Introduction to Data Science

Part 1: About Data Science

Dr. Nicolas Pfeuffer

Table of Contents

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

efl | the Data Science Institute

https://www.informs.org/ORMS-Today/Public-Articles/February-Volume-40-Number-1/Analytics-key-to-Obama-s-victory
https://www.cargill.com/2018/cargills-iqshrimp-helps-farmers-manage-risk
https://www.zdnet.com/article/nike-acquires-celect-adds-to-data-science-team/
https://www.forbes.com/sites/jacknera62/2019/08/27/data-science-will-drive-auto-industry-in-future-tata-consulting-says/#297ab8bb1c7f
https://www.ft.com/content/d9b3511c-90da-11e8-b639-7680cedcc421

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."

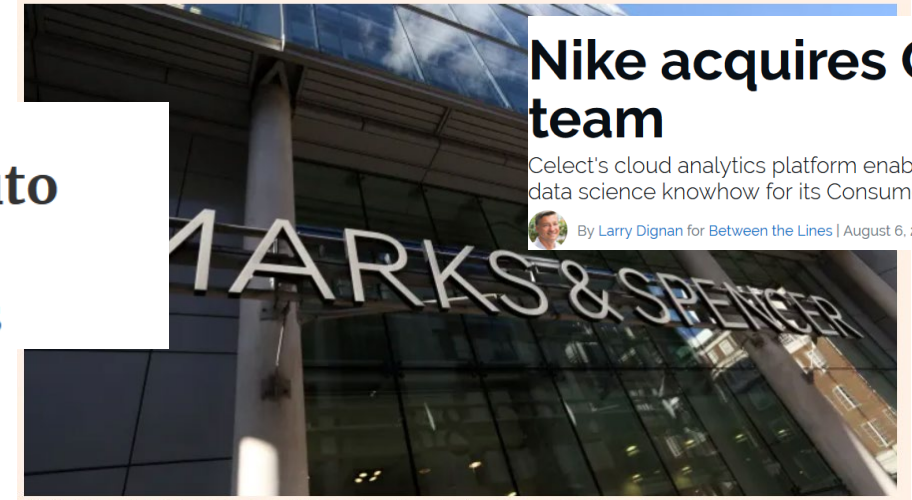
Cargill's iQShrimp helps farmers manage risk, make better decisions

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

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

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

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

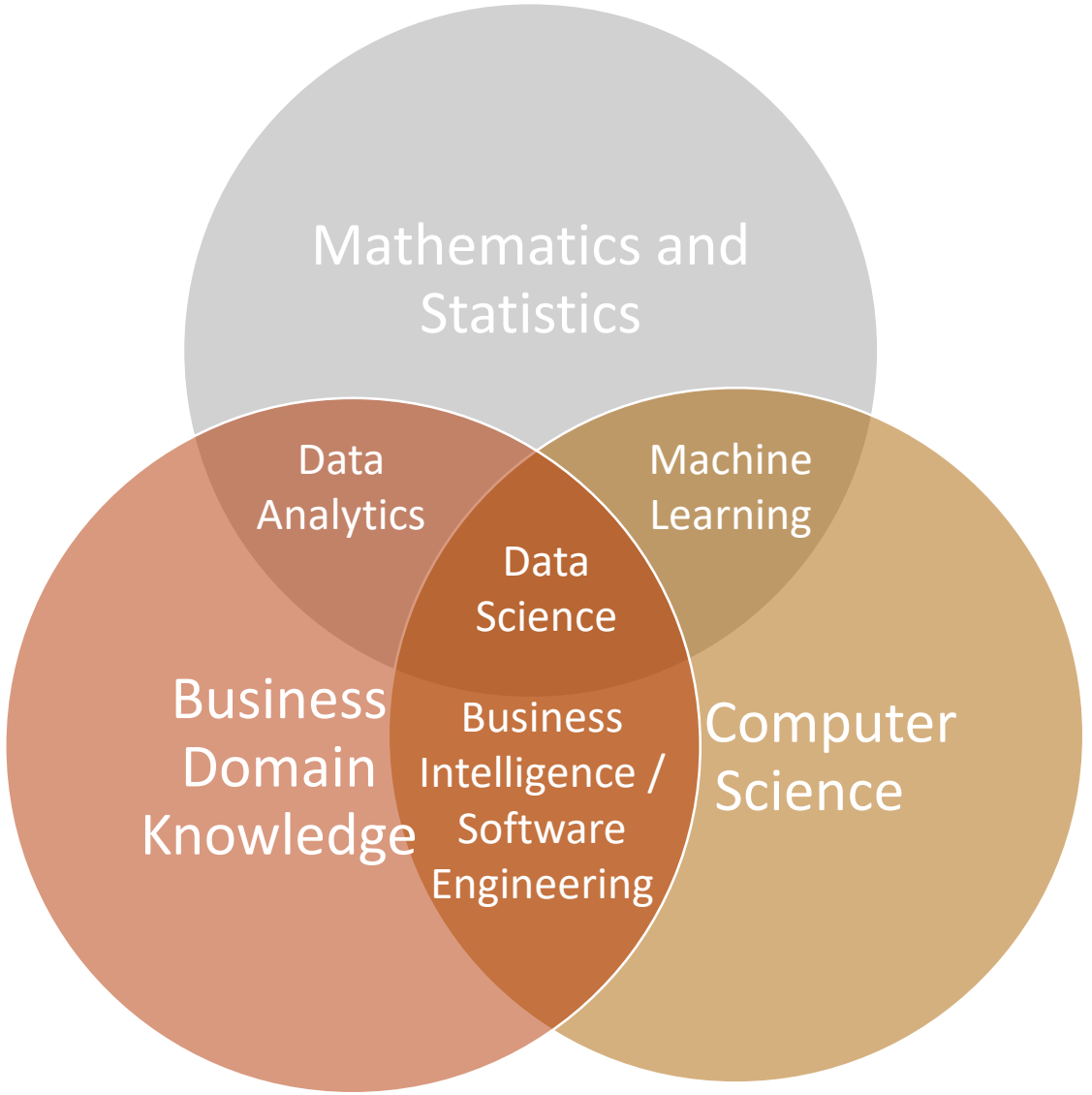
Data science is the **study of the generalizable extraction of knowledge from data**.

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

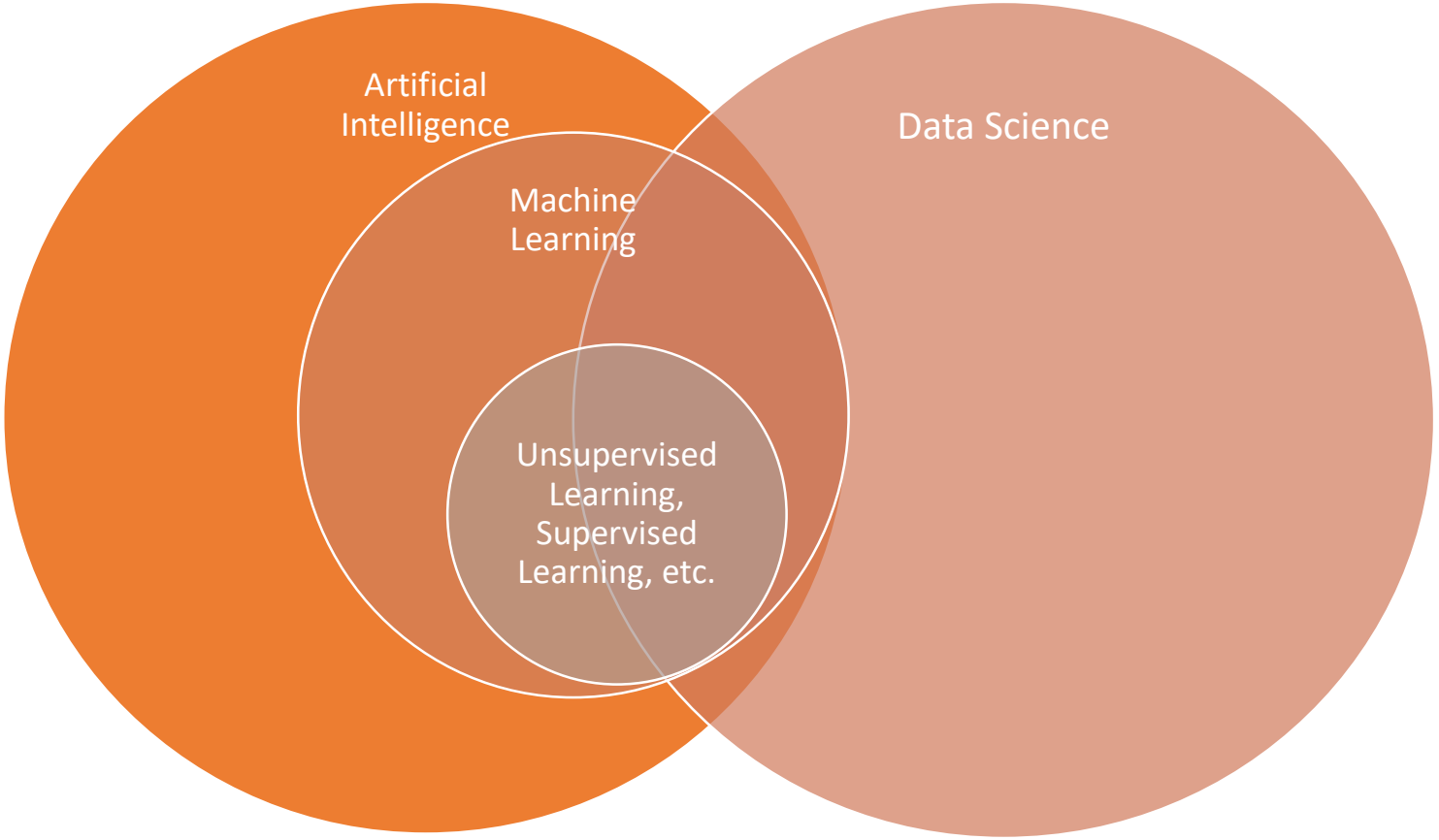
[…]Data Science is the **science of learning from data**; it studies the **methods involved in the analysis** and **processing of data** and **proposes technology** to **improve methods in an evidence-based manner**. The scope and impact of this science will expand enormously in coming decades as scientific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



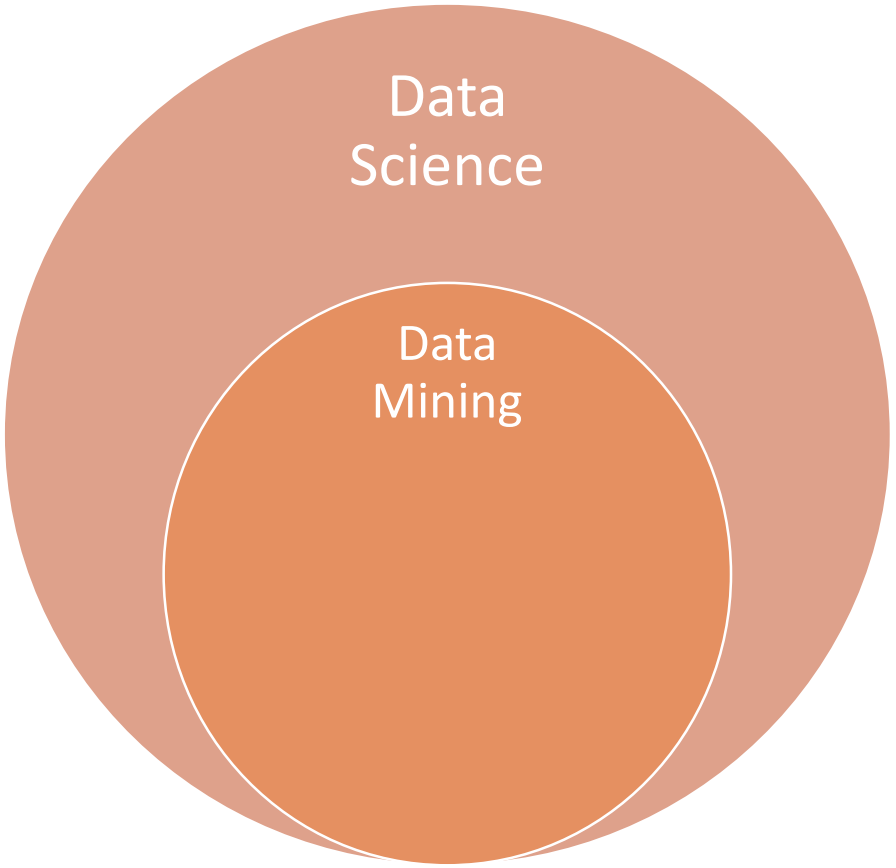
Data Mining

Data mining, [... is] the process of discovering interesting and useful patterns and relationships in large volumes of data.

The field combines tools from statistics and artificial intelligence (such as neural networks and machine learning) with database management to analyze large digital collections, known as data sets.

Clifton, Christopher (2019). "Encyclopædia Britannica: Definition of Data Mining". Retrieved 05.11.2020

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

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."

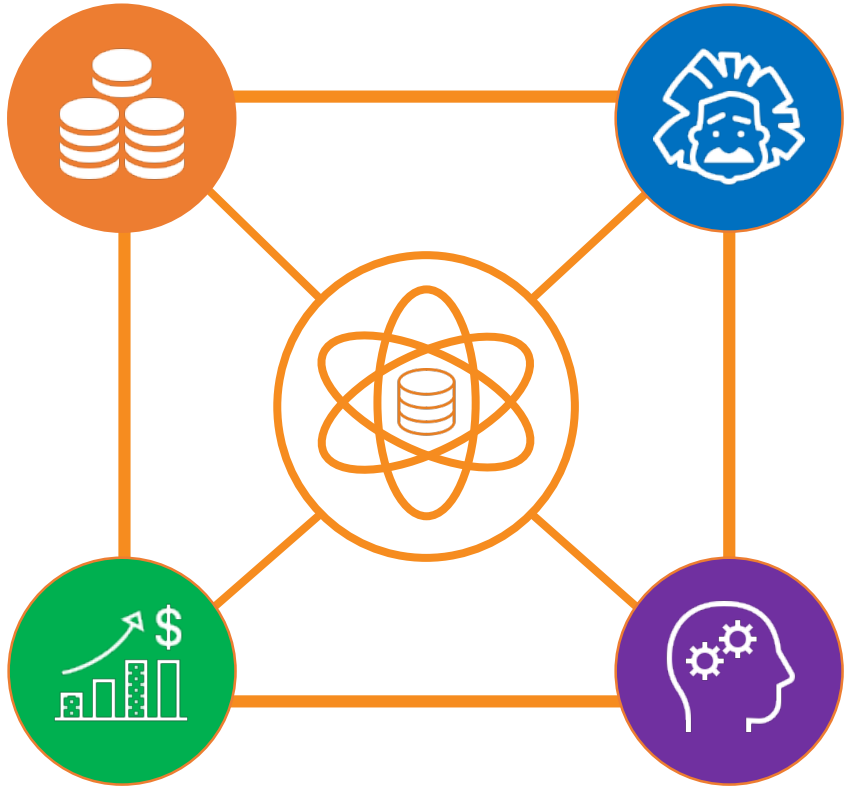
Joel Grus, in his book "Data Science from Scratch". O'Reilly, 2015.

Typical Roles in Data Science

Data Engineer

→ Responsible for the acquisition, organization and storage of data

→ Possible skills: Software Engineering, SQL, Hadoop, Python(Pandas)



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.

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.

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)
```

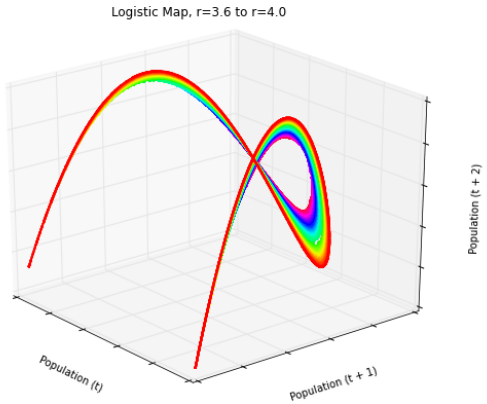
Functions

```
# 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')
```

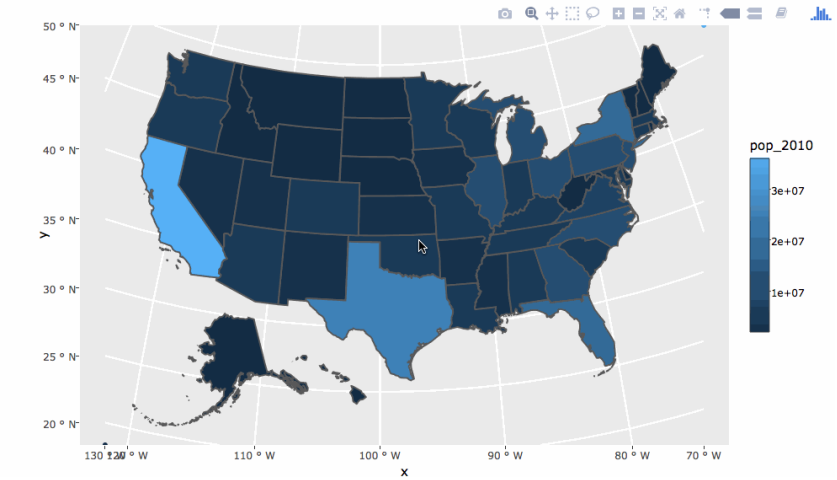
Basic Libraries

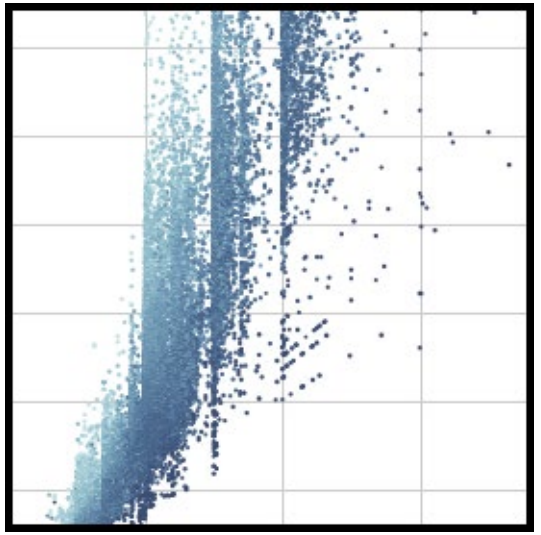
```
# os

import os

for i in range(10):
    os.makedir("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



matplotlib

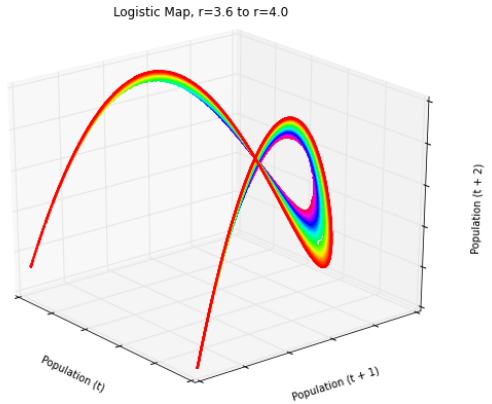




seaborn

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

Skills ..: Data Visualization



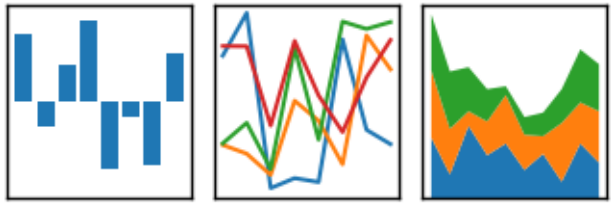
- Build affordances that prioritize data exploration and comparison.
- Emphasize clarity and transparency.
- Reduce cognitive load and focus on what matters.
- Experiences should be intuitive and easy to use.
- Provide context and help users navigate the data.
- Consider performance, polish, surprise,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

https://material.io/design/communication/data-visualization.html

Skills …: Data Wrangling, Statistics and Calculus

pandas

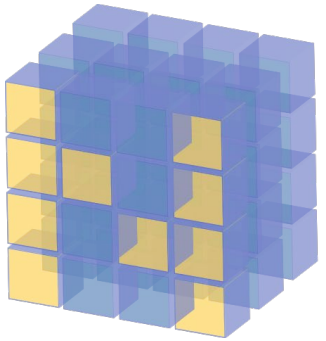
$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



```
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:

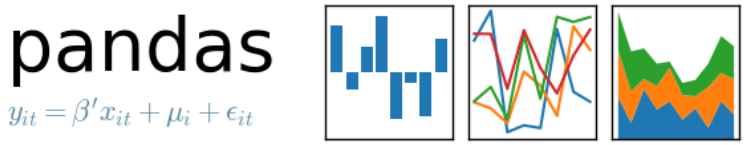
- the index of the DataFrame is the same as the index of the CARTO table (`cartodb_id`)
- `the_geom` column stores the geometry. This can be decoded if we set the



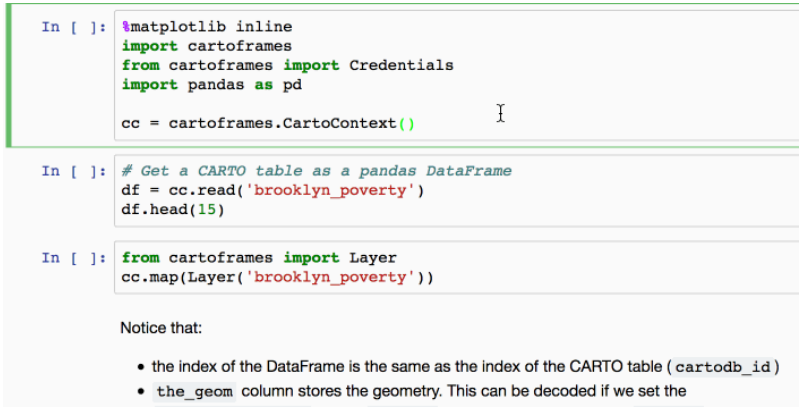
NumPy

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

Skills …: Data Wrangling, Statistics and Calculus



- 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/

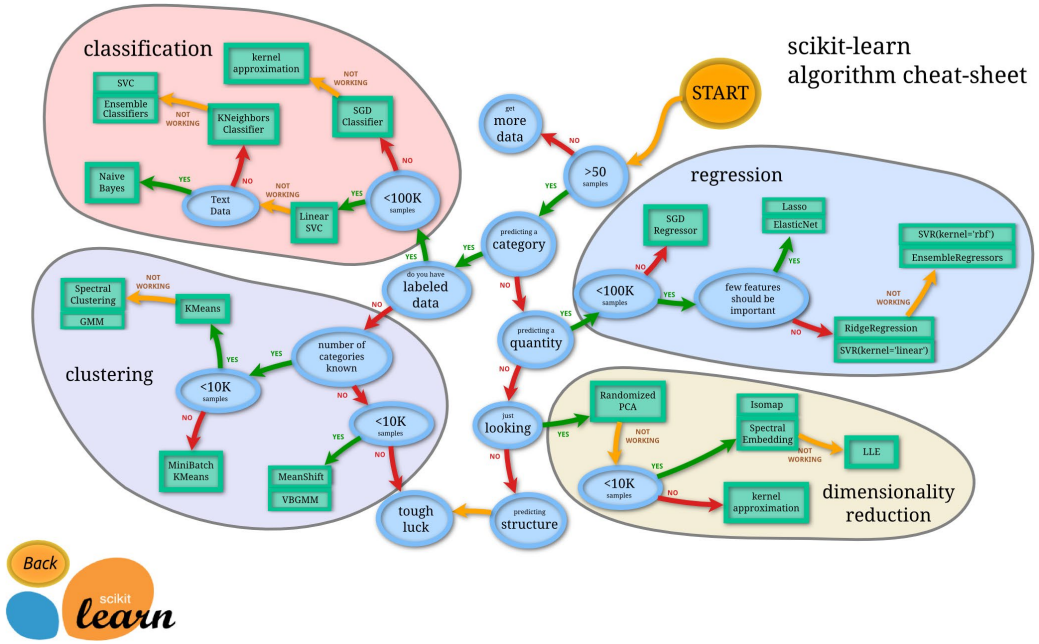
Skills …: Inference, Machine Learning and Deep Learning

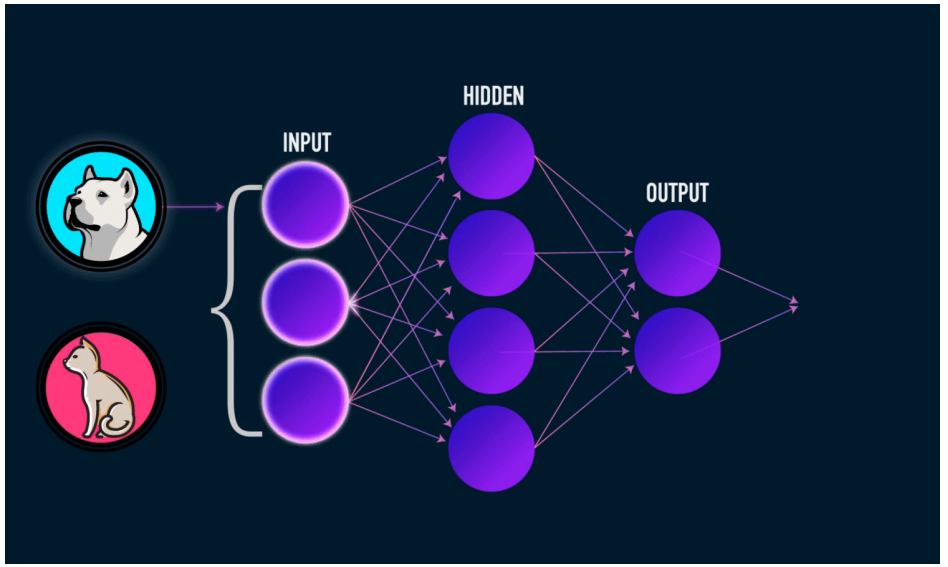
NumPy

TensorFlow

Keras

PyTorch





https://towardsdatascience.com/an-introduction-to-deep-learning-af63448c122c

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

James, Gareth, et al. *An introduction to statistical learning*. Vol. 112. New York: springer, 2013.
Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.
Sutton, Richard S., et al. *Introduction to reinforcement learning*. Cambridge: MIT press, 1998.
Upton, G., Cook, I. (2008) Oxford Dictionary of Statistics, OUP. ISBN 978-0-19-954145-4

Machine Learning

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

James, Gareth, et al. *An introduction to statistical learning*. Vol. 112. New York: springer, 2013.
Sutton, Richard S., et al. *Introduction to reinforcement learning*. Cambridge: MIT press, 1998.

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 and muffins. Which one is it?

https://www.freecodecamp.org/news/chihuahua-or-muffin-my-search-for-the-best-computer-vision-api-cbda4d6b425d/

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?

https://www.google.com/search?q=weather

Machine Learning

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

James, Gareth, et al. *An introduction to statistical learning*. Vol. 112. New York: springer, 2013.
Sutton, Richard S., et al. *Introduction to reinforcement learning*. Cambridge: MIT press, 1998.

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?

https://esadoctors.com/how-many-emotional-support-animals-can-you-have/

Machine Learning

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)

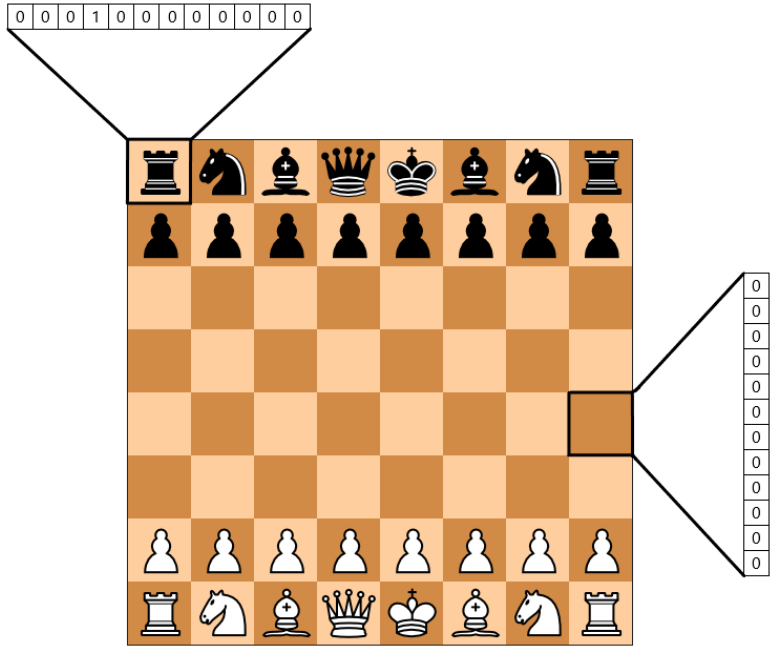
- Algorithm learns optimum strategy for given problem to maximize reward
- Translation of a situation in action
- E.g., Q-Learning

James, Gareth, et al. *An introduction to statistical learning*. Vol. 112. New York: springer, 2013.
Sutton, Richard S., et al. *Introduction to reinforcement learning*. Cambridge: MIT press, 1998.

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.

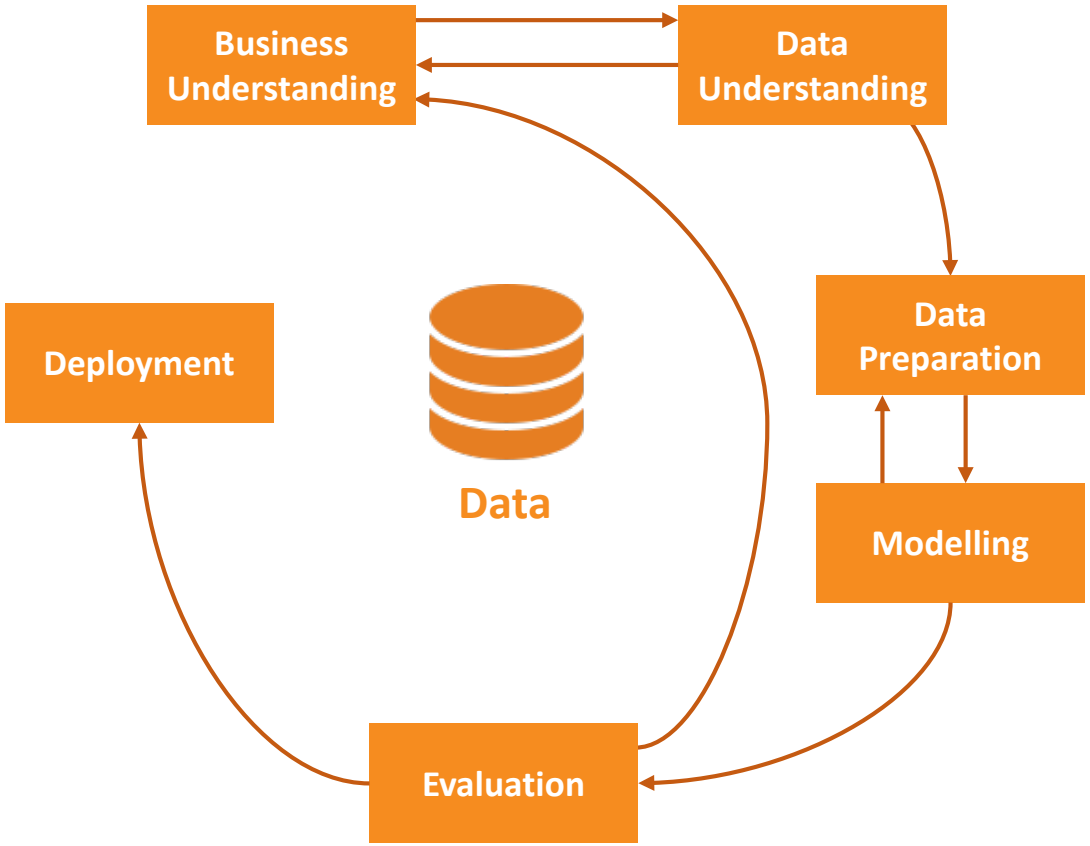


How can I win the game (against my opponent) ?

https://towardsdatascience.com/predicting-professional-players-chess-moves-with-deep-learning-9de6e305109e

Table of Contents

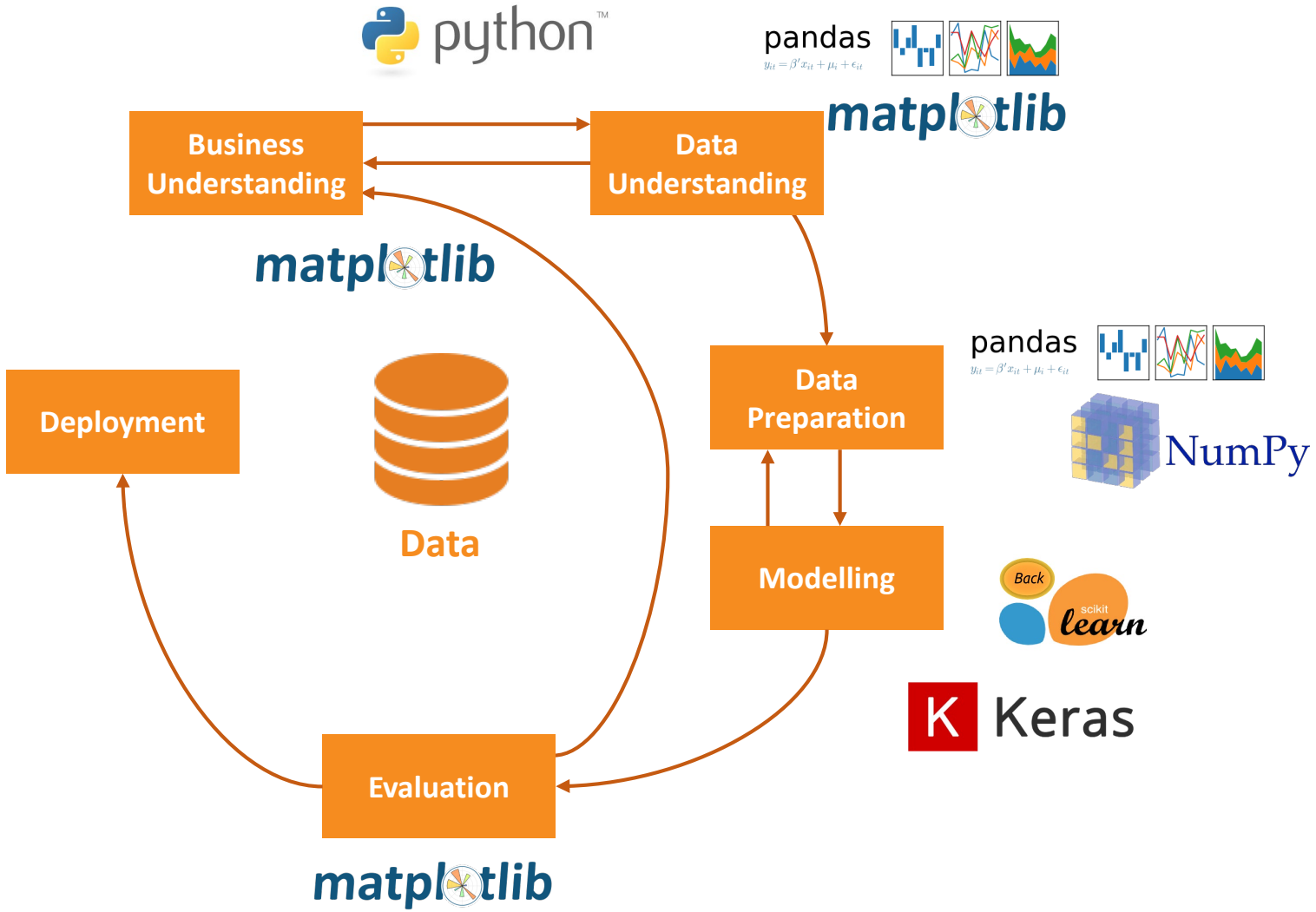
The Data Science Cycle: CRISP-DM



Cross Industry Standard Process for Data Mining

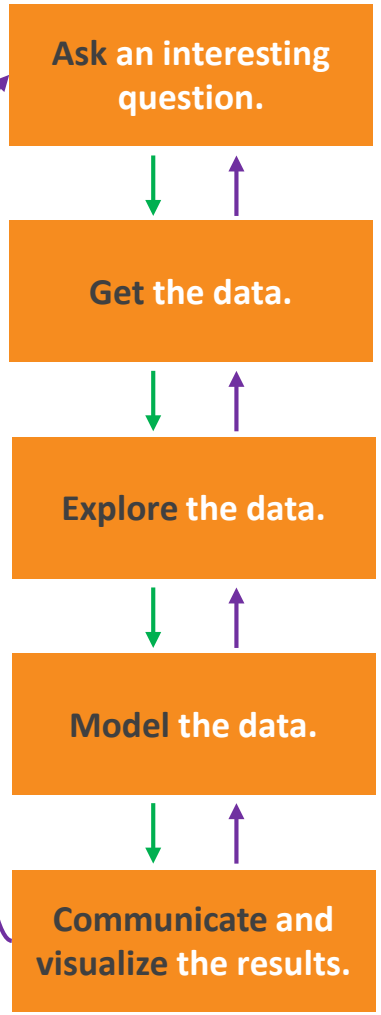
Shearer C., *The CRISP-DM model: the new blueprint for data mining*, J Data Warehousing (2000); 5:13—22.

The Data Science Cycle: CRISP-DM



Shearer C., *The CRISP-DM model: the new blueprint for data mining*, J Data Warehousing (2000); 5:13—22.

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**?

Graphic based on the idea and visualization of the data science process by Pfister, Blitzstein and Kaynig for their course „CS109 Data Science“ at Harvard John A. Paulson School of Engineering

Table of Contents

Case Study: Lending Club

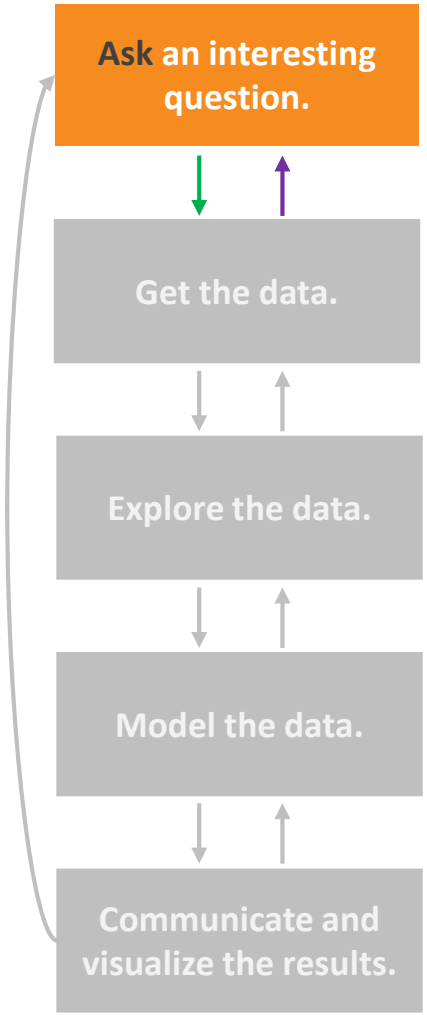
LendingClub is a US peer-to-peer lending company, headquartered in San Francisco, California.[3] It was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC), and to offer loan trading on a secondary market. LendingClub is the world's 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



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.

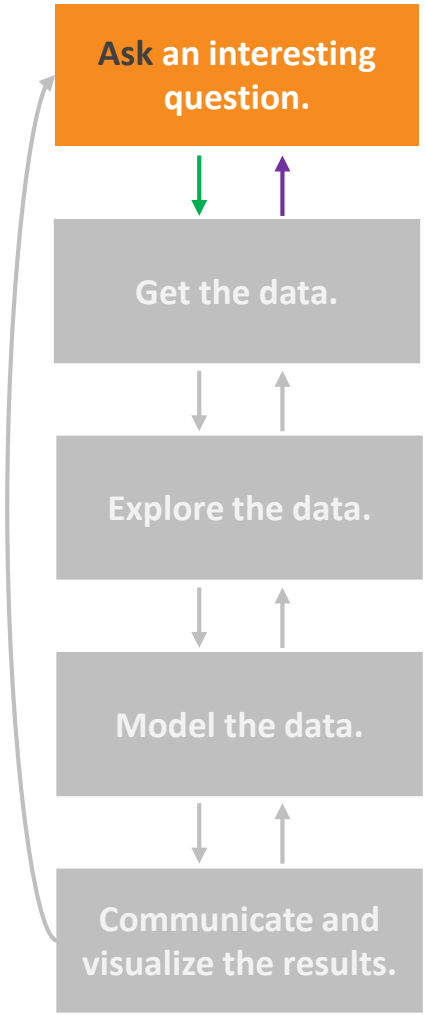
What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

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?

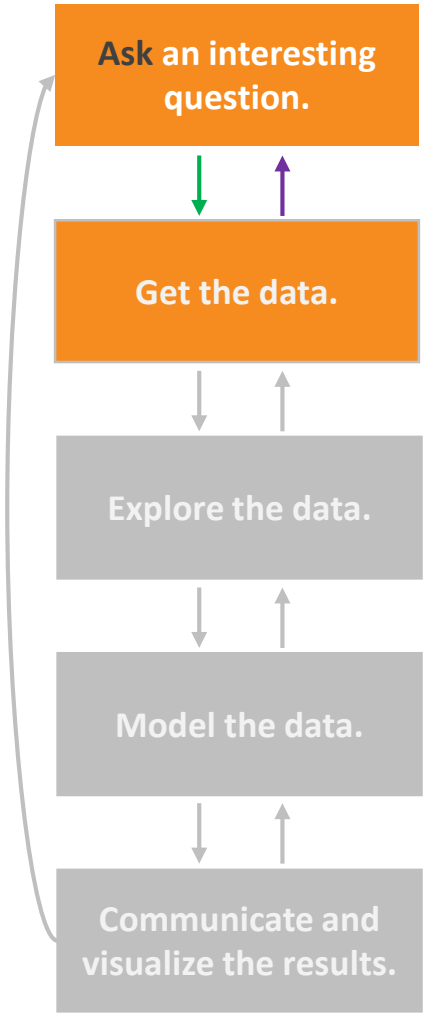
Plot the data.
Are there **anomalies**?
Are there **patterns**?

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

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?

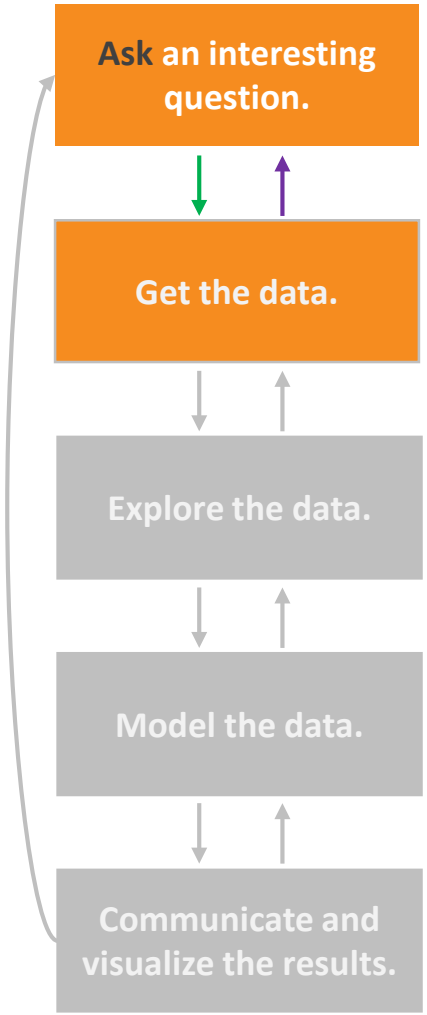
We got the data from our firm.
No privacy problems.
Let's go and explore!

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?

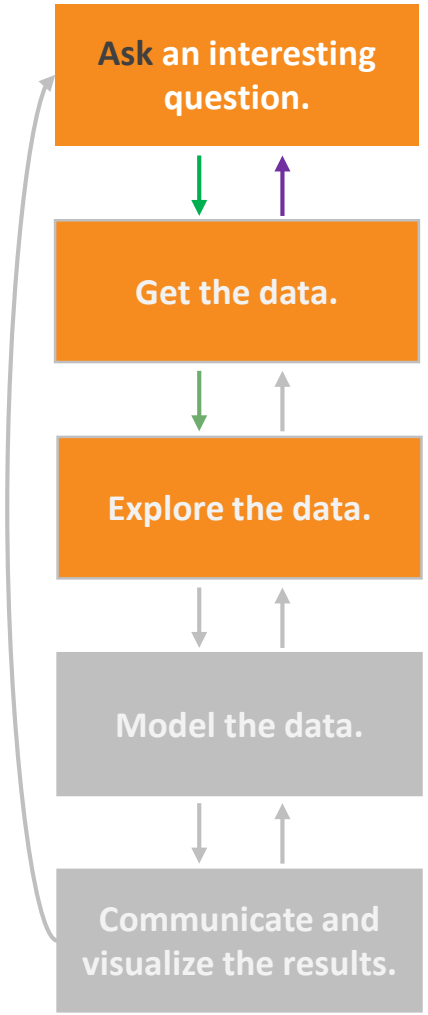
We got the data from our firm.
No privacy problems.
Let‘s go and explore!

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 3 – Explore the data (Part 1)



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**?