Data Science 2

Vorlesung - Bot Detection

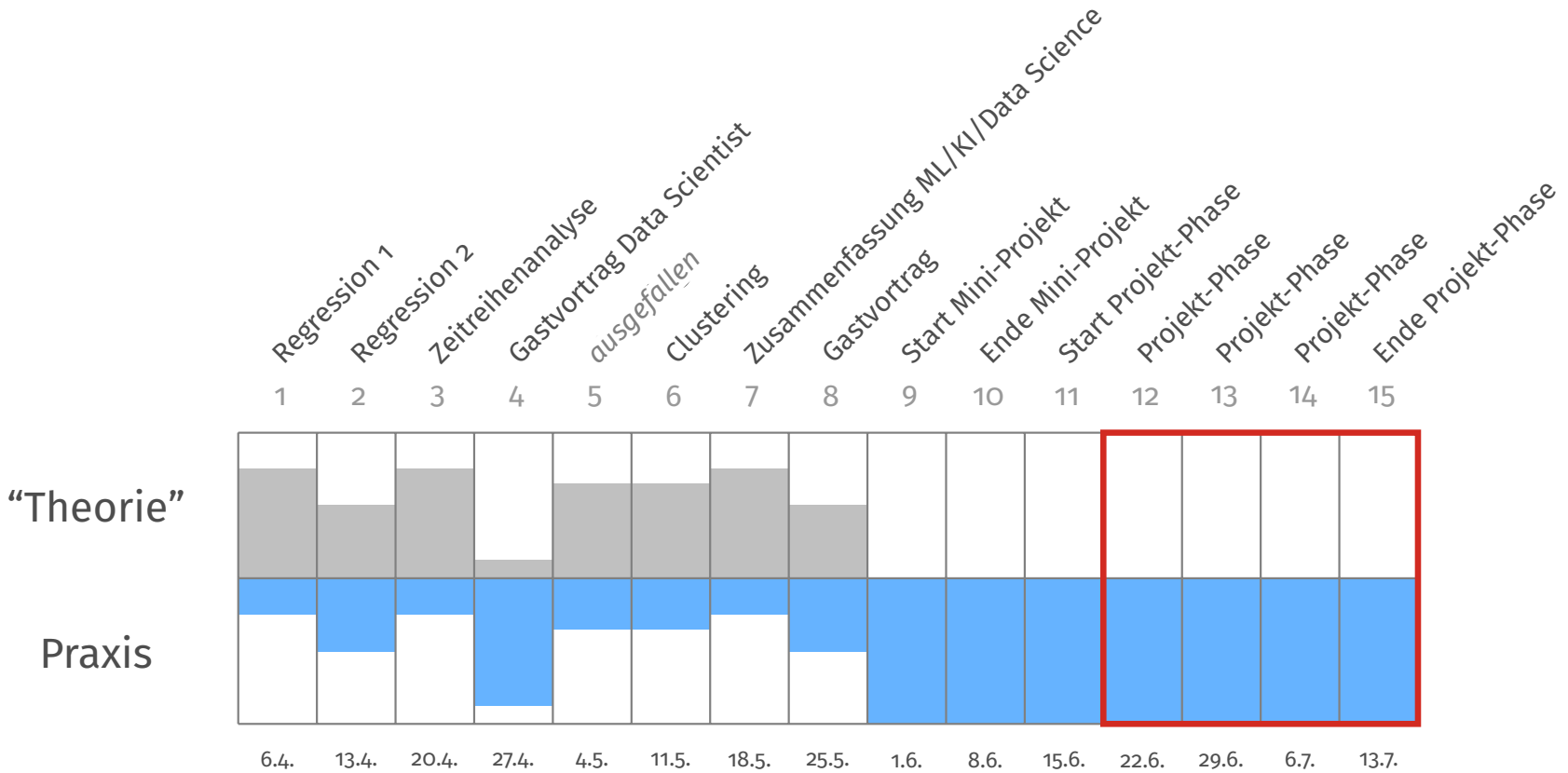
Prof. Dr. Christian Bockermann

Hochschule Bochum

Sommersemester 2021

Inhalt

Projektphase



22.6. um 13 Uhr

- Vorstellung/Besprechung der Gruppen Projekte
- Vortrag: Bot-Erkennung in Web-Daten (Projekt)

29.6. um 13 Uhr

- Projektarbeit, Fragerunde,...
- Vortrag: Big Data?

6.7. um 13 Uhr

- Vortrag/Diskussion: Datenanalyse, Datenschutz und Ethik

13.7. um 13 Uhr

- Gastvortrag der Fa. CuraCon (Wirtschaftsprüfung)
  Das Geheimnis von der Datenanalyse - und warum sie den Wirtschaftsprüfer (noch) nicht ersetzt!

Vorstellung Gruppen-Projekte

Gruppenprojekte

- Welcher Datensatz wird bearbeitet?
- Welche Fragestellungen haben sich ergeben?
- Ungefähre Aufteilung geplant?

Gruppenprojekte

- Welcher Datensatz wird bearbeitet?
- Welche Fragestellungen haben sich ergeben?
- Ungefähre Aufteilung geplant?

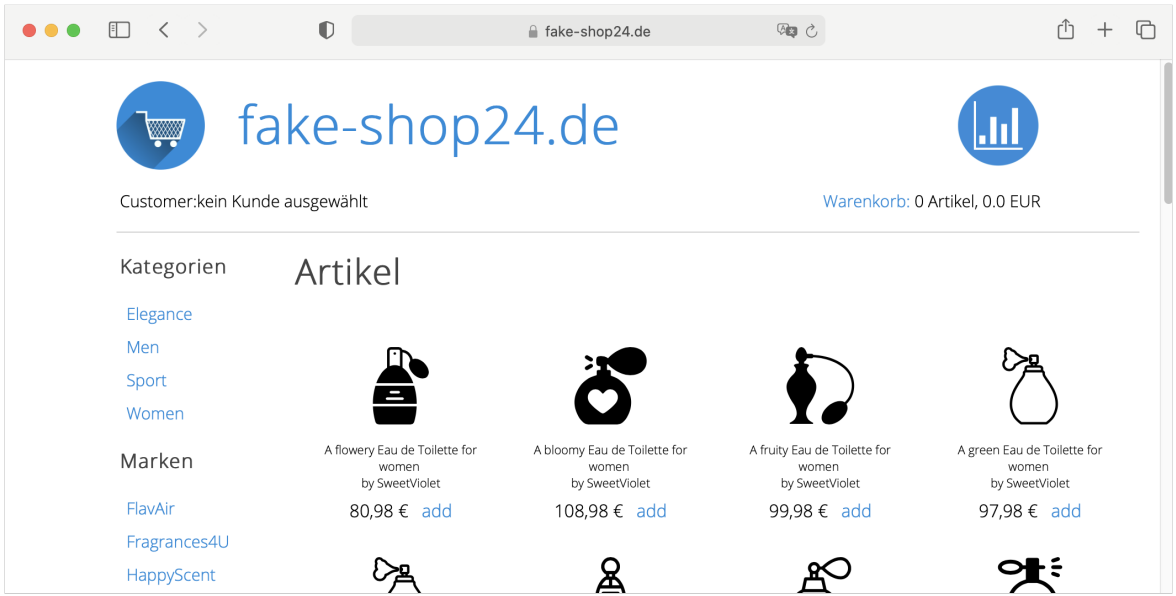
Fragen / Probleme

- Gibt es Fragen/Probleme mit der Aufgabe/den Daten?
- Unterstützung notwendig?

Bot-Erkennung in Web-Daten

e-Commerce Handel

- Primär: Verkauf von Waren
- Marketing-Instrument (Tracking, Kundenprofile,...)
- Kundenbindung (z.B. durch App, Newsletter, Neugier)



Wie messen wir Erfolg?

**Nur was wir messen,
können wir auch verbessern!**

Wie messen wir Erfolg?

**Nur was wir messen,
können wir auch verbessern!**

Mögliche Kennzahlen (KPIs):

- Besucherzahlen (*Visits*)
- *Conversion Rate* – tatsächliche Einkäufe pro Besucher
- Verweildauer im Shop
- Häufigkeit der Wiederkehr von Besuchern

Wie messen wir Erfolg?

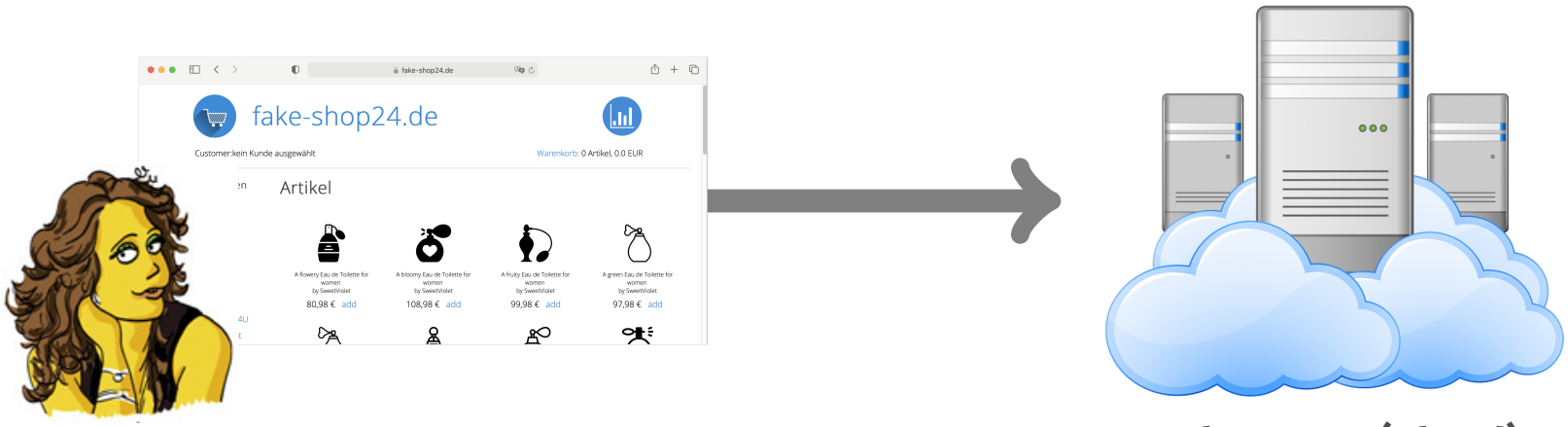
**Nur was wir messen,
können wir auch verbessern!**

Mögliche Kennzahlen (KPIs):

- Besucherzahlen (*Visits*)
- *Conversion Rate* – tatsächliche Einkäufe pro Besucher
- Verweildauer im Shop
- Häufigkeit der Wiederkehr von Besuchern

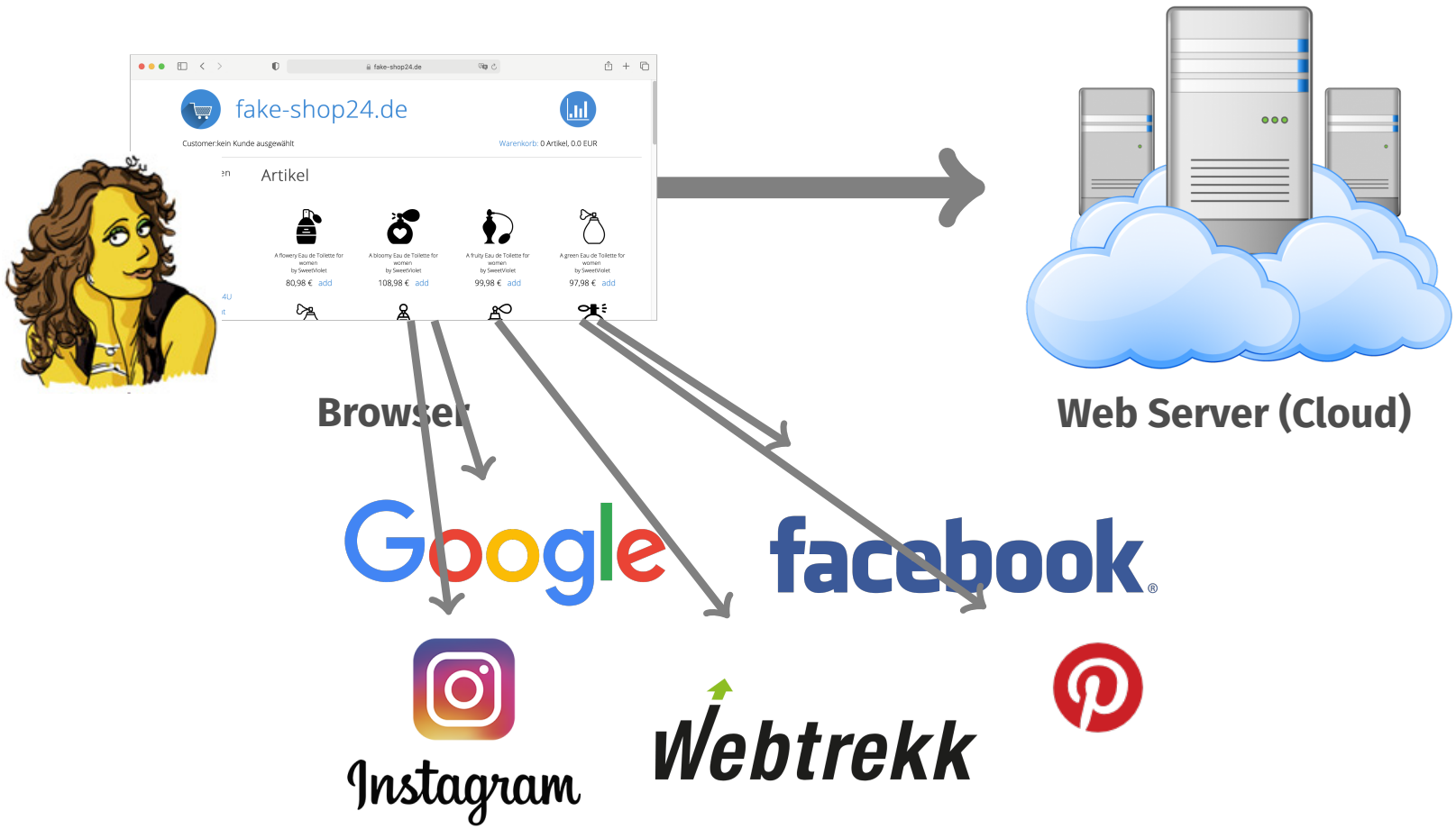
Wie können wir das messen?

Funktionsweise Online-Shop

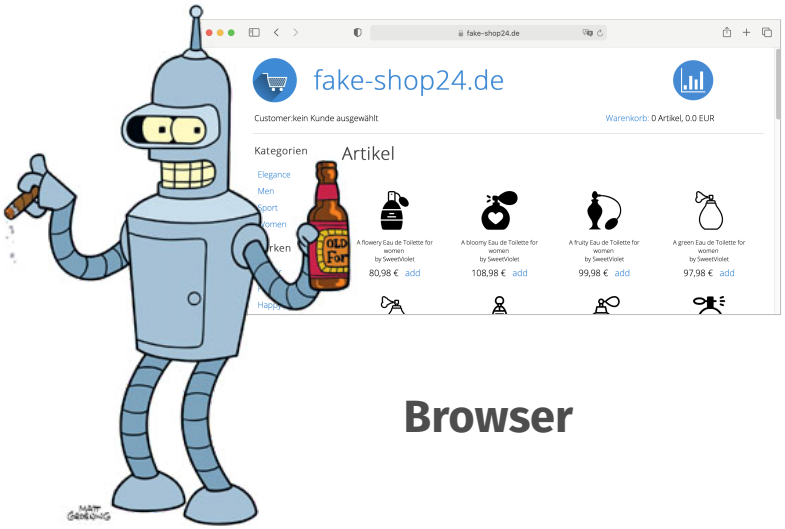


- Browser ruft Web-Seite auf
- Web-Seite besteht aus hunderten Seiten/URLs, Grafiken,...
- Seite enthält zusätzlich URLs zu Google, Adobe,...

Funktionsweise Online-Shop



Funktionsweise Online-Shop



Browser

Web Server (Cloud)

Zentrale Frage:

Sind alle Besucher auch wirklich Menschen = potentielle Käufer?

Bot Erkennung

Bot Zugriffe

Bots sind Programme die auf Webseiten zugreifen, z.B.

- Crawler: Google durchsucht alle Webseiten für die Suche
- Monitoring: Überwachung ob Webseite noch funktioniert

Bot Zugriffe

Bots sind Programme die auf Webseiten zugreifen, z.B.

- Crawler: Google durchsucht alle Webseiten für die Suche
- Monitoring: Überwachung ob Webseite noch funktioniert

Aber auch *Bad Bots*:

- Scraping: Abgreifen von Informationen (Preise, Passwörter)
- Spam-Bots: Spam in Blogs/Foren verbreiten
- Click-Fraud: Werbe-Clicks/Traffic erzeugen

Bot Zugriffe

Bots sind Programme die auf Webseiten zugreifen, z.B.

- Crawler: Google durchsucht alle Webseiten für die Suche
- Monitoring: Überwachung ob Webseite noch funktioniert

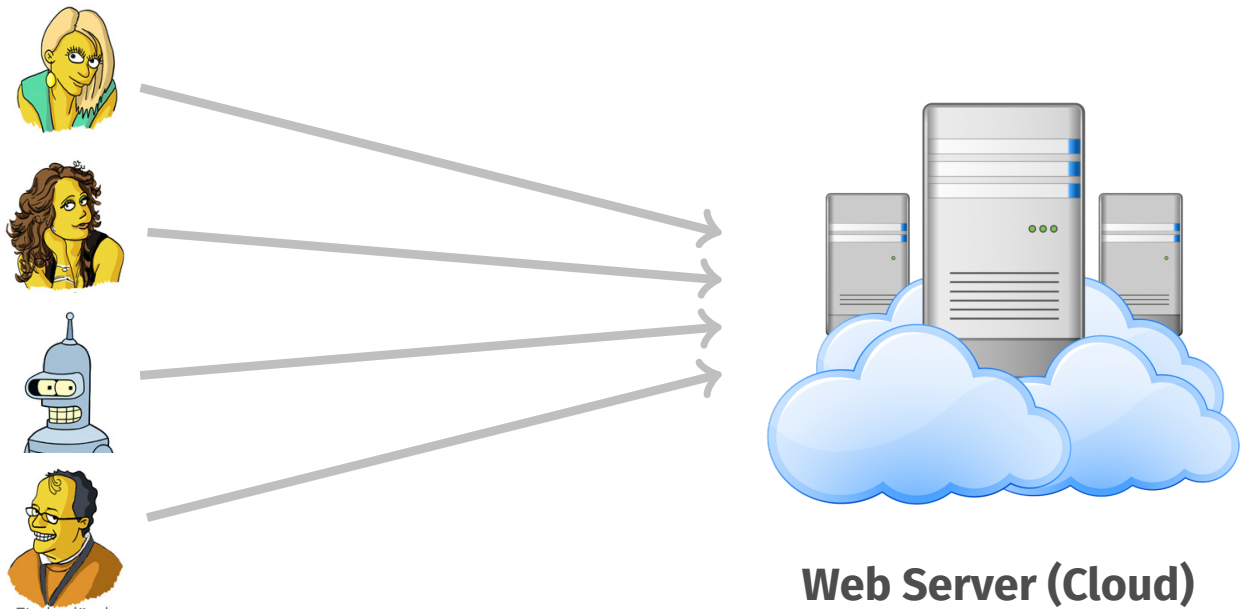
Aber auch *Bad Bots*:

- Scraping: Abgreifen von Informationen (Preise, Passwörter)
- Spam-Bots: Spam in Blogs/Foren verbreiten
- Click-Fraud: Werbe-Clicks/Traffic erzeugen

Bad Bots führen zu verfälschten KPIs!

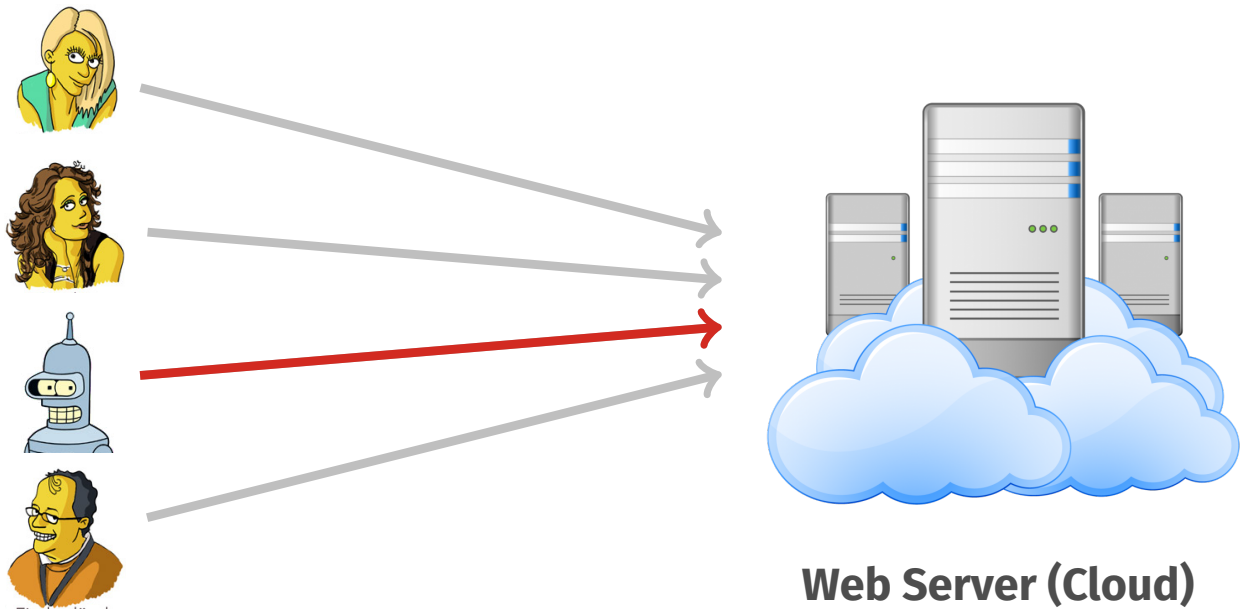
Unternehmerische Fragestellung(en)

- Wieviele meiner Besucher sind überhaupt Bots?
- Wie stark beeinflusst Bot-Traffic meine KPIs/Kampagnien?
- Wie können wir *bad bot traffic* blockieren?



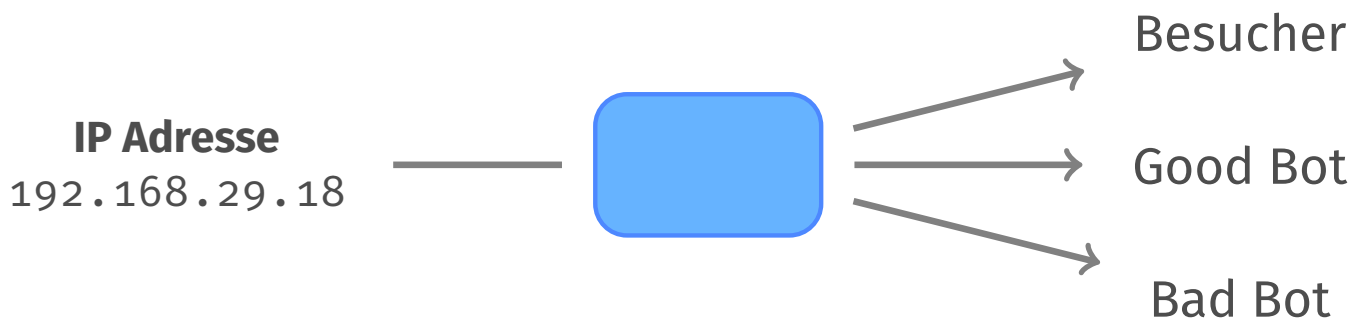
Unternehmerische Fragestellung(en)

- Wieviele meiner Besucher sind überhaupt Bots?
- Wie stark beeinflusst Bot-Traffic meine KPIs/Kampagnien?
- Wie können wir *bad bot traffic* blockieren?



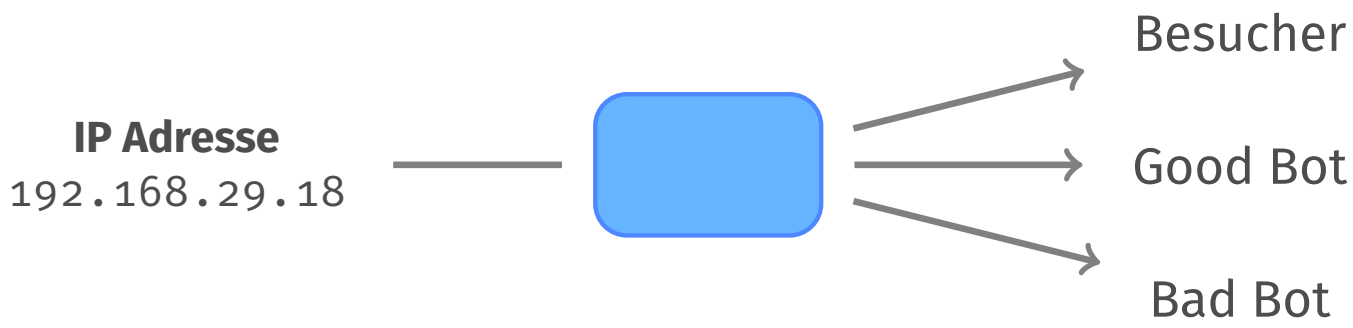
Machine Learning Problem:

- Zugriff auf Shop von IP-Adressen
- Hinter IP-Adresse kann Kunde/Besucher oder Bot stecken



Machine Learning Problem:

- Zugriff auf Shop von IP-Adressen
- Hinter IP-Adresse kann Kunde/Besucher oder Bot stecken

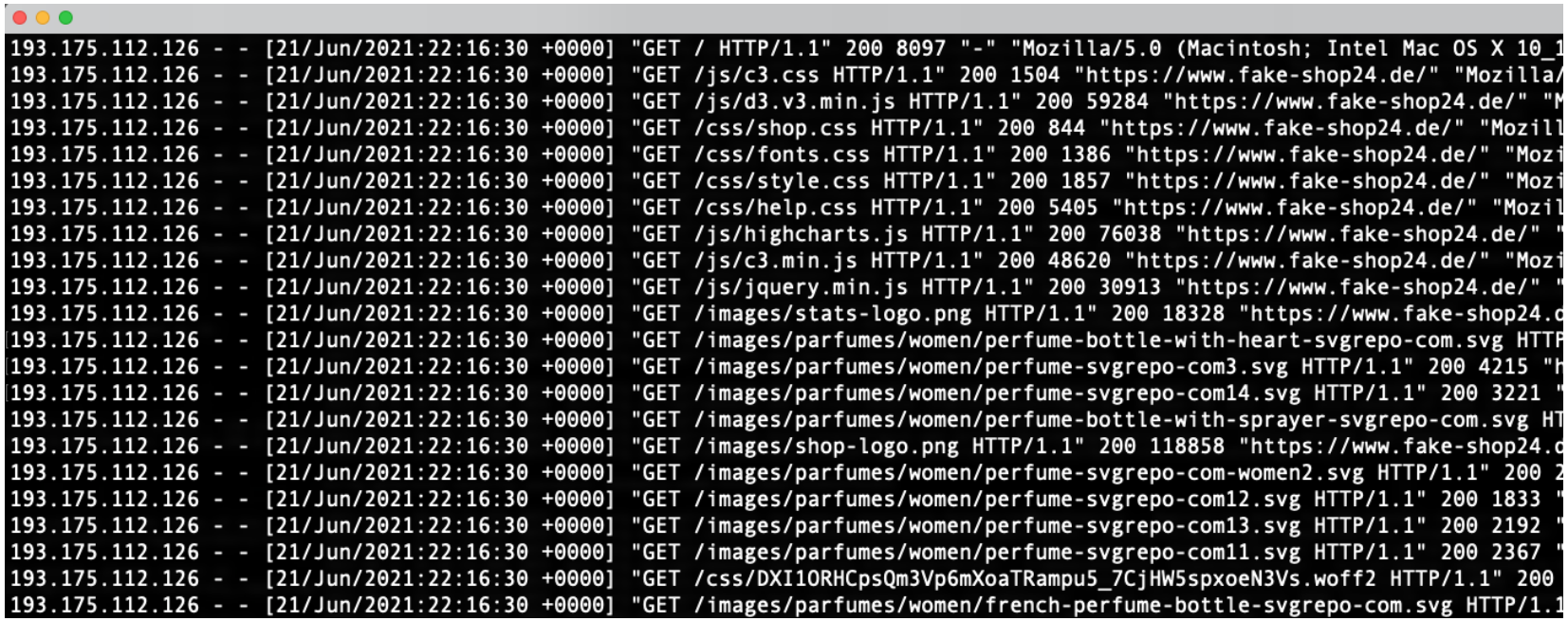


Klassifikationsaufgabe

Web Server Log-daten

Daten: www.fake-shop24.de

- Jeder Web-Server schreibt Log-Daten für Zugriffe
- Jeder einzelne Aufruf (Seite, Bild,...) wird protokolliert
- Log-Daten enthalten IP-Adresse, Browser-Typ, Uhrzeit, uvm.



```
193.175.112.126 - - [21/Jun/2021:22:16:30 +0000] "GET / HTTP/1.1" 200 8097 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_1
193.175.112.126 - - [21/Jun/2021:22:16:30 +0000] "GET /js/c3.css HTTP/1.1" 200 1504 "https://www.fake-shop24.de/" "Mozilla/
193.175.112.126 - - [21/Jun/2021:22:16:30 +0000] "GET /js/d3.v3.min.js HTTP/1.1" 200 59284 "https://www.fake-shop24.de/" "M
193.175.112.126 - - [21/Jun/2021:22:16:30 +0000] "GET /css/shop.css HTTP/1.1" 200 844 "https://www.fake-shop24.de/" "Mozill
193.175.112.126 - - [21/Jun/2021:22:16:30 +0000] "GET /css/fonts.css HTTP/1.1" 200 1386 "https://www.fake-shop24.de/" "Mozi
193.175.112.126 - - [21/Jun/2021:22:16:30 +0000] "GET /css/style.css HTTP/1.1" 200 1857 "https://www.fake-shop24.de/" "Mozi
193.175.112.126 - - [21/Jun/2021:22:16:30 +0000] "GET /css/help.css HTTP/1.1" 200 5405 "https://www.fake-shop24.de/" "Mozil
193.175.112.126 - - [21/Jun/2021:22:16:30 +0000] "GET /js/highcharts.js HTTP/1.1" 200 76038 "https://www.fake-shop24.de/" "
193.175.112.126 - - [21/Jun/2021:22:16:30 +0000] "GET /js/c3.min.js HTTP/1.1" 200 48620 "https://www.fake-shop24.de/" "Mozi
193.175.112.126 - - [21/Jun/2021:22:16:30 +0000] "GET /js/jquery.min.js HTTP/1.1" 200 30913 "https://www.fake-shop24.de/" "
193.175.112.126 - - [21/Jun/2021:22:16:30 +0000] "GET /images/stats-logo.png HTTP/1.1" 200 18328 "https://www.fake-shop24.c
193.175.112.126 - - [21/Jun/2021:22:16:30 +0000] "GET /images/parfumes/women/perfume-bottle-with-heart-svgrepo-com.svg HTTP
193.175.112.126 - - [21/Jun/2021:22:16:30 +0000] "GET /images/parfumes/women/perfume-svgrepo-com3.svg HTTP/1.1" 200 4215 "h
193.175.112.126 - - [21/Jun/2021:22:16:30 +0000] "GET /images/parfumes/women/perfume-svgrepo-com14.svg HTTP/1.1" 200 3221 "
193.175.112.126 - - [21/Jun/2021:22:16:30 +0000] "GET /images/parfumes/women/perfume-bottle-with-sprayer-svgrepo-com.svg HT
193.175.112.126 - - [21/Jun/2021:22:16:30 +0000] "GET /images/shop-logo.png HTTP/1.1" 200 118858 "https://www.fake-shop24.c
193.175.112.126 - - [21/Jun/2021:22:16:30 +0000] "GET /images/parfumes/women/perfume-svgrepo-com-women2.svg HTTP/1.1" 200 2
193.175.112.126 - - [21/Jun/2021:22:16:30 +0000] "GET /images/parfumes/women/perfume-svgrepo-com12.svg HTTP/1.1" 200 1833 "
193.175.112.126 - - [21/Jun/2021:22:16:30 +0000] "GET /images/parfumes/women/perfume-svgrepo-com13.svg HTTP/1.1" 200 2192 "
193.175.112.126 - - [21/Jun/2021:22:16:30 +0000] "GET /images/parfumes/women/perfume-svgrepo-com11.svg HTTP/1.1" 200 2367 "
193.175.112.126 - - [21/Jun/2021:22:16:30 +0000] "GET /css/DXI1ORHCpsQm3Vp6mXoaTRampu5_7CjHW5spxoeN3Vs.woff2 HTTP/1.1" 200
193.175.112.126 - - [21/Jun/2021:22:16:30 +0000] "GET /images/parfumes/women/french-perfume-bottle-svgrepo-com.svg HTTP/1.1
```

Web-Server Access Logs

Typische Felder sind:

IP Adresse	Quelle IP Adresse der Verbindung
Datum	Datum+Uhrzeit des Zugriffs
Method+URI	Zugriffsmethode und URL
Status-Code	Status der Antwort (Ok, Fehler,..)
Referer	Vorangegangene URL
User-Agent	Bezeichnung des Browsers

Access Logs

Beispiel: `datascience.hs-bochum.de`

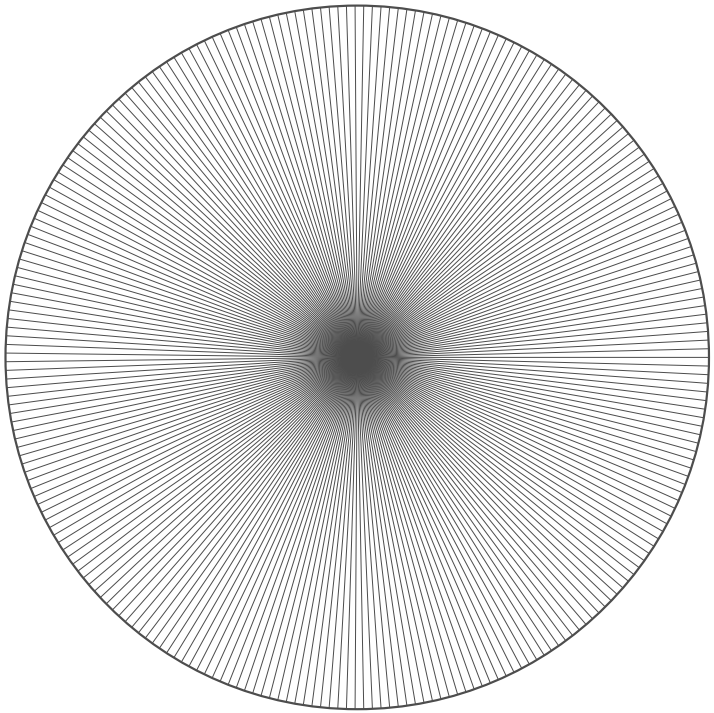
- Log Daten von 5.10.2020 bis 21.6.2021
- 1.405.942 Anfragen in ca. 260 Tagen
- von 1200 verschiedene IP-Addressen
- 24.419 verschiedene URLs
- 326 verschiedene Browser-Typen

IP Blöcke

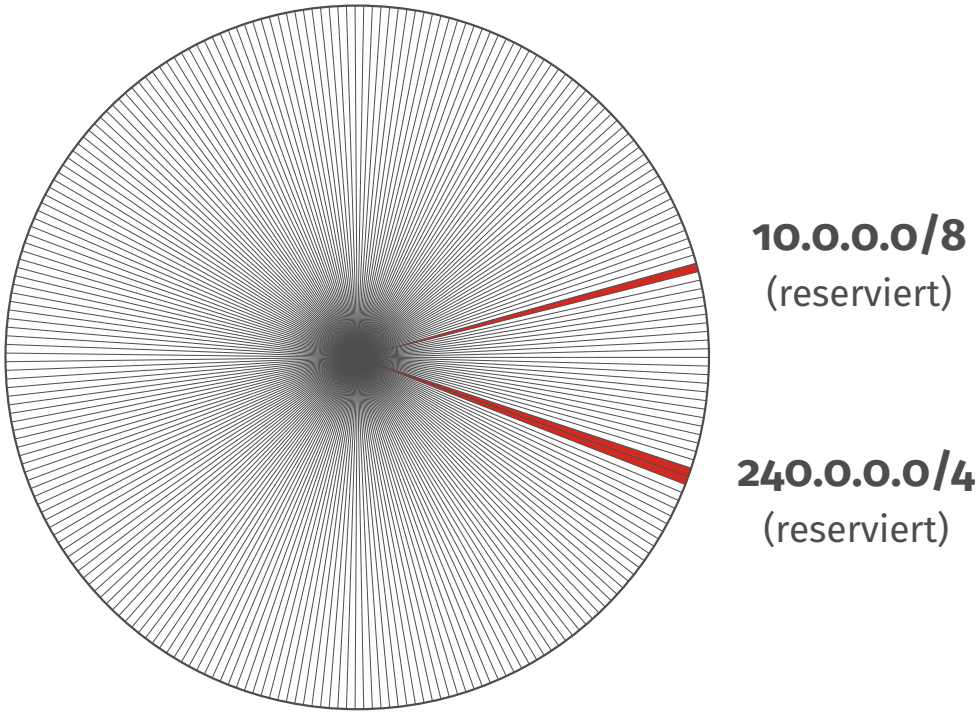
Aufteilung von IP-Adressen: IANA, RIPE, APNIC,...

- IPv4 hat 4.294.967.296 mögliche Adressen (alle vergeben)
- IPv6 hat $3.4 \cdot 10^{38}$ = 340 Sextillionen Adressen
- Verteilung von Adressblöcken durch Organisationen IANA, RIPE (EU),...
- Organisationen verteilen weiter an Internet-Provider, Unternehmen, Hochschulen,...

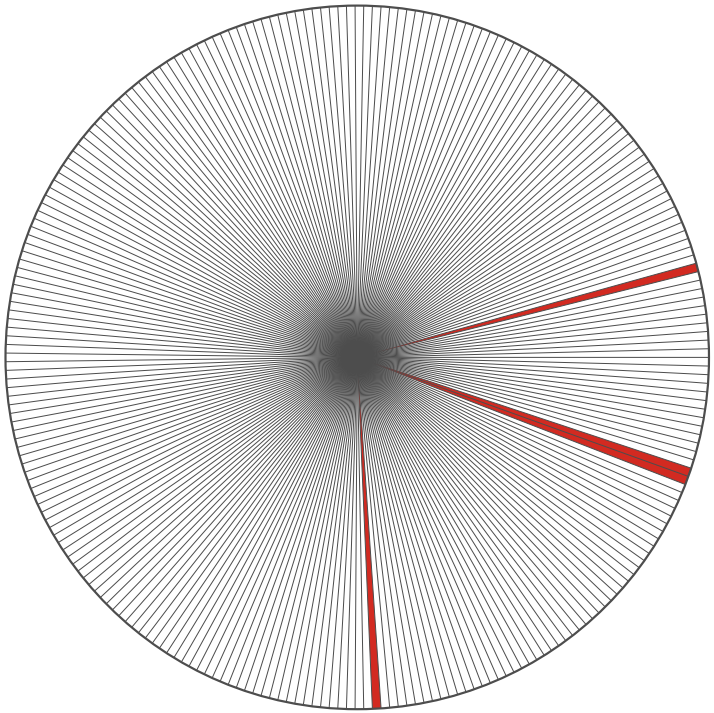
Aufteilung von IP-Adressen: IANA, RIPE, APNIC,...



Aufteilung von IP-Adressen: IANA, RIPE, APNIC,...



Aufteilung von IP-Adressen: IANA, RIPE, APNIC,...



10.0.0.0/8
(reserviert)

240.0.0.0/4
(reserviert)

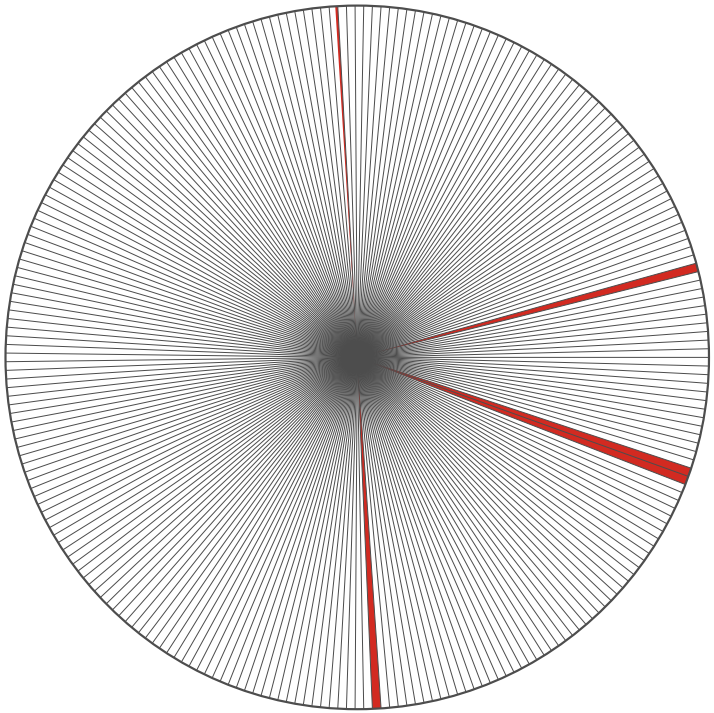
193.0.0.0/8
(reserviert durch RIPE)

→

193.175.112.0 - 193.175.115.0
(FH Bochum)

Aufteilung von IP-Adressen: IANA, RIPE, APNIC,...

66.249.64.0/19
(Google)



10.0.0.0/8
(reserviert)

240.0.0.0/4
(reserviert)

193.0.0.0/8
(reserviert durch RIPE)

→

193.175.112.0 - 193.175.115.0
(FH Bochum)

Informationen über IP-Blöcke in WHOIS Datenbank

- Organisation, der ein IP-Block zugewiesen wurde
- Land/Adress-Informationen

`datascience.hs-bochum.de` ⟶ `193.175.84.83`

```
# whois.ripe.net

inetnum:        193.175.84.0 - 193.175.85.255
netname:        FH-BOCHUM
descr:          Hochschule Bochum
admin-c:        DP12937-RIPE
tech-c:         HR668-RIPE
country:        DE
status:         ASSIGNED PA
mnt-by:         DFN-LIR-MNT
mnt-irt:        IRT-DFN-CERT
created:        1970-01-01T00:00:00Z
last-modified:  2015-12-11T10:05:28Z
source:         RIPE
```

Beispiel: Google Bot

```
66.249.64.63  - -
    [21/Jun/2021:20:12:11 +0000]
    GET / HTTP/1.1
    200
    7318
    -
    Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google...)
```

Beispiel: Google Bot

IP Adresse

```
66.249.64.63 - -
    [21/Jun/2021:20:12:11 +0000]
    GET / HTTP/1.1
    200
    7318
    -
    Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google...)
```

Beispiel: Google Bot

IP Adresse

```
66.249.64.63 - -
    [21/Jun/2021:20:12:11 +0000]
    GET / HTTP/1.1
    200
    7318
    -
    Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google...)
```

Browser Type (*User-Agent*)

Whois Eintrag

```
# whois.arin.net

NetRange:       66.249.64.0 - 66.249.95.255
CIDR:           66.249.64.0/19
NetName:        GOOGLE
NetHandle:      NET-66-249-64-0-1
Parent:         NET66 (NET-66-0-0-0-0)
NetType:        Direct Allocation
OriginAS:
Organization:   Google LLC (GOGL)
RegDate:        2004-03-05
Updated:        2012-02-24
Ref:            https://rdap.arin.net/registry/ip/66.249.64.0

OrgName:        Google LLC
OrgId:          GOGL
Address:        1600 Amphitheatre Parkway
City:           Mountain View
StateProv:      CA
PostalCode:     94043
Country:        US
RegDate:        2000-03-30
Updated:        2019-10-31
```

Beispiel: Google Bot ??

IP Adresse

```
172.86.75.115 - -
    [13/Mar/2021:13:57:34 +0000]
    GET /n/ HTTP/1.1
    302
    2892
    http://193.175.84.83
    Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google...)
```

Browser Type (*User-Agent*)

Whois Eintrag

```
# whois.arin.net

NetRange:       172.86.64.0 - 172.86.127.255
CIDR:           172.86.64.0/18
NetName:        PONYNET-16
NetHandle:      NET-172-86-64-0-1
Parent:         NET172 (NET-172-0-0-0-0)
NetType:        Direct Allocation
OriginAS:       AS53667
Organization:   FranTech Solutions (SYNDI-5)
RegDate:        2015-05-26
Updated:        2015-05-26
Ref:            https://rdap.arin.net/registry/ip/172.86.64.0

OrgName:        FranTech Solutions
OrgId:          SYNDI-5
Address:        1621 Central Ave
City:           Cheyenne
StateProv:      WY
PostalCode:     82001
Country:        US
RegDate:        2010-07-21
Updated:        2017-01-28
Ref:            https://rdap.arin.net/registry/entity/SYNDI-5
```

Feature Engineering

Klassifikation von IP-Adressen

Beispiel:

- `fake-shop24.de` ist Shop in Deutschland
- IP-Adresse aus Block, der zu BR / VN / CN gehört?
- Zugriff von Google-Bot aus nicht-Google Netzblock?

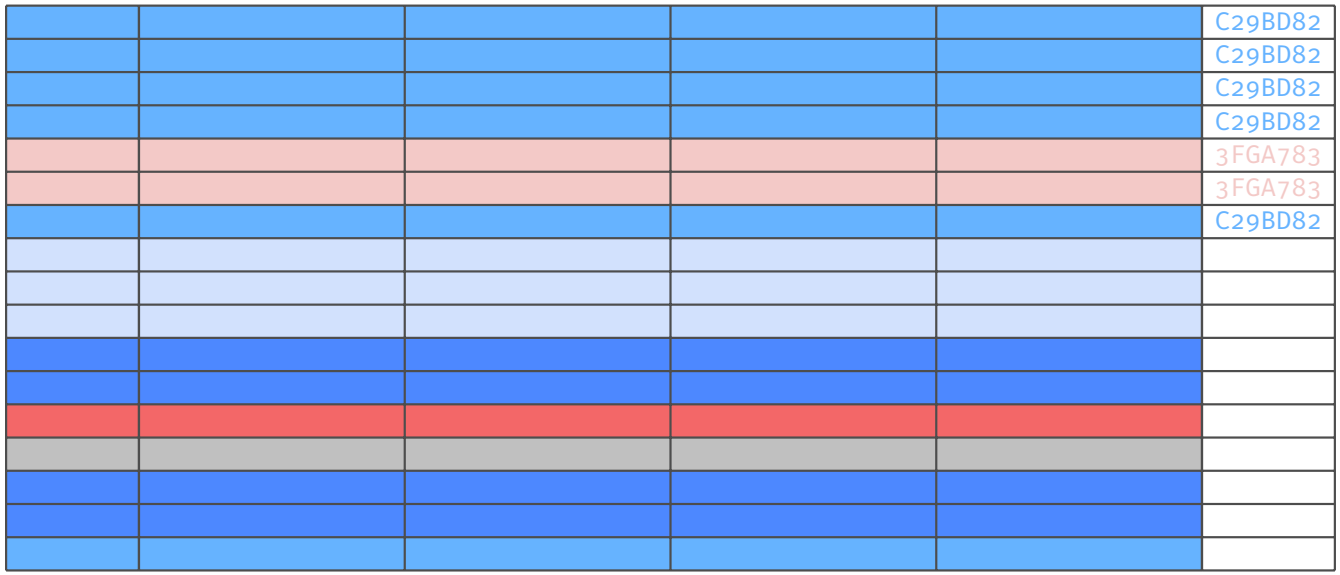
Klassifikation von IP-Adressen

Beispiel:

- `fake-shop24.de` ist Shop in Deutschland
- IP-Adresse aus Block, der zu BR / VN / CN gehört?
- Zugriff von Google-Bot aus nicht-Google Netzblock?
- IP-Addresse die in 1 Stunde 47 verschiedene Browser benutzt?
- Sehr viele Zugriffe in sehr kurzer Zeit?

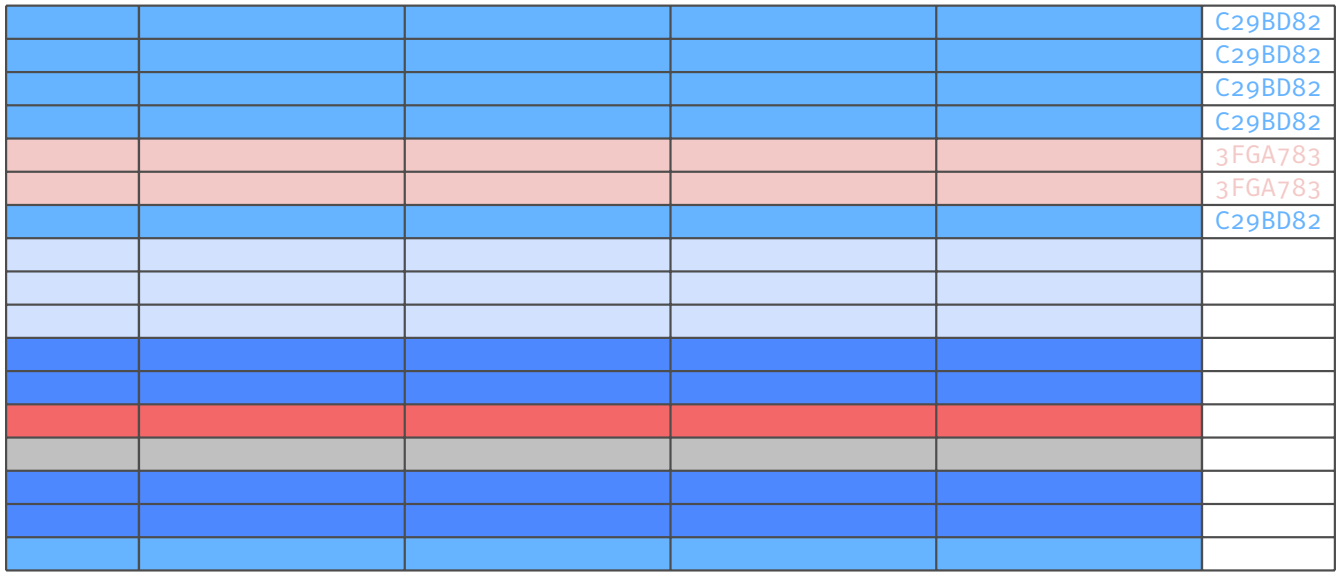
Web Sessions statt Einzel-Anfragen

- Log-Daten: pro Zugriff eine Zeile
- Web-Nutzung in Sessions über Session ID (Cookies)



Web Sessions statt Einzel-Anfragen

- Log-Daten: pro Zugriff eine Zeile
- Web-Nutzung in Sessions über Session ID (Cookies)



Cookies in der Regel nicht in Web-Log Daten enthalten.

Sessions über IP-Adresse

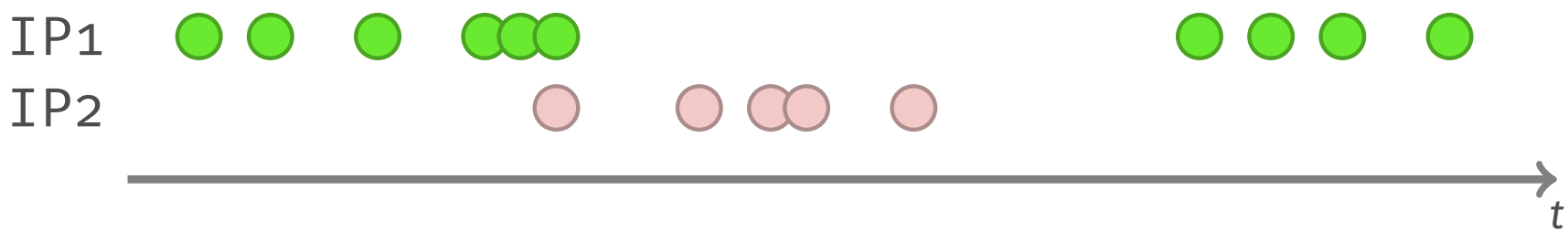
- Wie definieren wir einen Besuch/Session?
- Wann ist eine Session zu Ende?
- Was machen wir, wenn keine Cookies vorhanden?

Sessions über IP-Adresse

- Wie definieren wir einen Besuch/Session?
- Wann ist eine Session zu Ende?
- Was machen wir, wenn keine Cookies vorhanden?
- Ablaufzeit (Timeout) definieren!

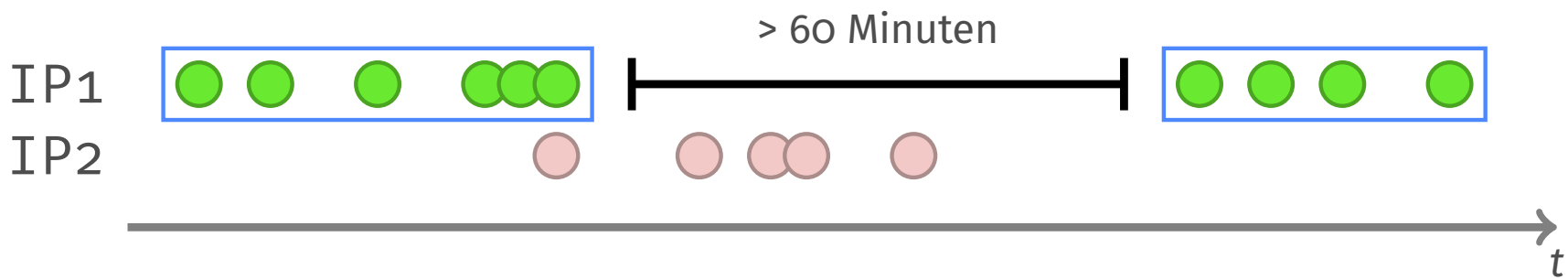
Sessions über IP-Adresse

- Wie definieren wir einen Besuch/Session?
- Wann ist eine Session zu Ende?
- Was machen wir, wenn keine Cookies vorhanden?
- Ablaufzeit (Timeout) definieren!



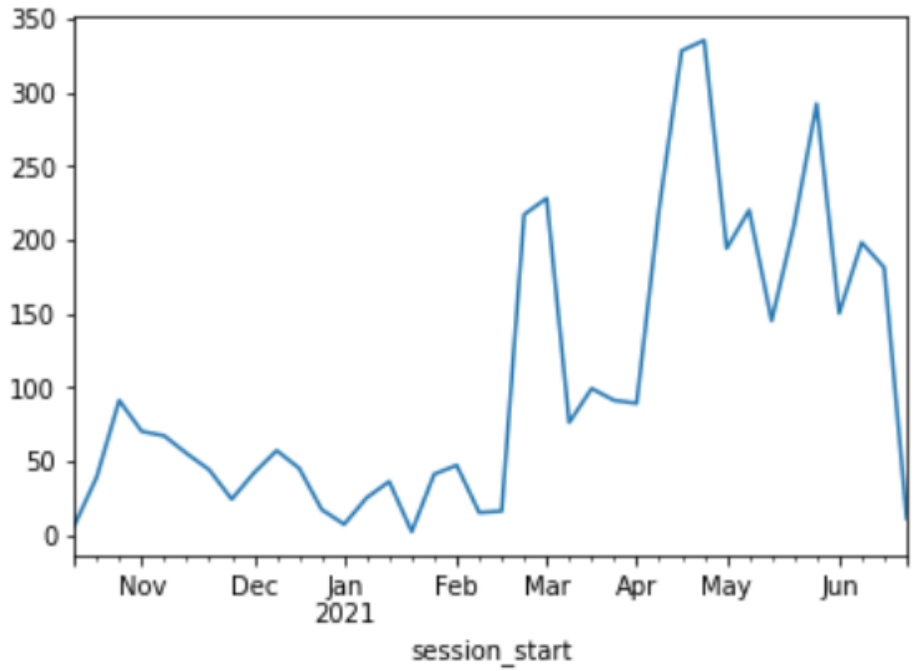
Sessions über IP-Adresse

- Wie definieren wir einen Besuch/Session?
- Wann ist eine Session zu Ende?
- Was machen wir, wenn keine Cookies vorhanden?
- Ablaufzeit (Timeout) definieren!



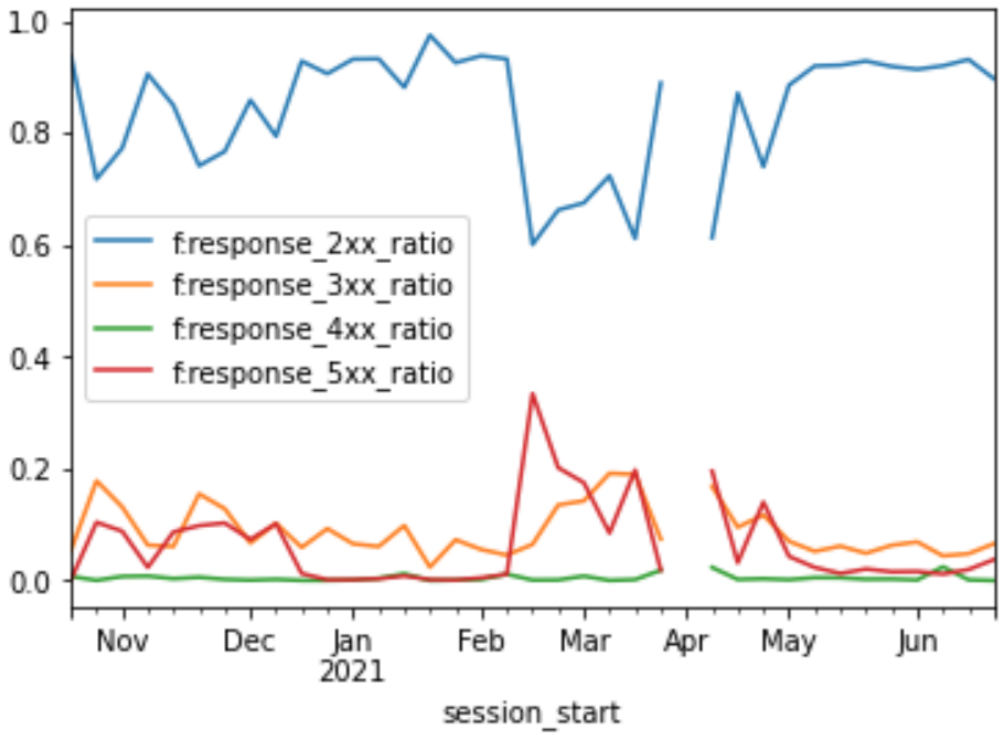
Sessions als Basis für ML

- Wieviel Anfragen pro Session?
- Wann begann die Session? (Tageszeit)
- Mehrere Browser in einer Session?
- Durchschnittliche Zeit zwischen Anfragen der Session?



Sessions als Basis für ML

Status-Meldungen (anteilig je Session, gemittelt)



Literatur Recherche

Google Scholar

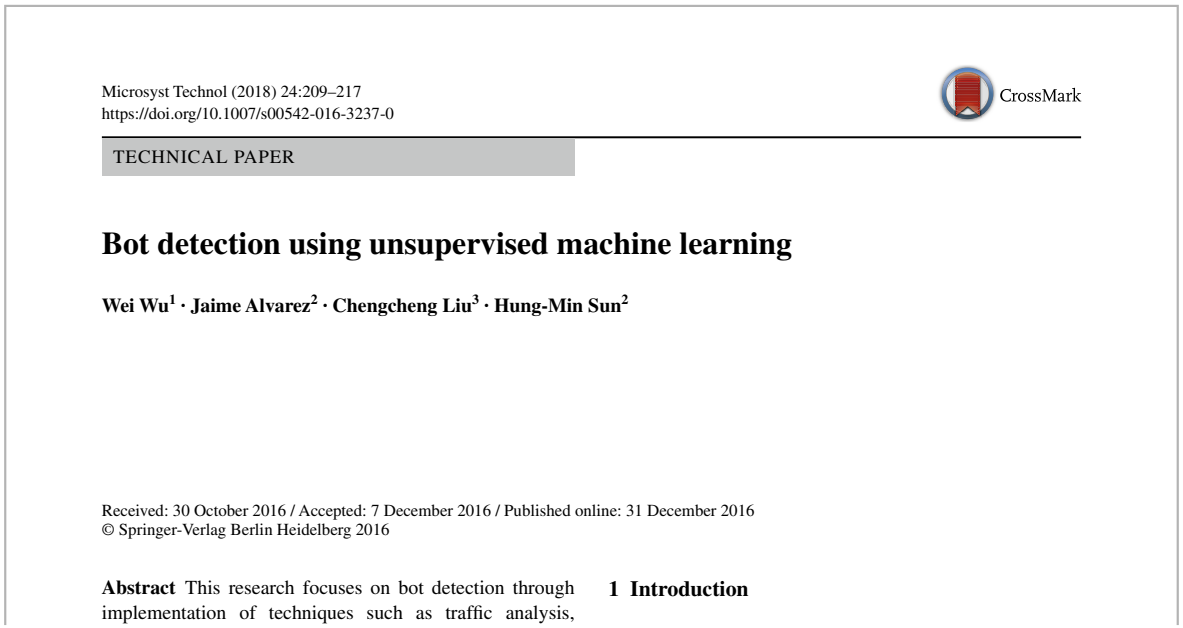
Suche für wissenschaftliche Veröffentlichungen:

`https://scholar.google.com`

Google Scholar

Suche für wissenschaftliche Veröffentlichungen:

https://scholar.google.com



Microsyst Technol (2018) 24:209–217
https://doi.org/10.1007/s00542-016-3237-0

CrossMark

TECHNICAL PAPER

Bot detection using unsupervised machine learning

Wei Wu[1] · Jaime Alvarez[2] · Chengcheng Liu[3] · Hung-Min Sun[2]

Received: 30 October 2016 / Accepted: 7 December 2016 / Published online: 31 December 2016

Abstract This research focuses on bot detection through implementation of techniques such as traffic analysis,

1 Introduction

Zitieren mit BibTeX?

BibTeX Format für Literatur-Angaben:

```
@article{wu2018bot,
  title={Bot detection using unsupervised machine learning},
  author={Wu, Wei and Alvarez, Jaime and Liu,...},
  journal={Microsystem Technologies},
  volume={24},
  number={1},
  pages={209--217},
  year={2018},
  publisher={Springer}
}
```

Kann im Text mit `\cite{wu2018bot}` referenziert werden und führt dann zu [1].

Zitieren mit BibTeX?

BibTeX Format für Literatur-Angaben:

```
@article{wu2018bot,
  title={Bot detection using unsupervised machine learning},
  author={Wu, Wei and Alvarez, Jaime and Liu,...},
  journal={Microsystem Technologies},
  volume={24},
  number={1},
  pages={209--217},
  year={2018},
  publisher={Springer}
}
```

Kann im Text mit `\cite{wu2018bot}` referenziert werden und führt dann zu [1].

BibTex mit Word:

`https://www.microsoft.com/en-us/research/project/academic/articles/new-feature-cite/`

Wei Wu, Jaime Alvarez, Chengcheng Liu, and Hung-Min Sun.
Bot detection using unsupervised machine learning.
Microsystem Technologies, 24(1):209–217, 2018.