DRESDEN concept

Datenanalyse und Machine Learning: R versus Python

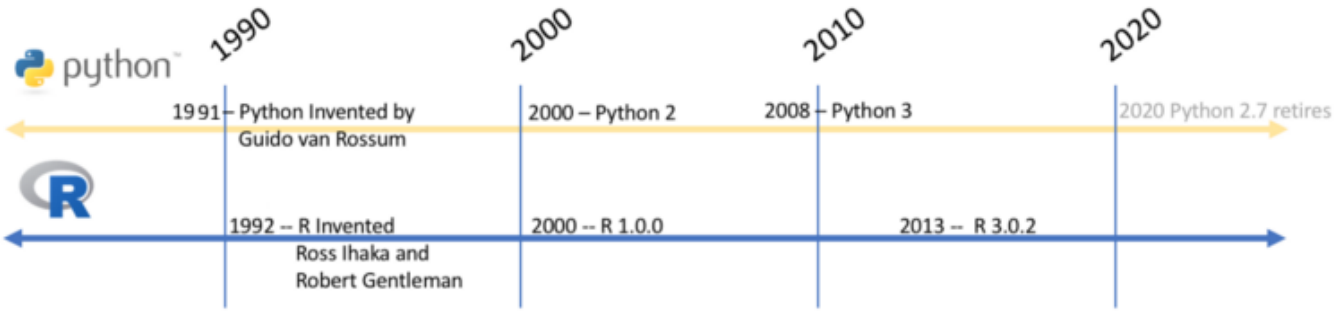
Tim Schmittmann

10 Dezember 2018

Gliederung

- Grundlagen
- Packages
- Trends
- Beispielproblem
- Fazit

Grundlagen



- Universelle Skriptsprachen
- Interaktive Kommandozeileninterpreter

Grundlagen

- Multiparadigmatisch
- Fokus auf Einfachheit und Produktivität
- There should be one obvious way to do

- Eher Funktional
- Fokus auf Datenanalyse, Statistik und Grafiken
- Many ways to do it

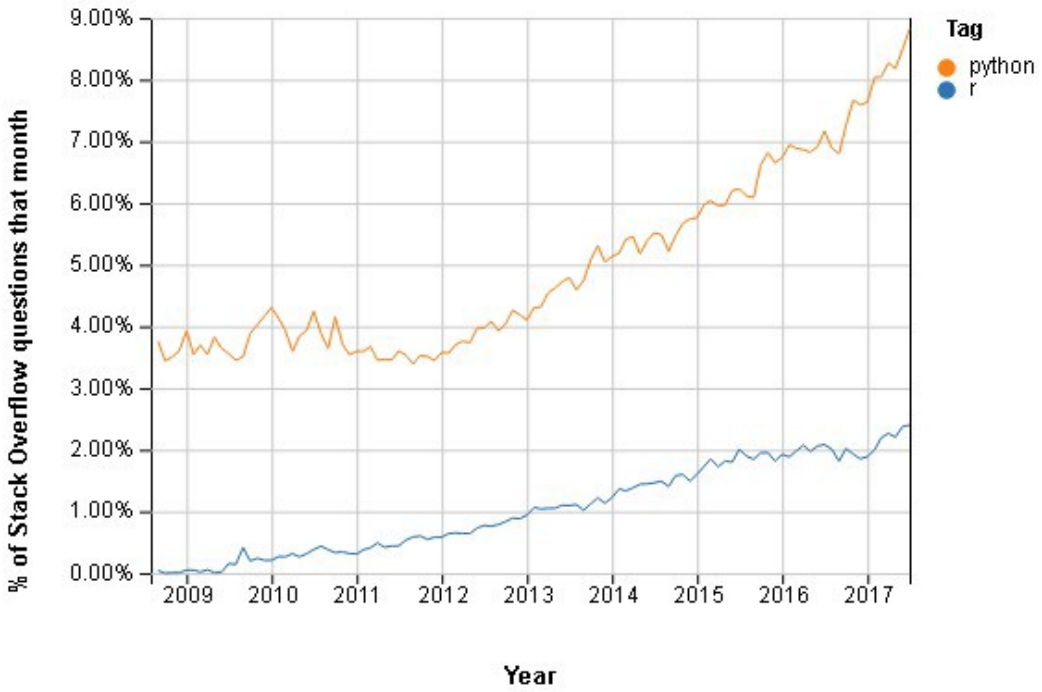
Packages

- PyPi
- 161k packages
  7,8k scientific/engineering
- Groß, "Standardpackages"

- CRAN
- 13,5k packages
- Klein, funktional

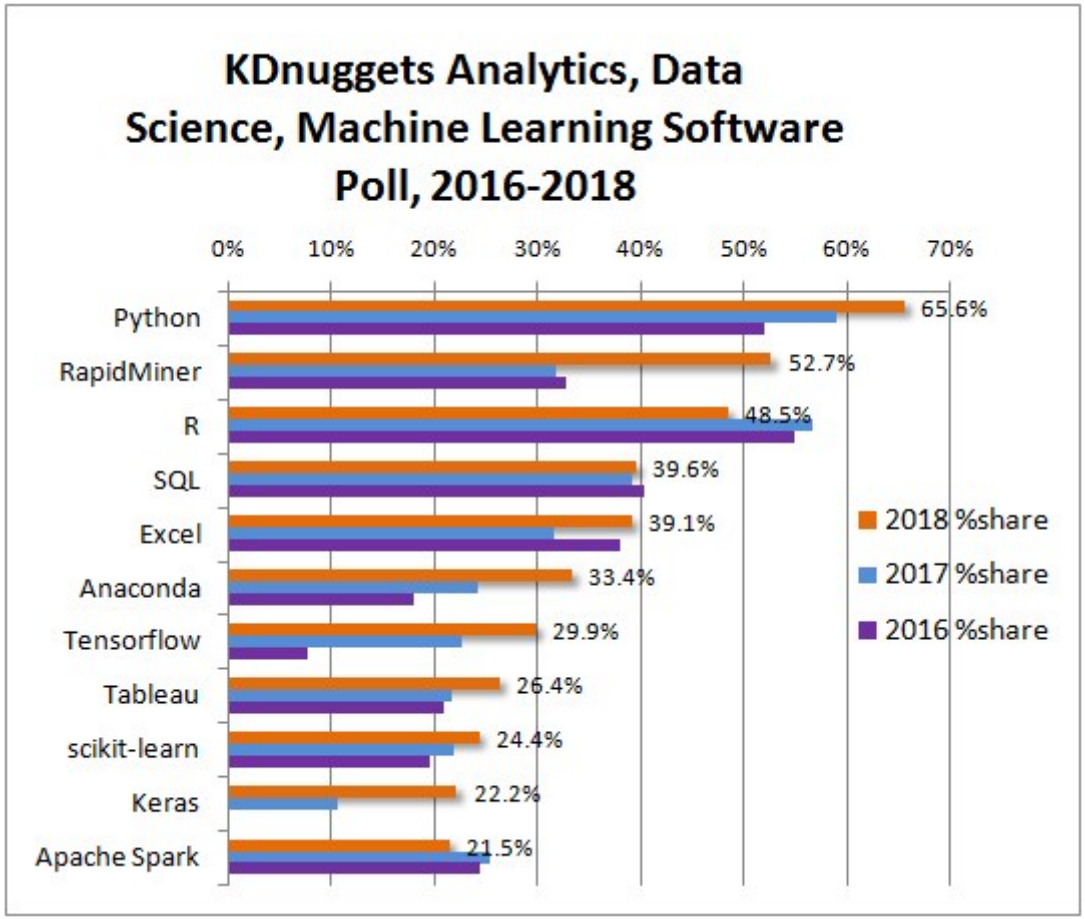
5/16

Trends


9.00%
8.00%
7.00%
6.00%
5.00%
4.00%
3.00%
2.00%
1.00%
0.00%
% of Stack Overflow questions that month
2009
2010
2011
2012
2013
2014
2015
2016
2017
Year
Tag
python
r

6/16

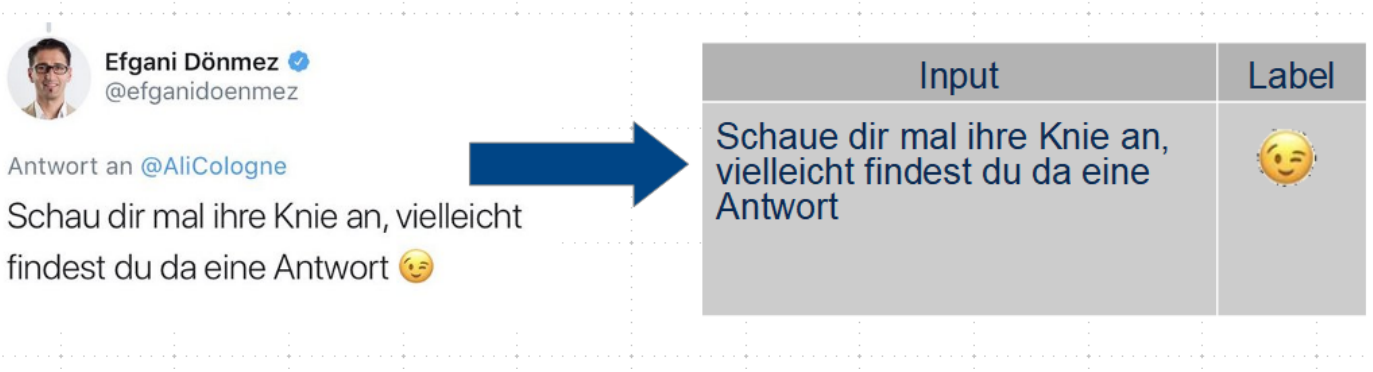
Trends


KDnuggets Analytics, Data Science, Machine Learning Software Poll, 2016-2018
0%
10%
20%
30%
40%
50%
60%
70%
Python
RapidMiner
R
SQL
Excel
Anaconda
Tensorflow
Tableau
scikit-learn
Keras
Apache Spark
65.6%
52.7%
48.5%
39.6%
39.1%
33.4%
29.9%
26.4%
24.4%
22.2%
21.5%
2018 %share
2017 %share
2016 %share

7/16

Beispielproblem

- Sentiment Analysis auf Tweets mit Emojis
- Varianten Binary Class und Multi-Class



Daten sammeln

```
import twitter
import pandas as pd
api = twitter.Api(consumer_key, consumer_secret, access_token, access_secret)
result = api.GetSearch(raw_query="q=♥&count=2")
df = pd.DataFrame([[x.AsDict()["id"],x.AsDict()["text"]] for x in result])
```

```
library(twitteR)
setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)
search = twitteR::searchTwitter("♥", n = 2)
df = twitteR::twListToDF(search)[,c("id", "text")]
```

```
##   id           text
## 1 1.057767e+18 Wenn das Herz liebt <U+2665><U+270C>
## 2 1.057766e+18 <U+263A> Gute Nacht Köln<U+2665><U+2665>
```

Daten speichern

```
df.to_csv("data.csv", sep=";", encoding="UTF-8",header=["id","text"], index=False)
df = pd.read_csv("data.csv", sep=";", encoding="UTF-8")
```

```
write.csv2(df, "data.csv", row.names = FALSE)

df = read.csv("data.csv", sep=';', skip=1, header=FALSE,
              colClasses = c("integer64", "character"), encoding="UTF-8")
names(df) = c("id","text")
df = df[,c("id","text")]
```

Daten aufbereiten

```
df = df.replace({'\n': ' ', '\r': ' '}, regex=True)
df = df.drop_duplicates(['text'])
df = df.sort_index(ascending=False)
```

```
df$text = gsub(pattern = "\n", replacement = " ", df$text)
df = df[!duplicated(df[,"text"]),]
df = df[order(df$id, decreasing = TRUE),]
```

Daten aufbereiten

```
import emoji
emoji_regex = "|".join(emoji.UNICODE_EMOJI).replace("*","\*")
df.loc[:,'target'] = df.loc[:,'text'].apply(extract_emojis)
df.loc[:,'text'] = df.loc[:,'text'].apply(lambda text: regex.sub(emoji_regex,"",text)
```

```
df[, "text"] = gsub("<U+FE0F>", "", df[, "text"], fixed=TRUE)
df[, "target"] = str_extract_all(df[,"text"], '<U\\+[0-9A-F]+>') %>%
                      lapply(paste, collapse = ",") %>%
                      unlist(use.names=FALSE)
df[,"text"] = gsub('<U\\+[0-9A-F]+>', "", df[,"text"])
```

```
##             id                 text                      target
## 1 1.057767e+18 Wenn das Herz liebt          <U+2665>,<U+270C>
## 2 1.057766e+18      Gute Nacht Köln <U+263A>,<U+2665>,<U+2665>
```

Emoji 1: red heart

Emoji 2: flexed biceps

Emoji 3: face with tears of joy

Error: kann png()-Gerät nicht starten

Fazit

- Python für Anfänger
- Wenn R, dann richtig
- Auf die Packages achten

Fragen?

Literatur

- https://www.dataquest.io/blog/python-vs-r/ (https://www.dataquest.io/blog/python-vs-r/)
- https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis (https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis)
- https://medium.com/@data_driven/python-vs-r-for-data-science-and-the-winner-is-3ebb1a968197 (https://medium.com/@data_driven/python-vs-r-for-data-science-and-the-winner-is-3ebb1a968197)
- https://jobsquery.it/stats/language/group (https://jobsquery.it/stats/language/group)
- https://www.kdnuggets.com/2018/06/ecosystem-data-science-python-victory.html (https://www.kdnuggets.com/2018/06/ecosystem-data-science-python-victory.html)
- https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html (https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html)