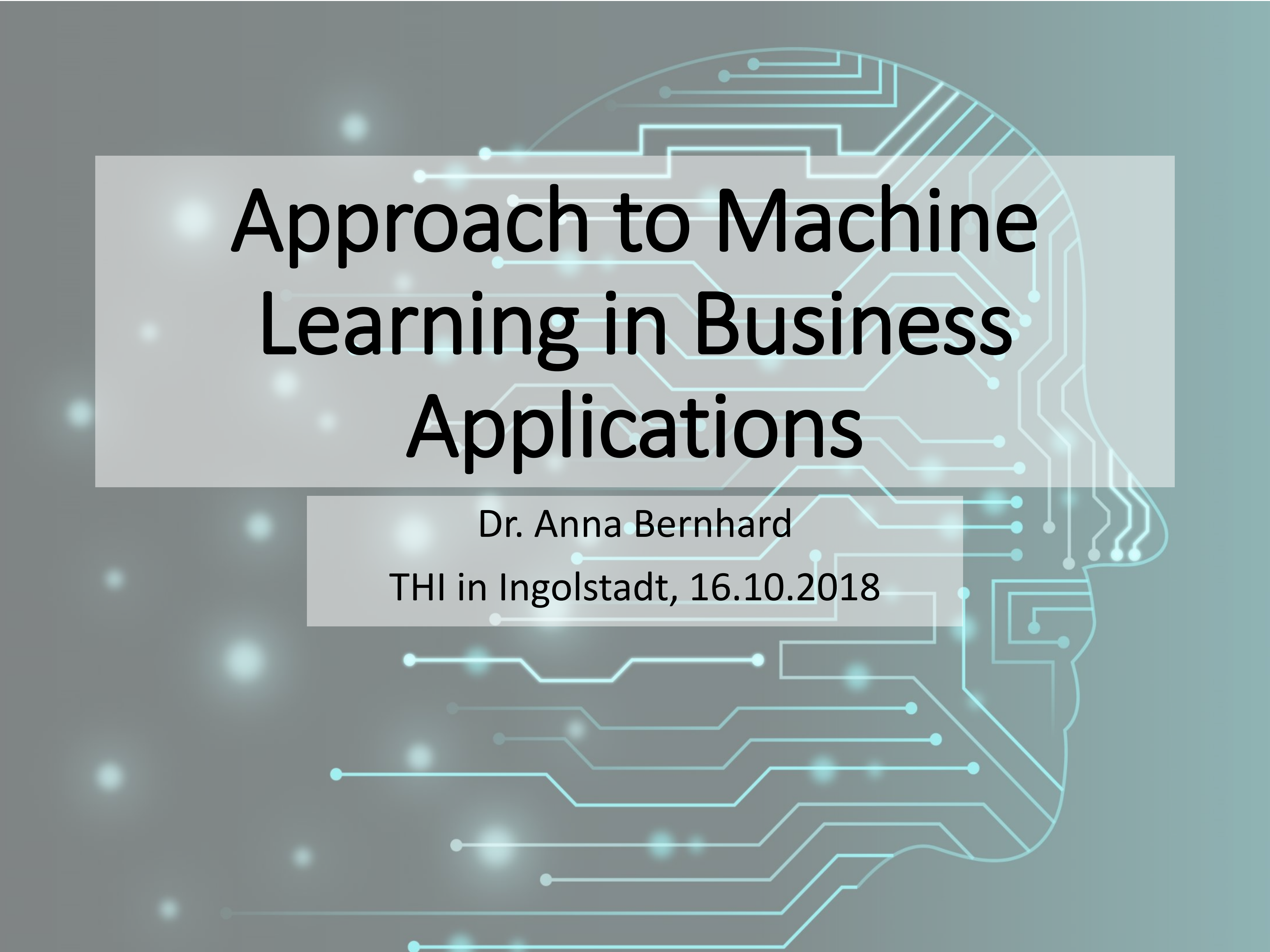


Approach to Machine Learning in Business Applications

Dr. Anna Bernhard

THI in Ingolstadt, 16.10.2018

Learning goals

- ✓ You know the **different steps of a machine learning project**
- ✓ You can bring the **steps** of such a project **in the right order** and **give examples** for each step
- ✓ You know the difference between **supervised** and **unsupervised learning** algorithms

Exercise: How would you start?

Bring these steps in the right order!

Model data

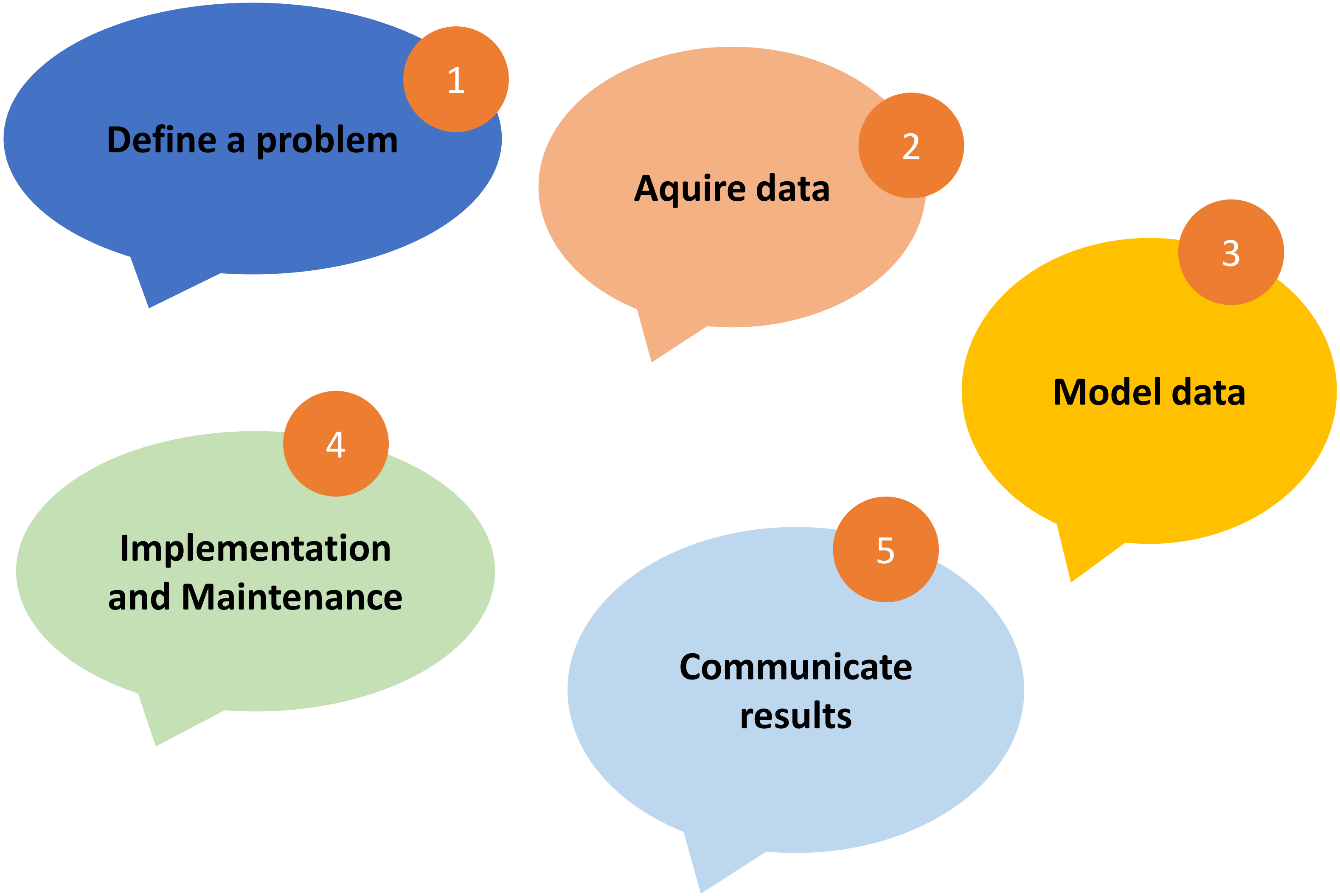
Communicate results

Define a problem

Implementation and Maintenance

Aquire data

Exercise: How would you start?



1 Define a problem

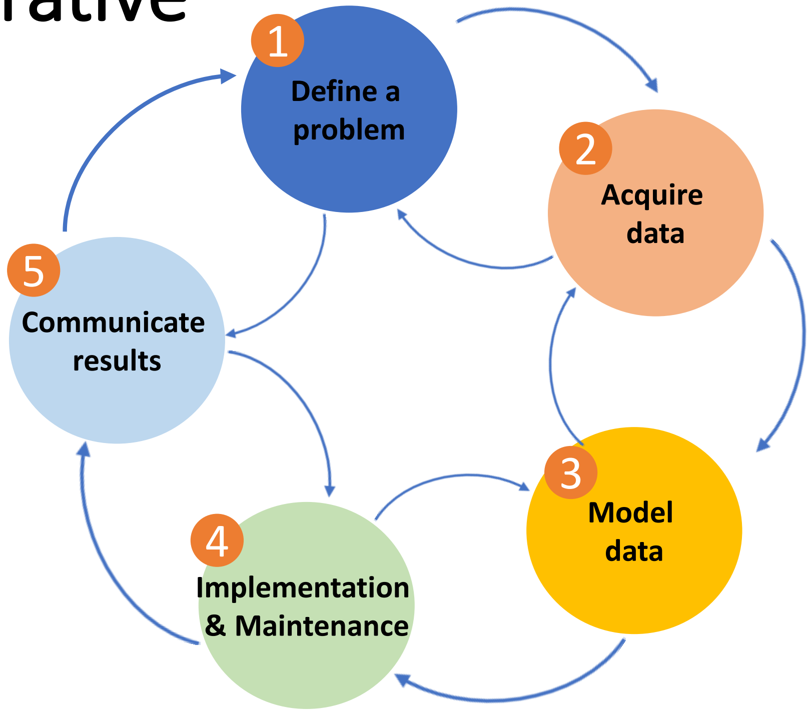
2 Aquire data

3 Model data

4 Implementation and Maintenance

5 Communicate results

A Project Lifecycle is iterative



A typical machine learning project should follow these steps

1. Define a problem
2. Acquire Data and explore data
3. Model data with ML algorithm
4. Implementation and Maintenance
5. Communicate results

Scale is central to the iterative approach

- Validate an approach with a small sample of data and use a large amount of data to refine a proven solution (POC, Proof of Concept)

The approach will be explained based on an example of a fictional booking platform that wants to create flexible pricing

stayhere

1

Define the Problem

The process begins by specifying a problem

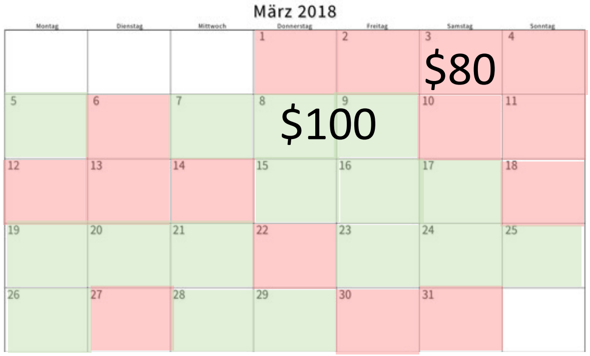
This is often directly related to revenue or costs:

- “People browse our site but don’t buy anything”
- “Subscribers aren’t renewing their service”
- “Our employees spend too much time searching for documents”

=> Translate a business case into a mathematical problem

1

Example: stayhere



März 2018

Montag	Dienstag	Mittwoch	Donnerstag	Freitag	Samstag	Sonntag
			1 🟥	2 🟥	3 🟥	4 🟥
5 🟩	6 🟥	7 🟩	8 🟩	9 🟩	10 🟥	11 🟥
12 🟥	13 🟥	14 🟥	15 🟩	16 🟩	17 🟩	18 🟥
19 🟩	20 🟩	21 🟩	22 🟥	23 🟩	24 🟩	25 🟩
26 🟩	27 🟥	28 🟩	29 🟩	30 🟥	31 🟥	

Problem: „There are too many unbooked nights“
Desired Outcome: reduce unbooked nights by 30%
Possible Solution: offer unbooked nights for a lower price

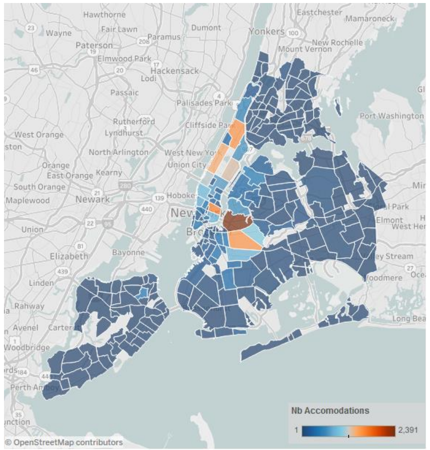
- „smart pricing model: suggest prices depending on seasonal variations“

2 Acquire and explore Data

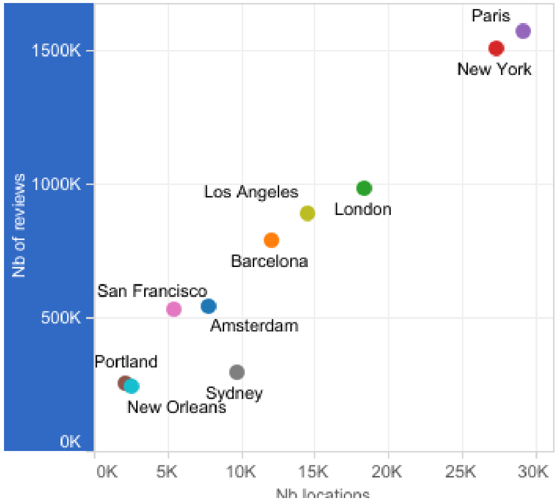
Approach:

- Collect all of the relevant data and assess quality
- Prepare and clean your data (e.g. filter out wrong data..)
- Getting it into a format suitable for analysis, most likely into a flat file format such as a .csv or in a database
- Do an exploratory data analysis (statistical overview) e.g. scatter plot, histograms…

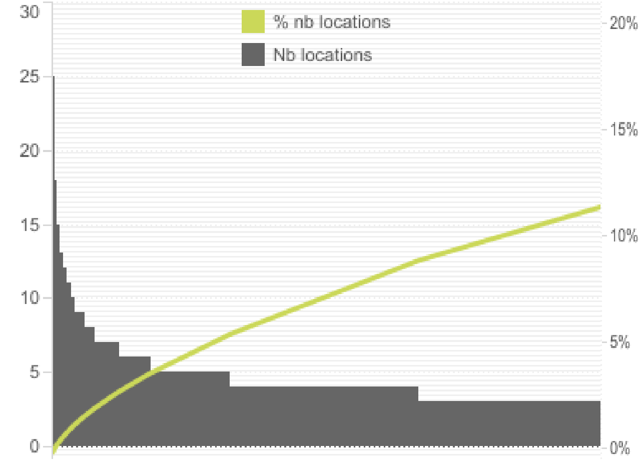
2 Example: stayhere



City by nb of reviews and nb locations



Distribution of locations by hosts



Source: Jonathan Trajkovic https://public.tableau.com/en-us/s/blog/2015/07/analyzing-airbnb-data

3

Model Data – two categories

Approach:

- Determine your target variable, the factor of which you are trying to gain deeper understanding.
- Define your ML method and language/tools (R, python…)
- Start with a POC (Proof of Concept) on a small amount of data
- Split your data sample into a test and training set for cross-validation

Supervised learning:

- The computer is presented with **example inputs** and their desired outputs, given by a "teacher“
- The goal is to learn a general rule that maps **inputs to outputs**.

Example: genre categorization of films

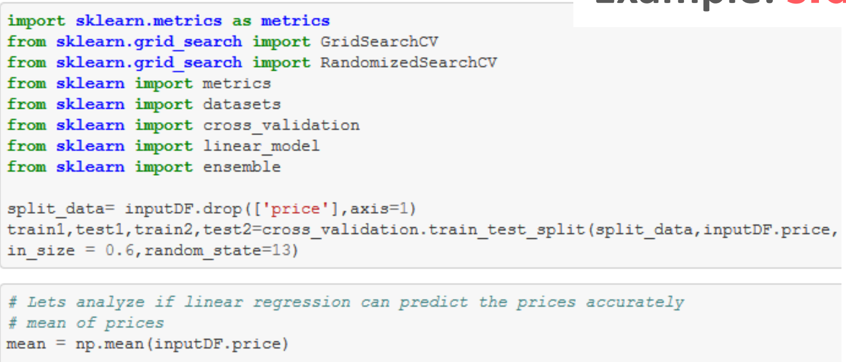
Unsupervised learning:

- **No labels** are given to the learning algorithm, leaving it **on its own** to find structure in its input.
- Unsupervised learning can be a goal in itself (discovering **hidden patterns** in data) or towards a specific outcome.

Example: customer segmentation

3

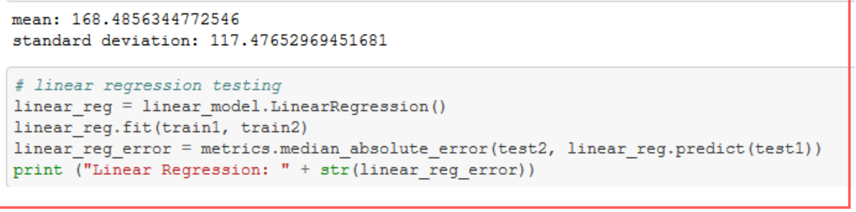
Example: stayhere



```
import sklearn.metrics as metrics
from sklearn.grid_search import GridSearchCV
from sklearn.grid_search import RandomizedSearchCV
from sklearn import metrics
from sklearn import datasets
from sklearn import cross_validation
from sklearn import linear_model
from sklearn import ensemble

split_data= inputDF.drop(['price'],axis=1)
train1,test1,train2,test2=cross_validation.train_test_split(split_data,inputDF.price,
in_size = 0.6,random_state=13)
```

```
# Lets analyze if linear regression can predict the prices accurately
# mean of prices
mean = np.mean(inputDF.price)
```



```
# standard deviation to compare
std = np.std(inputDF.price)

print("mean: " + str(mean))
print ("standard deviation: " + str(std))
```

```
mean: 168.4856344772546
standard deviation: 117.47652969451681
```

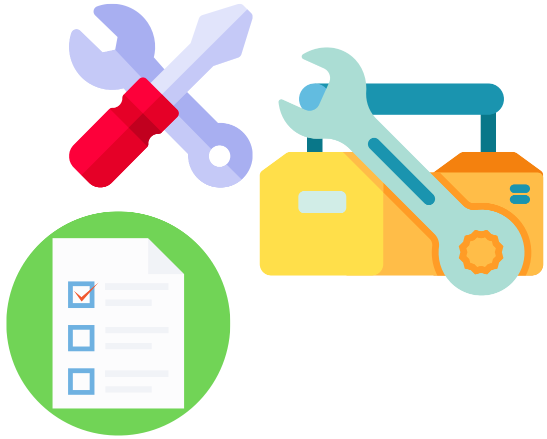
```
# linear regression testing
linear_reg = linear_model.LinearRegression()
linear_reg.fit(train1, train2)
linear_reg_error = metrics.median_absolute_error(test2, linear_reg.predict(test1))
print ("Linear Regression: " + str(linear_reg_error))
```

Code of supervised learning method „linear regression“

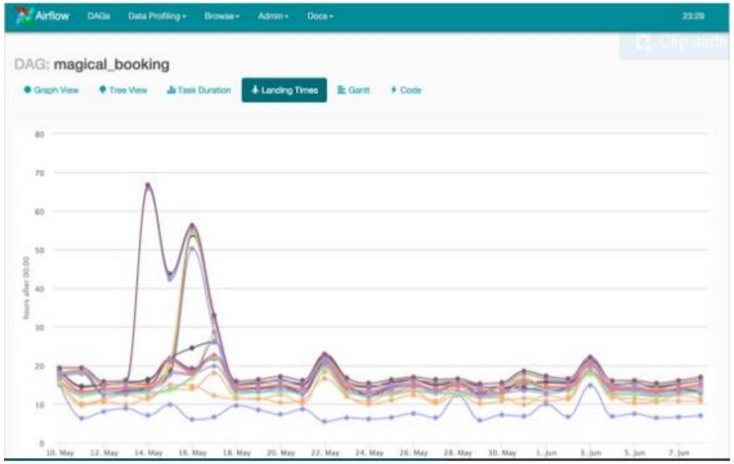
4 Implementation and Maintenance

Approach:

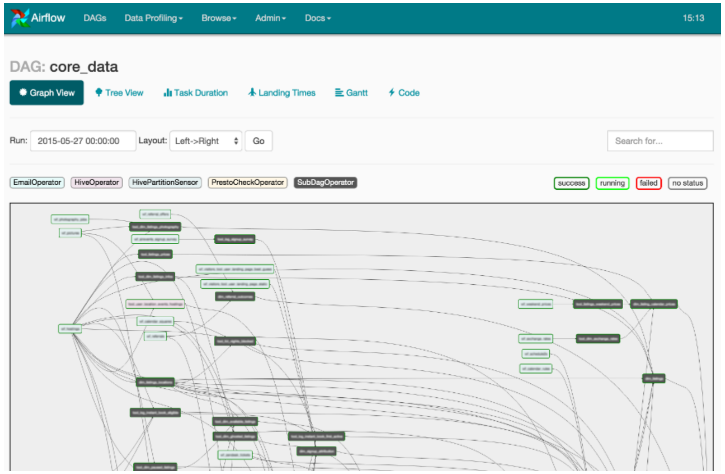
- Set up API (Application Programming Interface) system with an automated workflow
- Document modelling process for reproducibility
- Write tests for the code
- Create model for monitoring/logging and maintenance



4 Example: stayhere



DAG: magical_booking



DAG: core_data

Run: 2015-05-27 00:00:00 Layout: Left->Right

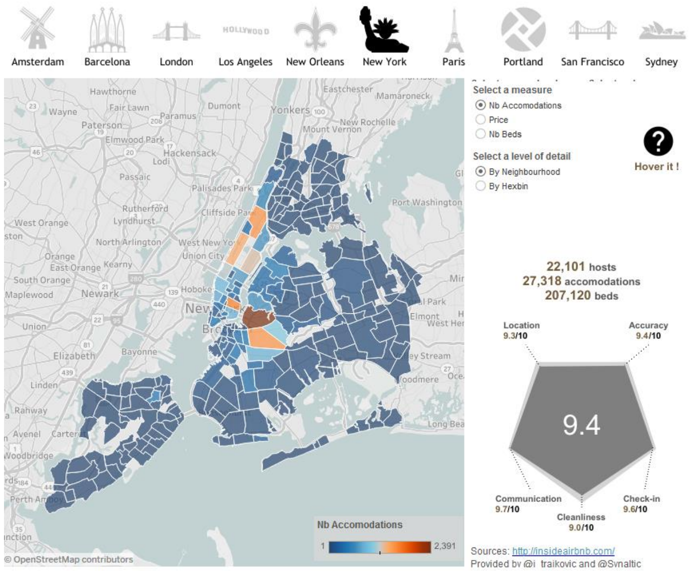
Source: Data Works Summit, Airflow - An Open Source Platform to Author and Monitor Data Pipelines

5 Communicate results

Approach:

- Communication is an essential part of the process
- Create meaningful visualizations that represent the data
- Dashboards are a common tool for communicating results
    - Good for Statistics, Summaries, Visualizations
- Get Customer feedback for further iterations

5 Example: stayhere

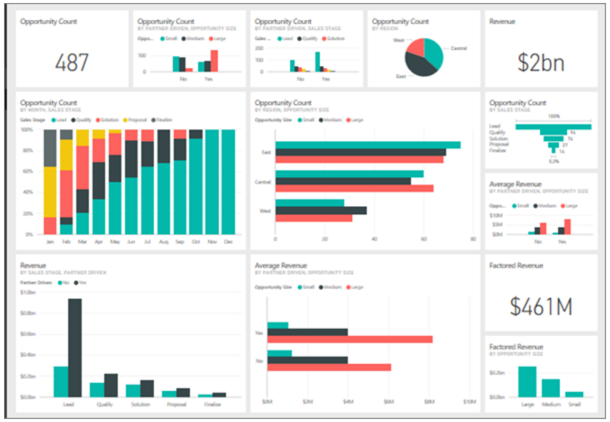


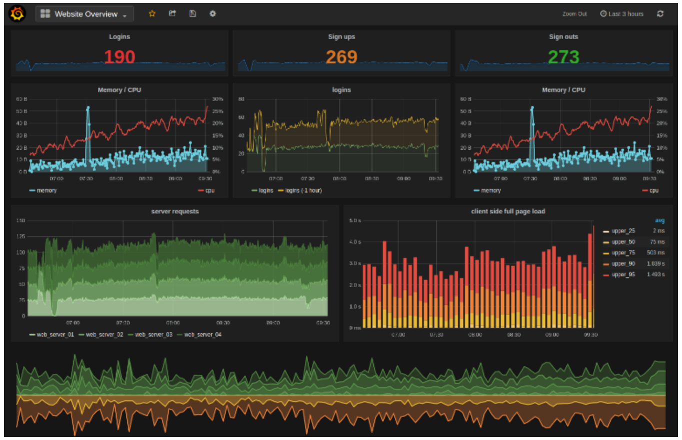
Select a measure
● Nb Accomodations
○ Price
○ Nb Beds

Select a level of detail
● By Neighbourhood
○ By Hexbin

22,101 hosts
27,318 accomodations
207,120 beds

Sources: http://insideairbnb.com/
Provided by @i_trajkovic and @Synaltic





Source: Jonathan Trajkovic https://public.tableau.com/en-us/s/blog/2015/07/analyzing-airbnb-data

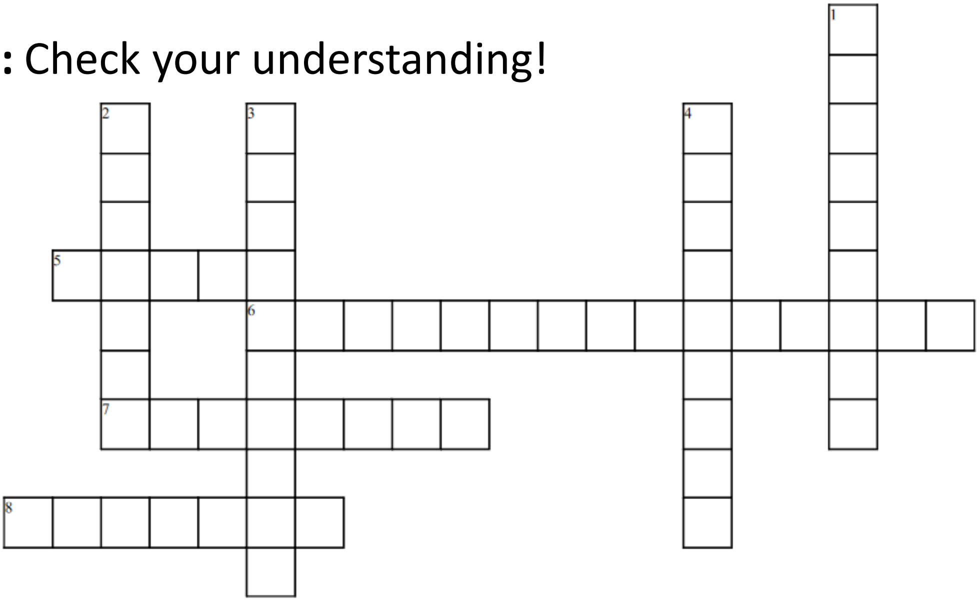
Outlook

Next lecture:

- Introduction into exploratory data analysis (EDA) and visualization

For those interested in the analysis of the Airbnb data:
https://github.com/ruchigupta19/Boston-Airbnb-data-analysis

Homework: Check your understanding!



Across

5. Central to the iterative approach

6. Why you need to document your code

7. Split your sample into a test and ... set

8. The first step is to specify a...

Down

1. A project lifecycle is...

2. Always start your data modelling with a Proof of ...

3. ...and unsupervised learning

4. Common tool for communicating results