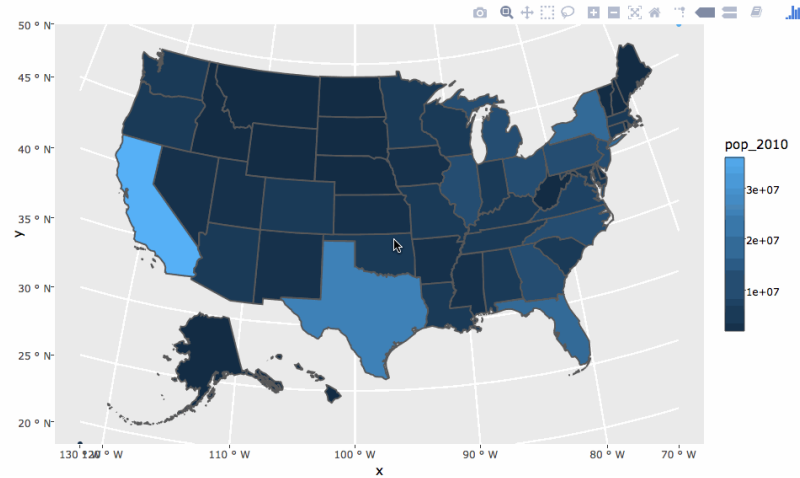
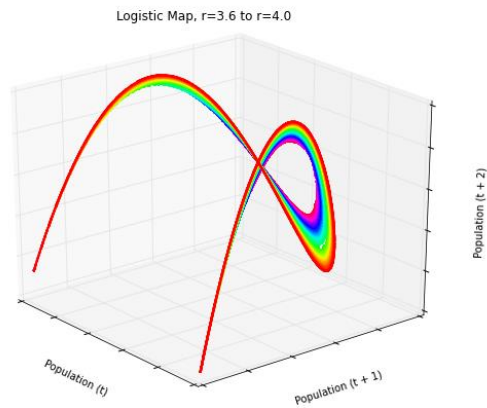
# re - regular expressions

import re

re.findall(pattern="[0-9]", string="4 plus 8
yields 12")
re.findall(pattern="[0-9]+", string="4 plus 8
yields 12")

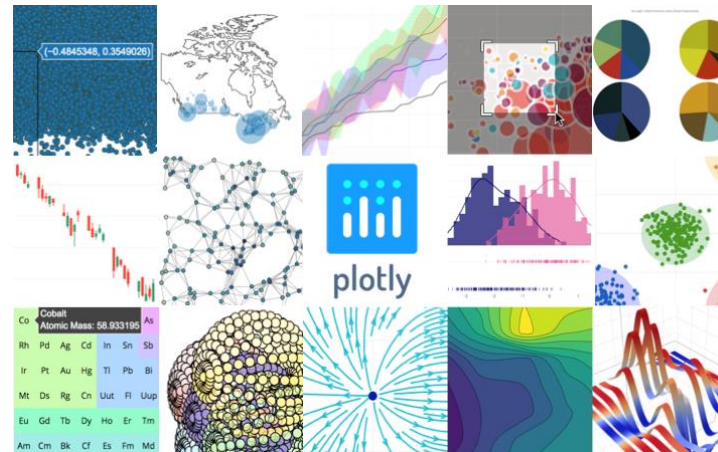
# further libs, see python core libraries
```

Skills ..: Data Visualization



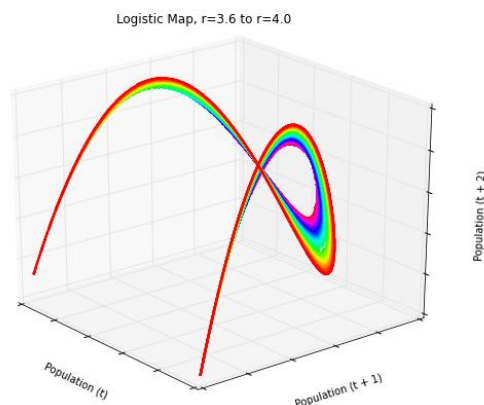
seaborn

matplotlib



<https://moderndata.plot.ly/plotly-4-7-0-now-on-cran/>
<https://geoffboeing.com/2015/04/animated-3d-plots-py/>

Skills ..: Data Visualization



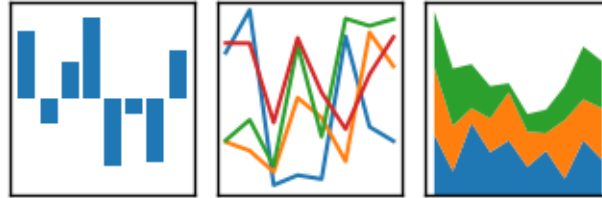
- Build affordances that prioritize data exploration and comparison.
- Emphasize clarity and transparency.
- Reduce cognitive load and focus on what matters.
- Experiences should be intuitive and easy to use.
- Provide context and help users navigate the data.
- Consider performance, polish, surprise, and innovation.



Skills ...: Data Wrangling, Statistics and Calculus

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



```
In [ ]: %matplotlib inline
import cartoframes
from cartoframes import Credentials
import pandas as pd

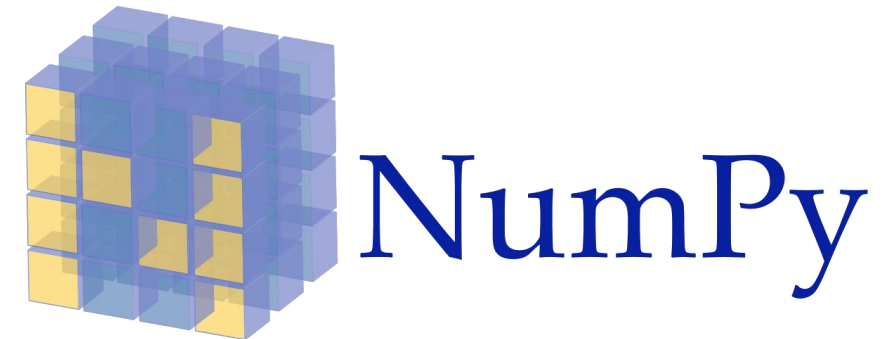
cc = cartoframes.CartoContext()
```

```
In [ ]: # Get a CARTO table as a pandas DataFrame
df = cc.read('brooklyn_poverty')
df.head(15)
```

```
In [ ]: from cartoframes import Layer
cc.map(Layer('brooklyn_poverty'))
```

Notice that:

- the index of the DataFrame is the same as the index of the CARTO table (cartodb_id)
- the_geom column stores the geometry. This can be decoded if we set the

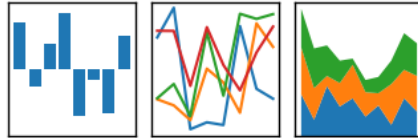


<https://cartoframes.readthedocs.io/en/stable/>

Skills ...: Data Wrangling, Statistics and Calculus

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



- Obtain the data from the right sources.
- Store and structure the data efficiently with suitable data structures.
- Plot the data to explore their nature and generate first insights.
- Perform statistical operations and data transformations to generate insights and prepare your data for inferential operations, such as machine learning.

```
In [ ]: %matplotlib inline
import cartoframes
from cartoframes import Credentials
import pandas as pd

cc = cartoframes.CartoContext()
```

```
In [ ]: # Get a CARTO table as a pandas DataFrame
df = cc.read('brooklyn_poverty')
df.head(15)
```

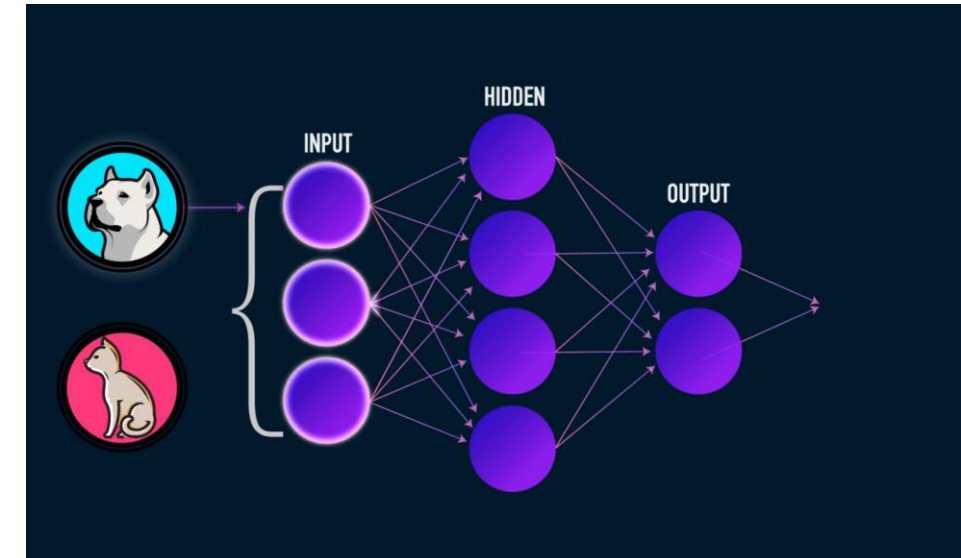
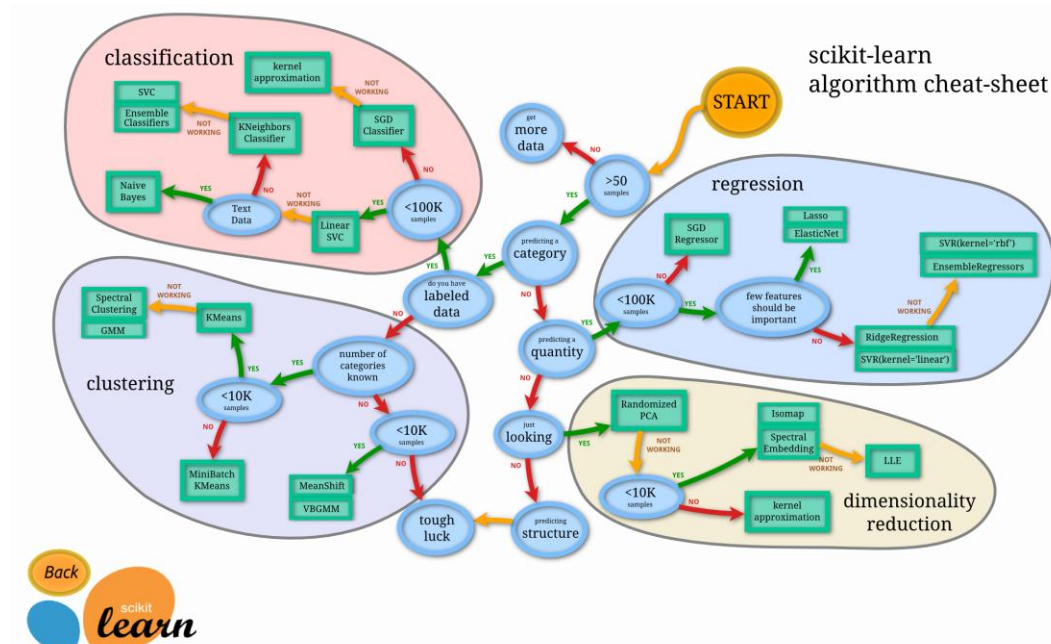
```
In [ ]: from cartoframes import Layer
cc.map(Layer('brooklyn_poverty'))
```

Notice that:

- the index of the DataFrame is the same as the index of the CARTO table (cartodb_id)
- the_geom column stores the geometry. This can be decoded if we set the

<https://cartoframes.readthedocs.io/en/stable/>

Skills ...: Inference, Machine Learning and Deep Learning



Inference and Machine Learning

Different Inference tools are appropriate for different kind of problems:

Statistical Inference (Inferenzstatistik)

- Statistical inference analyzes data to infer properties of a population, by testing hypotheses and deriving estimates.
- A conclusion of an inference is e.g., a probability estimate, a classification or clustering.
- Easier to implement than ML methods, focus on high transparency
- E.g., Bayesian Statistics, Likelihood statistics
- E.g., Linear Regression, Lasso

Machine learning (Maschinelles Lernen)

- Machine Learning deals with the development of and research on computer based algorithms capable of learning to improve their task performance (Jordan & Mitchell 2015).
- ML provides the capabilities to train powerful models on large amounts of data
- (Rather) Focus on model performance
- E.g., Decision Trees, Support Vector Machines, Neural Nets

Deep learning (ML mit tiefen Neuronalen Netzen)

- Special sub-area of machine learning that deals with architectures of neural nets that consist of multiple layers of a multitude of neurons.
- Models are highly non-linear and powerful, but also less transparent than simpler ML methods, such as Decision Trees or Naïve Bayes.
- Appropriate for very complex problems
- E.g., Convolutional Neural Nets, LSTM-Nets

In general, there are three major learning categories on how algorithms can learn:

Supervised learning (Überwachtes Lernen)

- For each observation (e.g. person, age, interests) there is an associated target value (purchased products).
- Two problems:
 - Target value is continuous, e.g. quantity of products purchased → Regression, e.g. Regression Trees
 - Target value is categorical, e.g. which products purchased → Classification, e.g. Decision Trees

Machine Learning- Supervised Learning

Preset: A set of labels(classes) is known

Task-Classification: Estimate to which class x belongs.

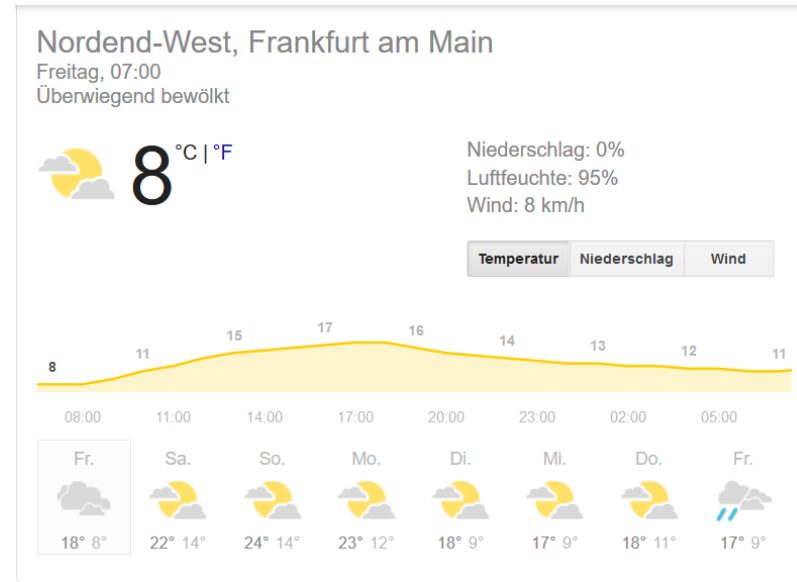


I know there are dogs, cats and muffins. Which one is it?

Machine Learning- Supervised Learning

Preset: Some info (attributes/features) that could give us an idea about the state\quantity\probability of an object or event of interest in the future.

Task-Regression/Prediction: Estimate the unknown quantity of interest in the future.



I have some weather sensor data, historical data...

Which weather is it going to be tomorrow?

In general, there are three major learning categories on how algorithms can learn:

Supervised learning (Überwachtes Lernen)

- For each observation (e.g. person, age, interests) there is an associated target value (purchased products).
- Two problems:
 - Target value is continuous, e.g. quantity of products purchased → Regression, e.g. Regression Trees
 - Target value is categorical, e.g. which products purchased → Classification, e.g. Decision Trees

Unsupervised learning (Unüberwachtes Lernen)

- There is no associated target variable for observations.
- Instead, focus on analyzing the relationships between variables, e.g. cluster analysis (grouping observations based on similarities in their properties)
- E.g., K-Means, K-Medians

Machine Learning- Unsupervised Learning

Preset: Some info (attributes/features) that could help us differentiate between objects of interest.

Task-Clustering: Try to cluster the data/ group it / find different categories.



I have some info, like color, size, running speed...

Are these really different breeds? / Are these breeds really that different?

In general, there are three major learning categories on how algorithms can learn:

Supervised learning (Überwachtes Lernen)

- For each observation (e.g. person, age, interests) there is an associated target value (purchased products).
- Two problems:
 - Target value is continuous, e.g. quantity of products purchased → Regression, e.g. Regression Trees
 - Target value is categorical, e.g. which products purchased → Classification, e.g. Decision Trees

Unsupervised learning (Unüberwachtes Lernen)

- There is no associated target variable for observations.
- Instead, focus on analyzing the relationships between variables, e.g. cluster analysis (grouping observations based on similarities in their properties)
- E.g., K-Means, K-Medians

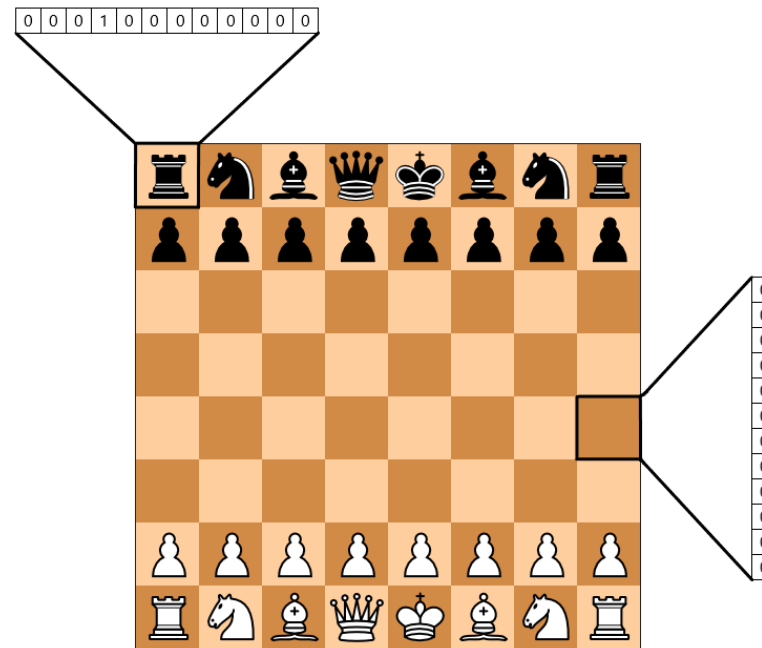
Reinforcement learning (Verstärkendes Lernen)

- Algorithm learns optimum strategy for given problem to maximize reward
- Translation of a situation in action
- E.g., Q-Learning

Machine Learning- Reinforcement Learning

Preset: Some info (attributes/features) about a problem environment.

Task: Find the best strategy to reach a certain goal with minimum effort.

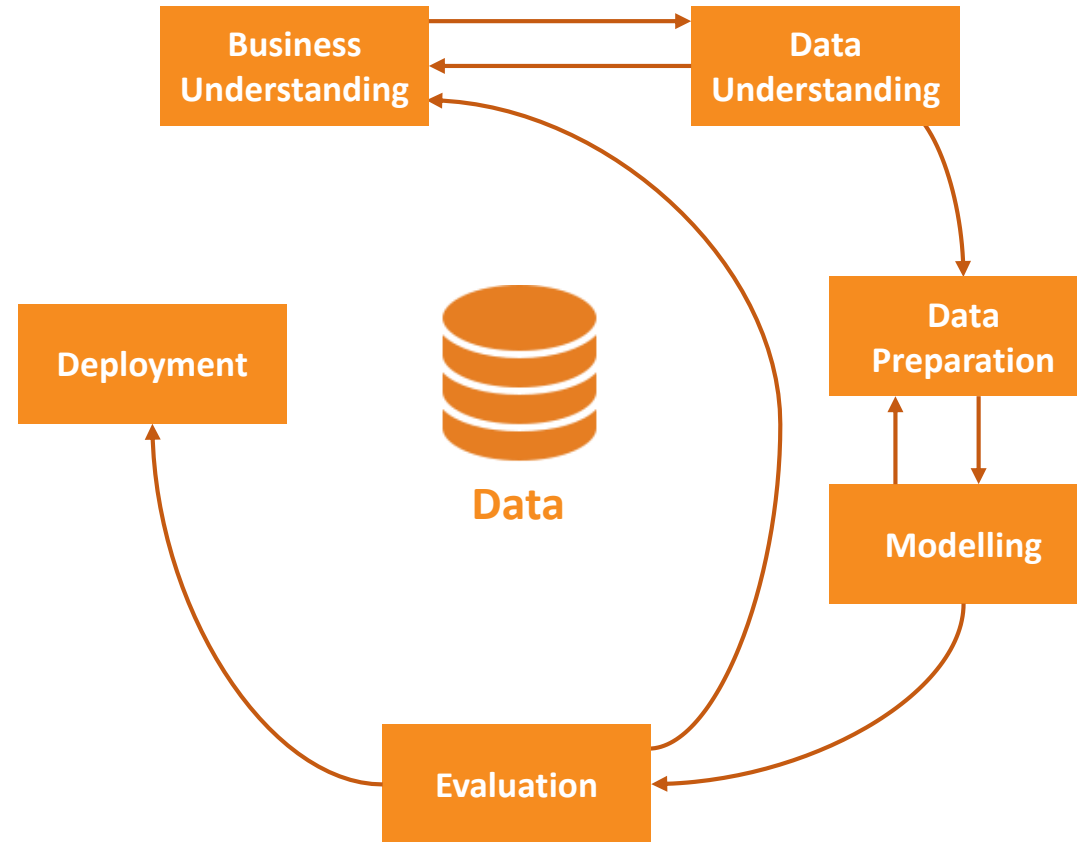


How can I win the game (against my opponent) ?

Table of Contents

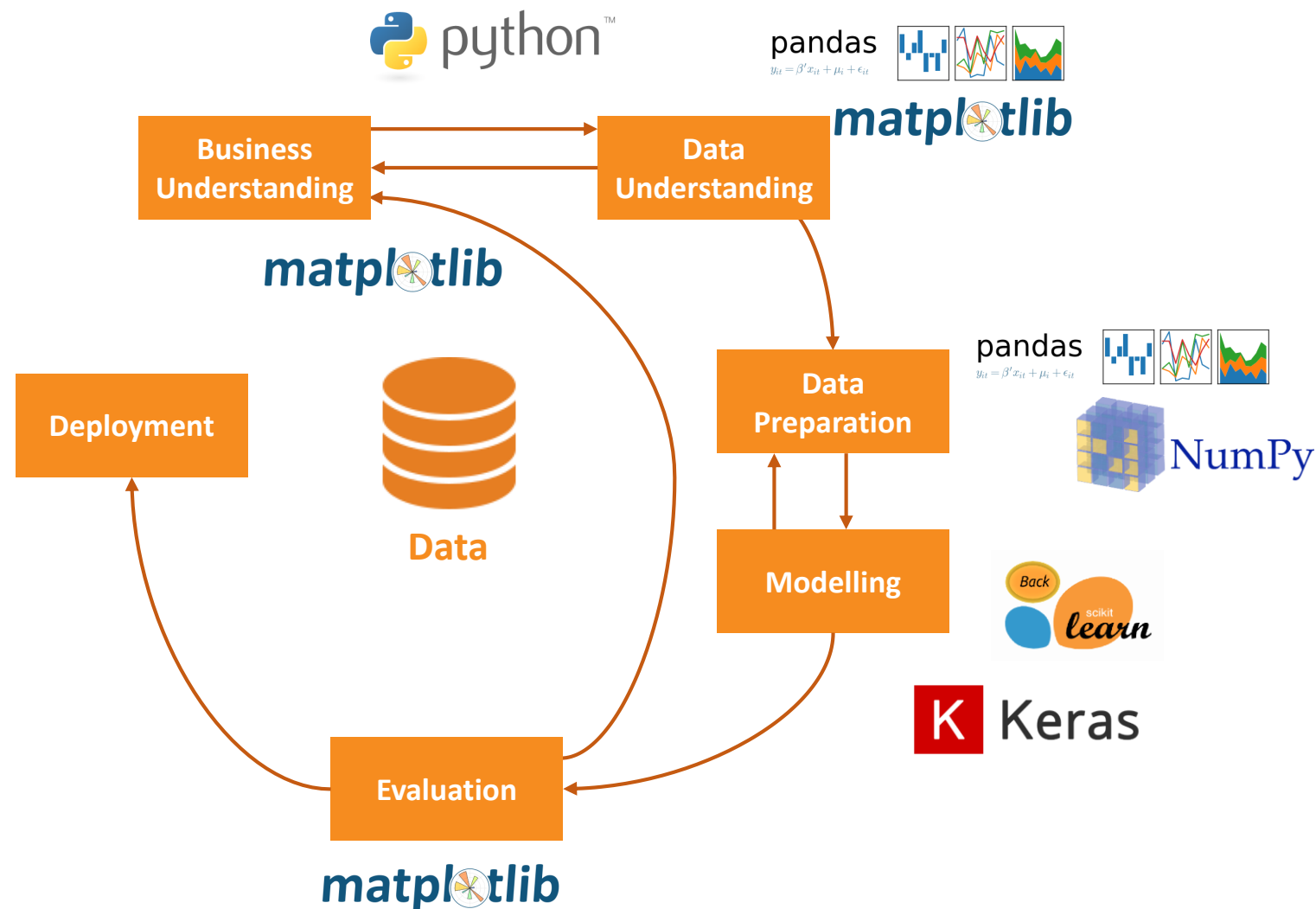
- 1. What is Data Science?**
- 2. Data Science Roles and Skills**
- 3. Data Science Process**
- 4. The first step: Asking the right Questions**
- 5. Step 2 and 2^{1/2} : A first glance at our dataset**

The Data Science Cycle: CRISP-DM

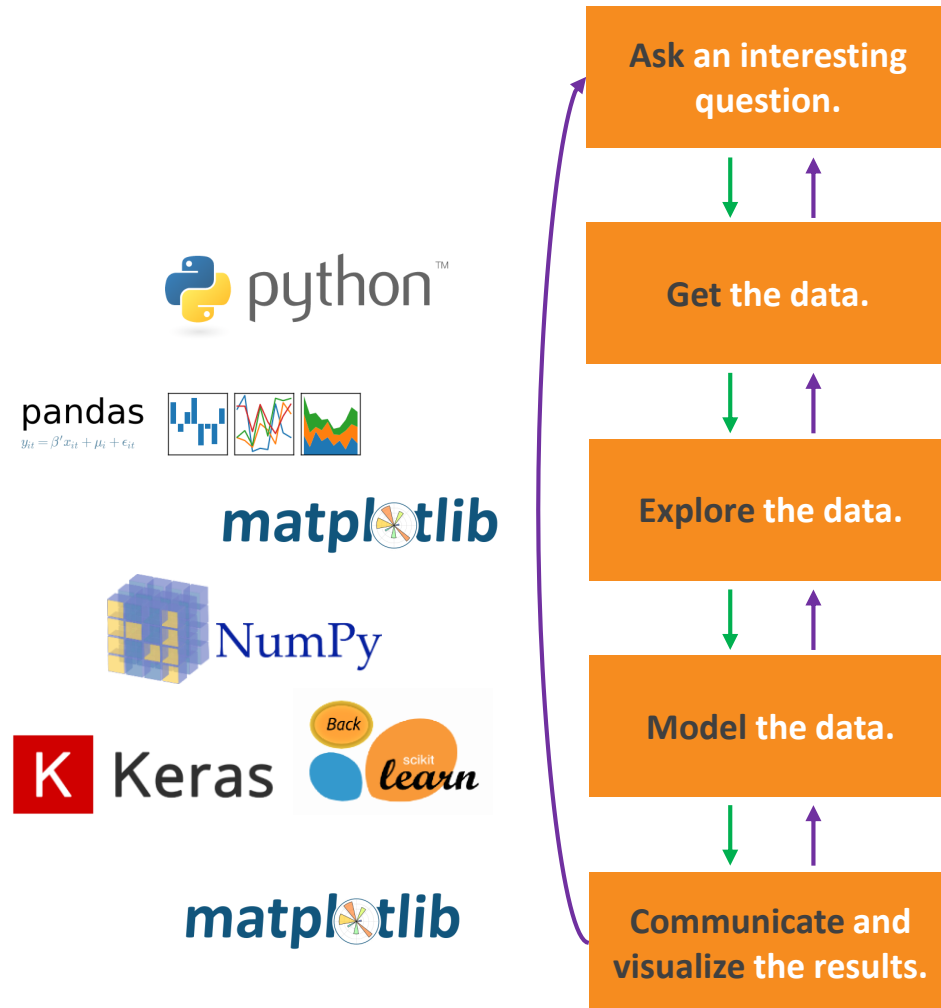


Cross Industry Standard Process for Data Mining

The Data Science Cycle: CRISP-DM



The Data Science Process



What is the **goal**?
 What would you do with the **data**?
 What do you want to **predict** or **estimate**?

How were the data **sampled**?
 Which data are **relevant**?
 Are there **privacy** issues?

Plot the data.
 Are there **anomalies**?
 Are there **patterns**?

Build a model.
Fit the model.
Validate the model.

What did we **learn**?
 Do the results make **sense**?
 Can we tell a **story**?

Table of Contents

- 1. What is Data Science?**
- 2. Data Science Roles and Skills**
- 3. Data Science Process**
- 4. The first step: Asking the right Questions**
- 5. Step 2 and 2^{1/2} : A first glance at our dataset**

Case Study: Lending Club



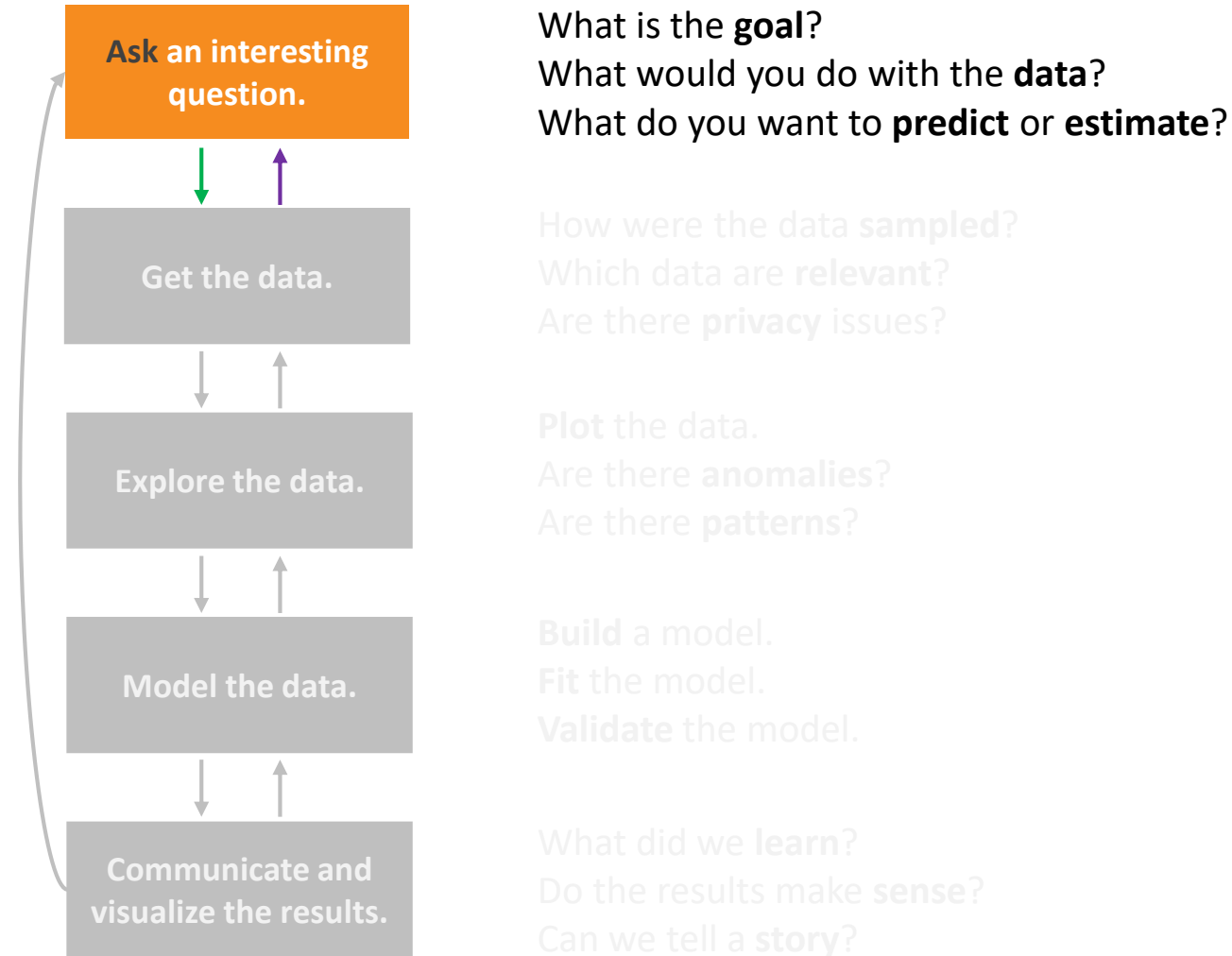
LendingClub is a US [peer-to-peer lending](#) company, headquartered in [San Francisco, California](#).^[3] It was the first peer-to-peer lender to register its offerings as [securities](#) with the [Securities and Exchange Commission](#) (SEC), and to offer loan trading on a secondary market. LendingClub is the world's largest peer-to-peer lending platform.^[4] The company claims that \$15.98 billion in loans had been originated through its platform up to December 31, 2015.^[5]

Source: Wikipedia

Your Role at LendingClub:

You are a data scientist at lending club, driving the platform efficiency and informing management decisions. Recently, your boss decided to leverage the power of the large amounts of data that are available.

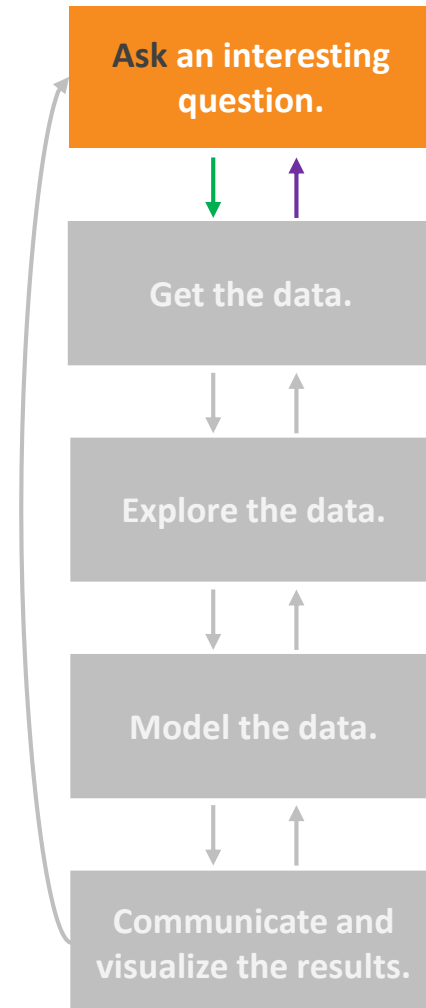
Step 1 – Asking the right questions



Step 1 – Asking the right questions

- During this day, we aim to solve two problems that Lending Club faces
 1. **Regression:** An interest rate is assigned to the debtor that has to make an interest payment to the creditor. We aim to compute the interest rate automatically for a given transaction by training a machine learning algorithm that uses data on past transactions as a training sample.
 2. **Classification:** Each debtor in the Lending Club database has a grade. To assign a grade to a new debtor, e.g., without prior information on his/her creditworthiness, we want to train a classifier that tell us the most probable grade of a new debtor.

Step 1 – Asking the right questions



How do we compute the informed interest rate automatically from a given datapoint?

What are the data we need to train a debt-class-classifier?

How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?

Plot the data.
Are there **anomalies**?
Are there **patterns**?

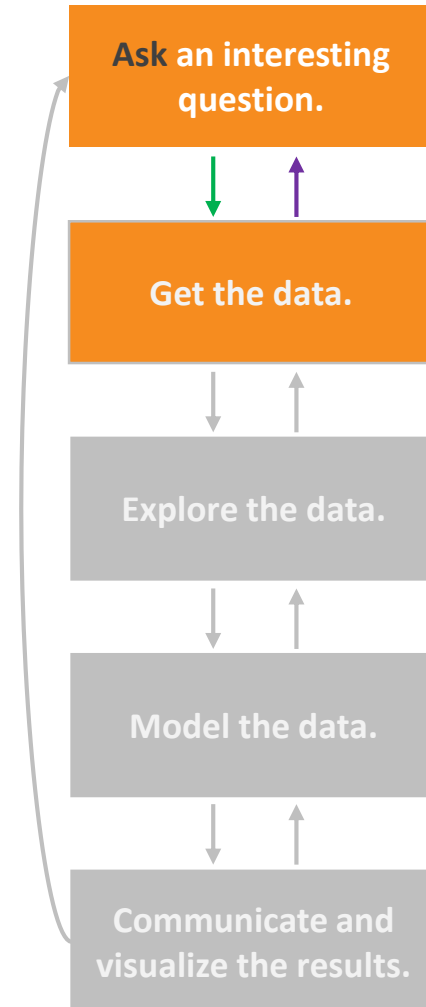
Build a model.
Fit the model.
Validate the model.

What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

Table of Contents

- 1. What is Data Science?**
- 2. Data Science Roles and Skills**
- 3. Data Science Process**
- 4. The first step: Asking the right Questions**
- 5. Step 2 and 2 ^{1/2} : A first glance at our dataset**

Step 2 – Get the data



How do we compute the interest rate automatically from a given datapoint?

What are the data we need to train a debt-class-classifier?

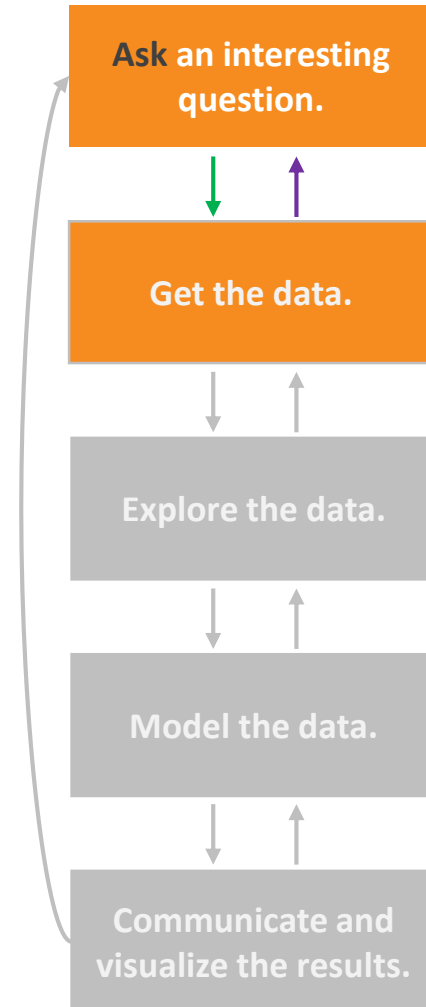
We got the data from our firm.
No privacy problems.
Let's go and explore!

Plot the data.
Are there **anomalies**?
Are there **patterns**?

Build a model.
Fit the model.
Validate the model.

What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

Step 2 – Get the data



How do we compute the interest rate automatically from a given datapoint?

What are the data we need to train a debt-class-classifier?

We got the data from our firm.

No privacy problems.

Let's go and explore!

Plot the data.

Are there **anomalies**?

Are there **patterns**?

Build a model.

Fit the model.

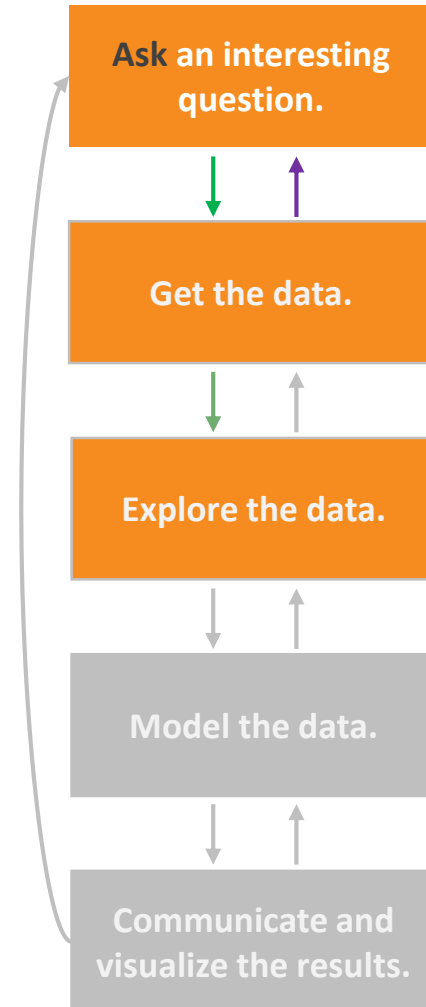
Validate the model.

What did we **learn**?

Do the results make **sense**?

Can we tell a **story**?

Step 2 – Get the data



How do we compute the interest rate automatically from a given datapoint?

What are the data we need to train a debt-class-classifier?

We got the data from our firm.

No privacy problems.

Let's go and explore!

Plot the data.

Are there **anomalies**?

Are there **patterns**?

Build a model.

Fit the model.

Validate the model.

What did we **learn**?

Do the results make **sense**?

Can we tell a **story**?