Data Science

Data Science ≠ Big Data

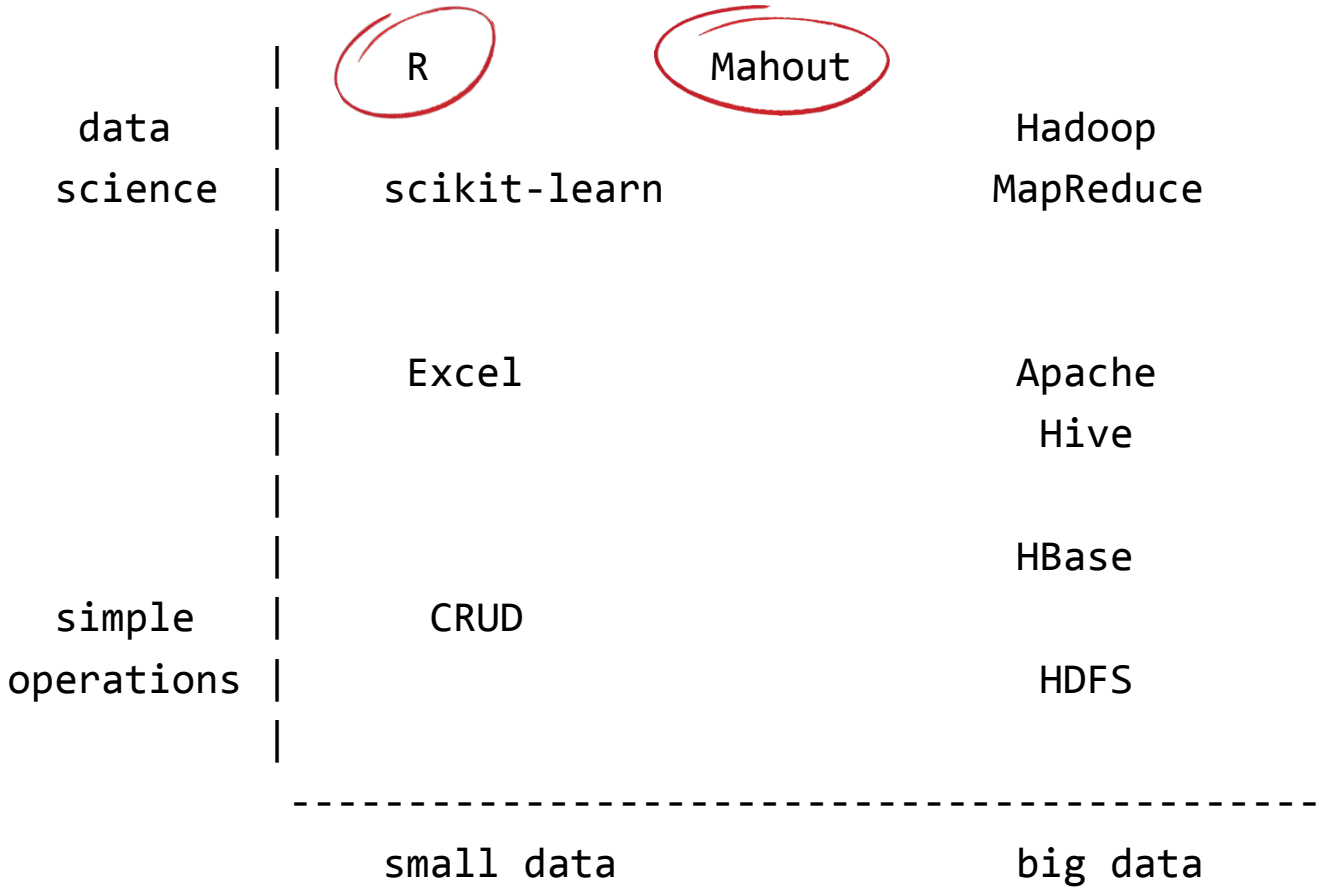
Data Science - het halen van informatie uit data met geavanceerde technieken.

- wiskunde, statistiek en informatica.
- o.a. kansmodellen, machine learning, patroonherkenning, voorspelling, etc.

Big Data - data die te groot is voor de traditionele opslag- en bewerkingsmethodes.

- opslag, analyse, visualisatie, etc.

Data Science ≠ Big Data




mahout

Mahout

Machine Learning library

- 2008
- Java
- Onderdeel van Hadoop
- Clustering, Classificatie, Recommendation
- Schaalt
  - Single-server
  - Hadoop
- Open-source

Mahout – clustering

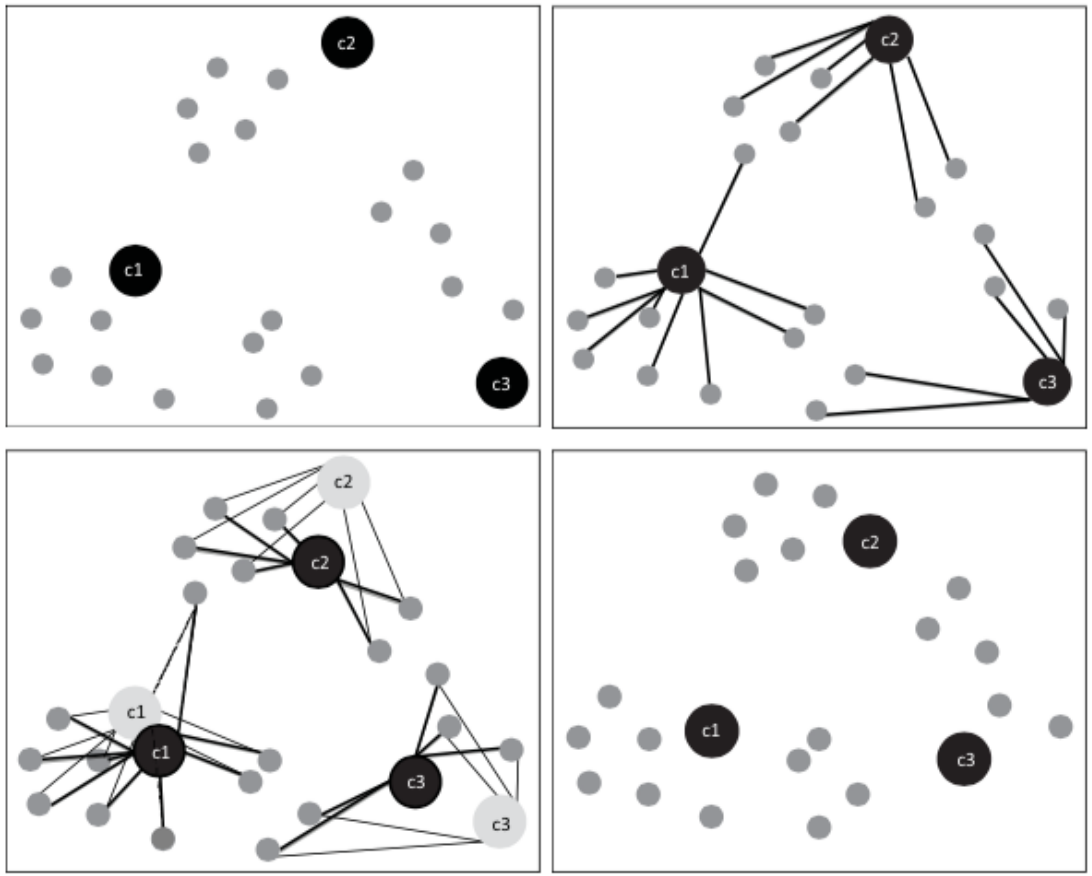
Verdeel x datapunten onder in groepen

Voorbeelden:

- Nieuwsberichten
  - frequency–inverse document frequency (TF-IDF)
- Klanten
- Apps
  - AdGoji
- etc.

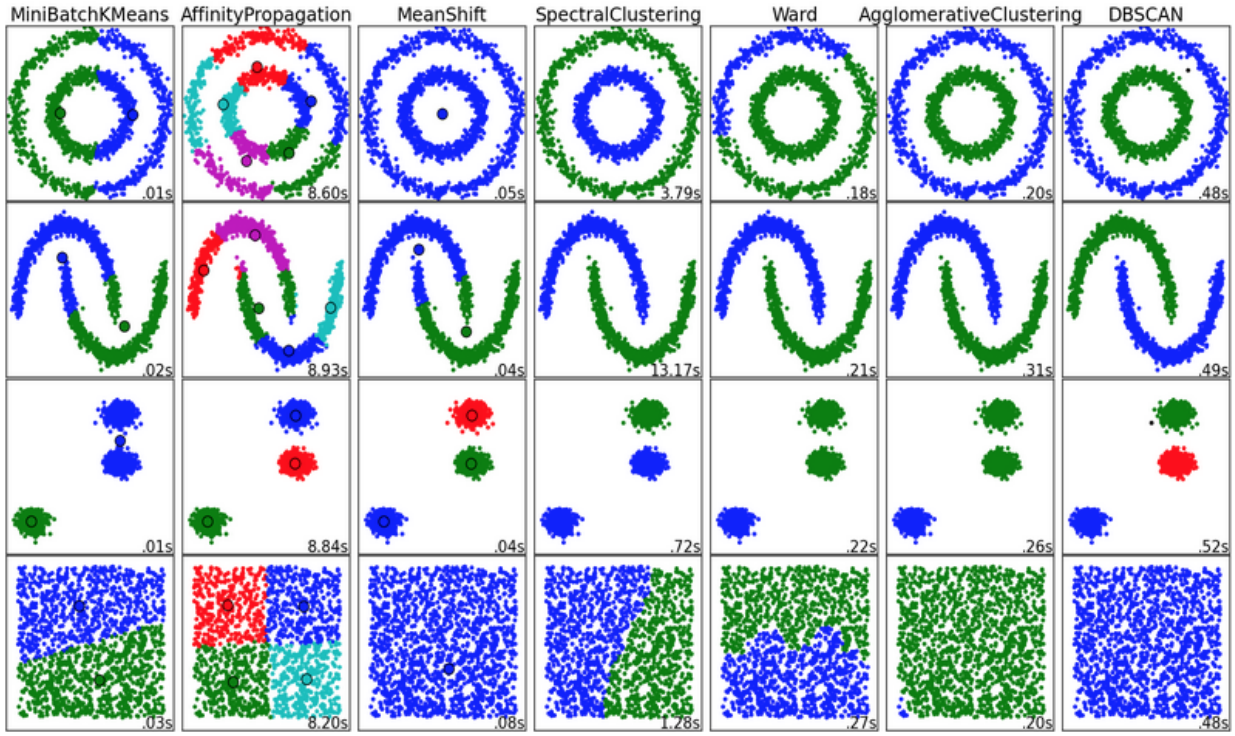
Unsupervised

Mahout – clustering



k-means clustering

Mahout - clustering



verschillende clusteringalgoritmes

Mahout – classificatie

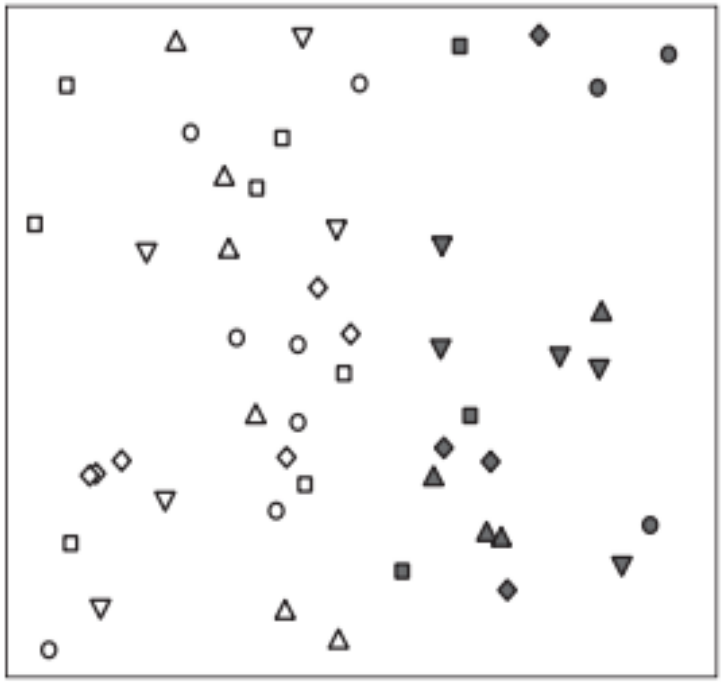
Bepaal van een datapunt x in welke klasse het valt

Voorbeelden:

- Spam filter
- Gezichtsherkenning
- Tekstherkenning
- etc.

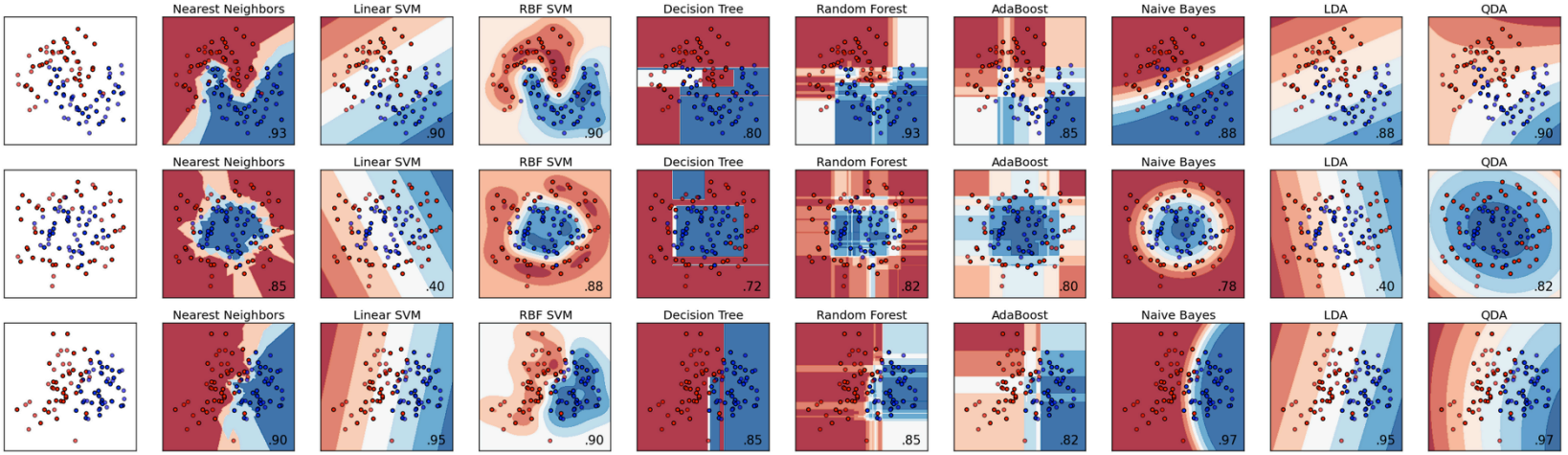
Supervised

Mahout – classificatie



welke kleur krijgt een nieuw punt?

Mahout – classificatie

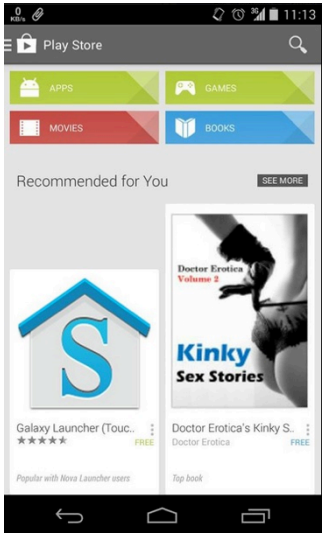


verschillende classificatiealgoritmes

Mahout – recommendation

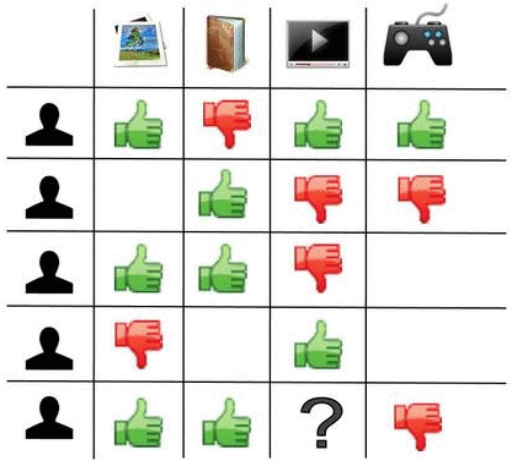
Geef een gebruiker een advies om iets te kopen/luisteren/kijken/etc.

- Boeken
  - "people who bought this, also bought…"
- Films
  - Netflix Prize 1 mln
- Muziek
  - Last.fm, Spotify, Pandora
- Mensen
  - Facebook, Twitter, dating sites
- Ebay
- etc. etc.



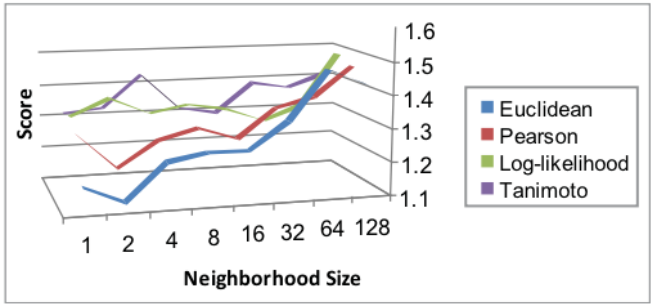
Mahout – recommendation

Collaborative filtering



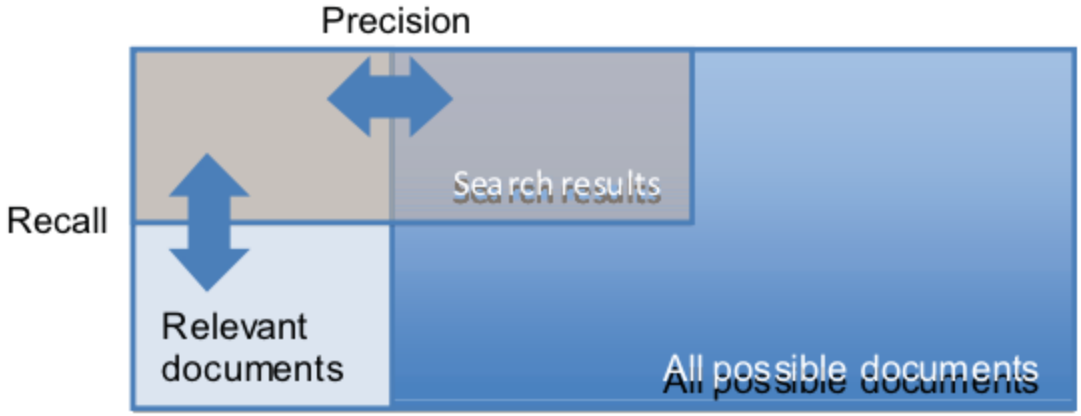
Vind rating *r* voor gebruiker *u* voor item *i*:

- Vind gebruikers die op *u* lijken
- Kijk naar hun ratings voor *i*
- Extrapoleer *u*'s rating voor *i*



Mahout - recommendation

Beoordelen kwaliteit

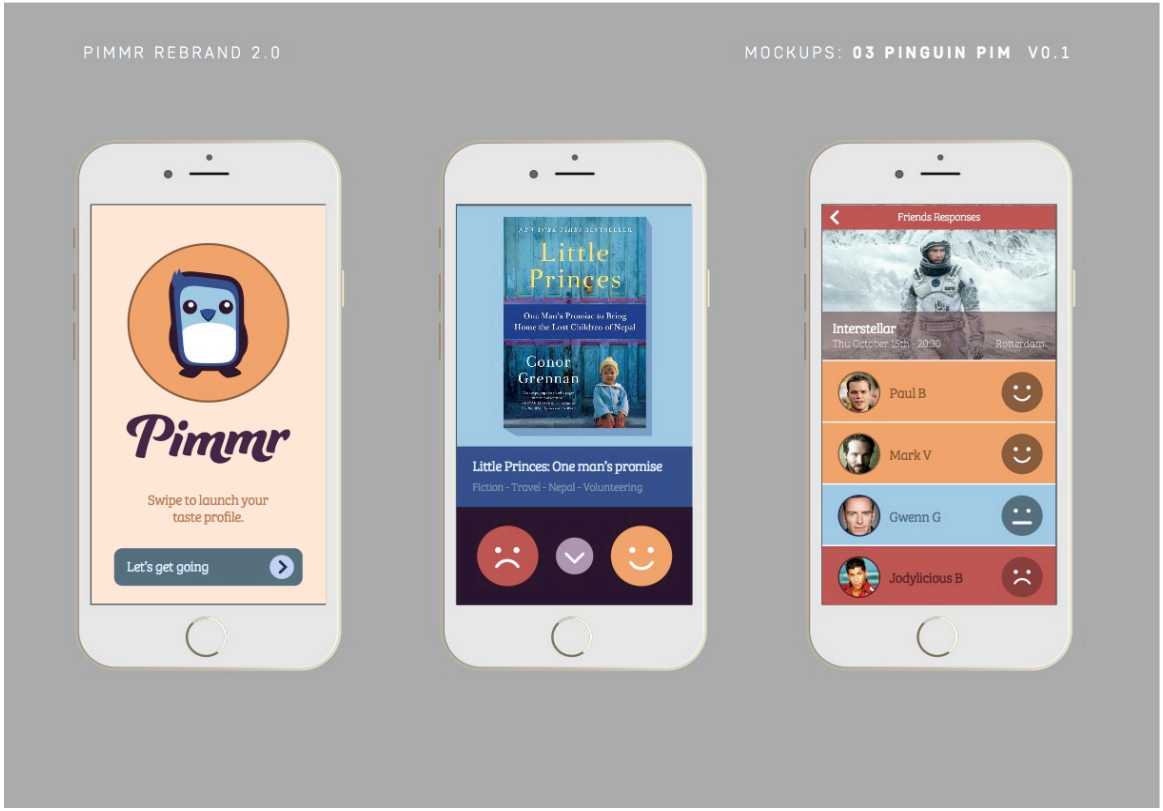


Mahout – recommendation

Voorbeeld: Pimmr

- persoonlijke buddy
- groepsvoorspelling
- cross-domain

Mahout - recommendation



Mahout - recommendation

ml-10k.csv	Random	0	1	0.002	1	0	1	0.05	384	1.46	0.038	0.58	0.428	0.002	0.003	0.008	0.003	0.004	0.92
ml-10k.csv	ItemAverage	0	3	0.012	4	0.004	4	3.5	384	0.84	0.026	0.64	0.242	0.003	0.003	0.008	0.003	0.003	1.47
ml-10k.csv	ItemUserAverage	0	4	0.011	4	0.001	4	2.97	384	0.82	0.043	0.65	0.213	0.003	0.003	0.008	0.003	0.003	1.53
ml-10k.csv	UB Pearson 1nn	0	6	0.002	6	0	6	0.69	384	NaN	0.207	1	0.264	NaN	0	0	NaN	NaN	1.45
ml-10k.csv	UB Pearson 2nn	0	6	0	6	0	6	0.38	384	1.17	0.228	1	0.489	0.037	0.014	0.0019	0.021	0.04	1.08
ml-10k.csv	UB Pearson 4nn	0	7	0.002	7	0	7	0.6	384	0.85	0.31	1	0.432	0.034	0.034	0.0058	0.034	0.041	1.41
ml-10k.csv	UB Pearson th.8	0	7	0.001	7	0	7	1.65	384	0.79	0.201	0.84	0.33	0.041	0.049	0.0076	0.045	0.042	1.08
ml-10k.csv	UB Pearson th.9	0	6	0	6	0	6	0.83	384	0.83	0.16	0.89	0.403	0.044	0.055	0.0076	0.049	0.046	0.92
ml-10k.csv	UB Pearson th.95	0	9	0	9	0	9	0.93	384	0.91	0.161	0.91	0.286	0.043	0.058	0.0076	0.049	0.048	0.89
ml-10k.csv	UB PearsonW 1nn	0	9	0	9	0	9	0.17	384	NaN	0.171	1	0.277	NaN	0	0	NaN	NaN	0.77
ml-10k.csv	UB PearsonW 2nn	0	9	0	9	0	9	0.22	384	1.36	0.158	1	0.275	0.037	0.014	0.0019	0.021	0.04	0.82
ml-10k.csv	UB PearsonW 4nn	0	9	0	9	0	9	0.3	384	0.9	0.177	1	0.268	0.034	0.034	0.0058	0.034	0.041	1.26
ml-10k.csv	UB PearsonW th.8	0	9	0.001	9	0	9	1.16	384	0.9	0.225	0.82	0.436	0.041	0.049	0.0076	0.045	0.042	1.78
ml-10k.csv	UB PearsonW th.9	0	9	0	9	0	9	2.05	384	0.83	0.2	0.89	0.698	0.044	0.055	0.0076	0.049	0.046	1.5
ml-10k.csv	UB PearsonW th.95	0	9	0	9	0	9	1.55	384	0.92	0.366	0.91	0.326	0.043	0.058	0.0076	0.049	0.048	1.07
ml-10k.csv	UB Euclidian 1nn	0	9	0.002	9	0	9	0.33	384	NaN	0.206	1	0.333	NaN	0	0	NaN	NaN	0.8
ml-10k.csv	UB Euclidian 2nn	0	9	0	9	0	9	0.19	384	0.9	0.153	1	0.241	0.043	0.006	0.001	0.01	0.042	0.83
ml-10k.csv	UB Euclidian 4nn	0	9	0	9	0	9	0.29	384	0.85	0.188	1	0.248	0.042	0.027	0.004	0.033	0.048	0.81
ml-10k.csv	UB Euclidian th.8	0	9	0.001	9	0.003	9	2.13	384	0.81	0.139	0.87	0.267	0.034	0.041	0.0078	0.037	0.031	0.93
ml-10k.csv	UB Euclidian th.9	0	8	0.001	8	0.002	8	1.96	384	0.9	0.127	0.87	0.279	0.034	0.041	0.0078	0.037	0.031	1.07
ml-10k.csv	UB Euclidian th.95	0	8	0	8	0.002	8	1.97	384	0.92	0.14	0.89	0.307	0.034	0.041	0.0078	0.037	0.031	0.91
ml-10k.csv	UB EuclidianW 1nn	0	8	0	8	0	8	0.18	384	NaN	0.175	1	0.235	NaN	0	0	NaN	NaN	0.79
ml-10k.csv	UB EuclidianW 2nn	0	8	0	8	0	8	0.22	384	1.17	0.139	1	0.288	0.043	0.006	0.001	0.01	0.042	0.86
ml-10k.csv	UB EuclidianW 4nn	0	9	0	9	0	9	0.33	384	0.75	0.156	1	0.257	0.042	0.027	0.004	0.033	0.048	0.87
ml-10k.csv	UB EuclidianW th.8	0	9	0	9	0.002	9	2.09	384	0.91	0.173	0.89	0.265	0.034	0.041	0.0078	0.037	0.031	0.95
ml-10k.csv	UB EuclidianW th.9	0	8	0.001	8	0.002	8	2.68	384	0.85	0.266	0.87	0.652	0.034	0.041	0.0078	0.037	0.031	2.13
ml-10k.csv	UB EuclidianW th.95	0	8	0	8	0.003	8	2.17	384	0.85	0.183	0.83	0.277	0.034	0.041	0.0078	0.037	0.031	0.93
ml-10k.csv	UB Spearman th.5	0	8	0.019	8	0.005	8	7.83	384	0.84	2.35	0.74	5.206	0.021	0.025	0.0079	0.023	0.018	2.15
ml-10k.csv	UB Spearman th.7	0	5	0.011	5	0.002	5	0.88	2.819	0.8	4.51	0.032	0.039	0.0078	0.035	0.028	1.63		
ml-10k.csv	UB Spearman th.9	0	4	0.005	4	0.002	4	5.11	384	0.91	2.131	0.83	4.073	0.031	0.045	0.0078	0.037	0.032	1.74
ml-10k.csv	IB Pearson	0	4	0.107	14	0.01	17	47.58	384	1.15	0.11	0.69	0.138	0.023	0.03	0.0078	0.026	0.031	6.4
ml-10k.csv	IB PearsonW	0	5	0.078	15	0.009	15	55.47	384	1.17	0.048	0.68	0.152	0.023	0.03	0.0078	0.026	0.031	6.43
ml-10k.csv	IB Euclidian	0	3	0.108	15	0.015	15	54.65	384	0.82	0.041	0.6	0.226	0	0	0.008	NaN	0	6.47
ml-10k.csv	IB EuclidianW	0	5	0.072	16	0.008	16	48.2	384	0.84	0.08	0.62	0.404	0	0	0.008	NaN	0	7.83
ml-10k.csv	IB Tanimoto	0	4	0.104	15	0.009	15	47.15	384	0.82	0.108	0.62	0.18	0	0	0.008	NaN	0	6.25
ml-10k.csv	IB LogLikelihood	0	4	0.129	14	0.015	17	53.81	384	0.85	0.119	0.62	0.262	0	0	0.008	NaN	0	8.16
ml-10k.csv	BIB Pearson	0	4	0.224	20	0.009	23	48.87	384	0.84	0.143	0.75	0.166	0.023	0.028	0.0072	0.025	0.023	7.35
ml-10k.csv	BIB Euclidian	0	3	0.096	19	0.009	21	48.68	384	0.78	0.059	0.65	0.138	0.025	0.029	0.0077	0.027	0.026	9.23
ml-10k.csv	SlopeOne	0	3	0.271	25	0.024	25	9.05	384	0.9	0.296	0.67	0.423	0.001	0.002	0.008	0.002	0.002	18.91
ml-10k.csv	SlopeOneMem	0	6	0.138	27	0.001	27	8.72	384	0.85	0.108	0.7	0.273	0.001	0.002	0.008	0.002	0.002	21.11
ml-10k.csv	SVG_ALS	0	6	1.972	74	0	74	2.43	384	0.99	1.525	0.65	1.998						

Waarom Mahout?

- Volwassen machine learning library
- Open-source
- Makkelijk uitbreidbaar
- Single server en Hadoop

Waarom niet?

- Minder actief ontwikkeld
- Richting onduidelijk
- Steeds meer concurrentie
  - MLLib, Oryx, H2O, …

Mahout – hoe te beginnen

```
1,10,1.0
1,11,2.0
1,12,5.0
1,13,5.0
1,14,5.0
1,15,4.0
...
...
```

- Installeer Java
- Download Mahout jar
- Maak test set
-

```
DataModel model = new FileDataModel(new File("/path/to/dataset.csv"));
UserSimilarity similarity = new PearsonCorrelationSimilarity(model);
UserNeighborhood neighborhood =
    new ThresholdUserNeighborhood(0.1, similarity, model);
UserBasedRecommender recommender =
    new GenericUserBasedRecommender(model, neighborhood, similarity);
List recommendations = recommender.recommend(2, 3);
for (RecommendedItem recommendation : recommendations) {
  System.out.println(recommendation);
}
```

- Run

- quiz-

Welke drie dingen kan Mahout?

Geef van elk een voorbeeld

Is Mahout small of big data?

– hersengymnastiek –

Welke drie dingen kan Mahout?

Clustering, classificatie, recommendation

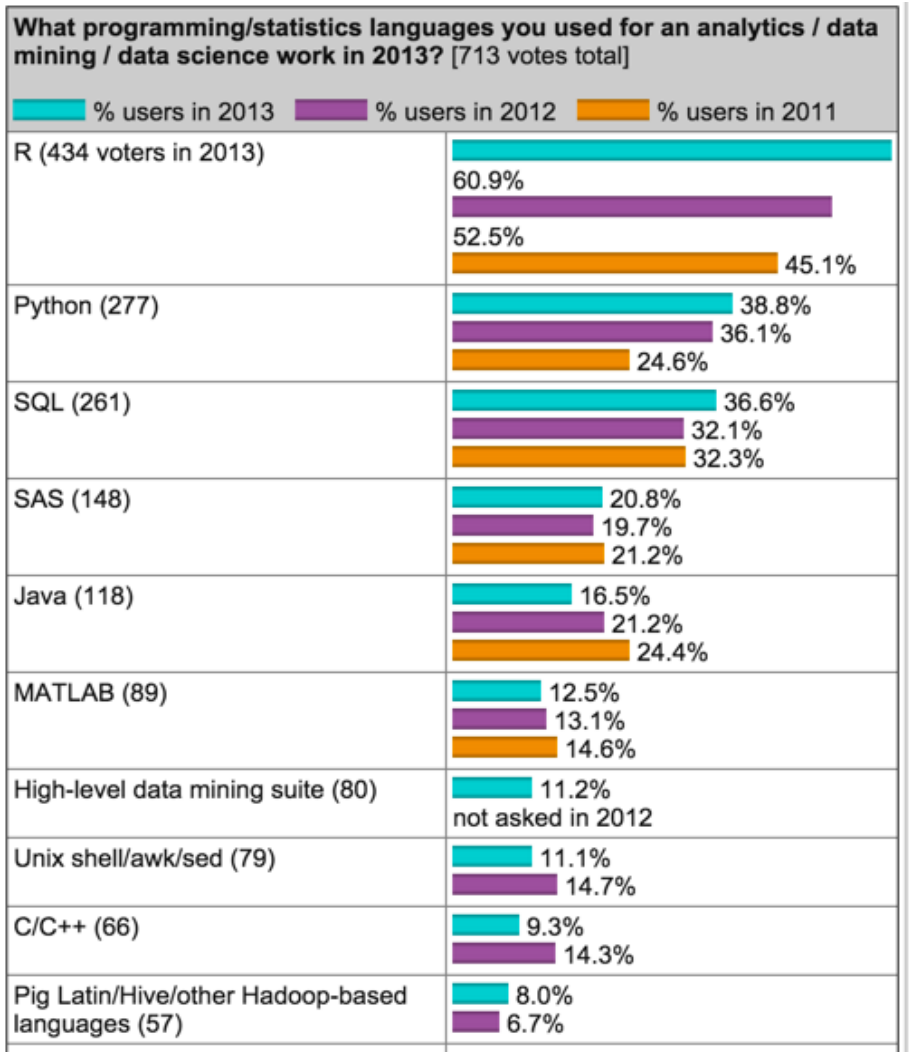
Geef van elk een voorbeeld

Klanten, Spam, Amazon

Is Mahout small of big data?

Beide, hoera!

What programming/statistics languages you used for an analytics / data mining / data science work in 2013? [713 votes total]



R

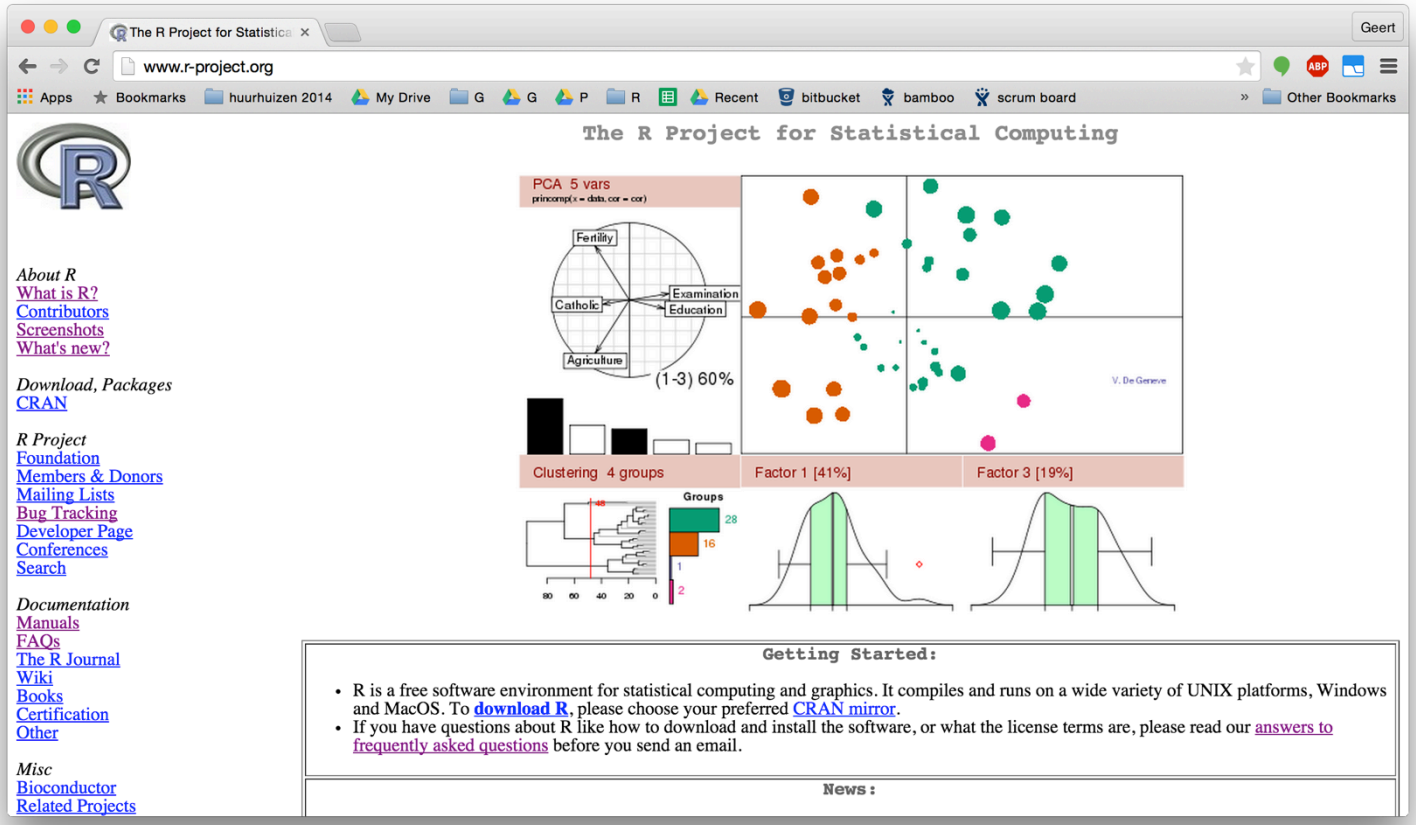
Programmeertaal en -omgeving voor statistische berekeningen en graphics

- 1993 (S: 1976)
- Steeds populairder
- Beetje gekke syntax
- Ongeëvenaard aantal packages voor allerlei doeleinden
- R programmeurs goed betaald, zelfs vergeleken met MapReduce, NoSQL, Cassandra

- Single-threaded
- Bedoeld om op een desktop machine te draaien
- Alles in-memory

-> R schaalt niet

R


The R Project for Statistica
Geert
www.r-project.org
Apps
Bookmarks
huurhuizen 2014
My Drive
G
G
P
R
Recent
bitbucket
bamboo
scrum board
Other Bookmarks
The R Project for Statistical Computing
About R
What is R?
Contributors
Screenshots
What's new?
Download, Packages
CRAN
R Project
Foundation
Members & Donors
Mailing Lists
Bug Tracking
Developer Page
Conferences
Search
Documentation
Manuals
FAQs
The R Journal
Wiki
Books
Certification
Other
Misc
Bioconductor
Related Projects
PCA 5 vars
Fertility
Catholic
Examination
Education
Agriculture
(1-3) 60%
Clustering 4 groups
Factor 1 [41%]
Factor 3 [19%]
Groups
Getting Started:
R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.
If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.
News:

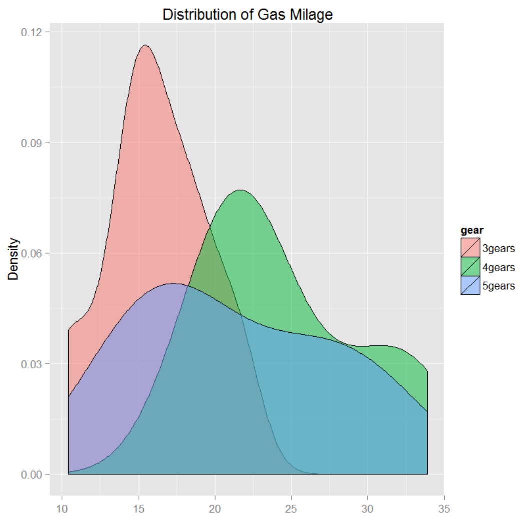
R

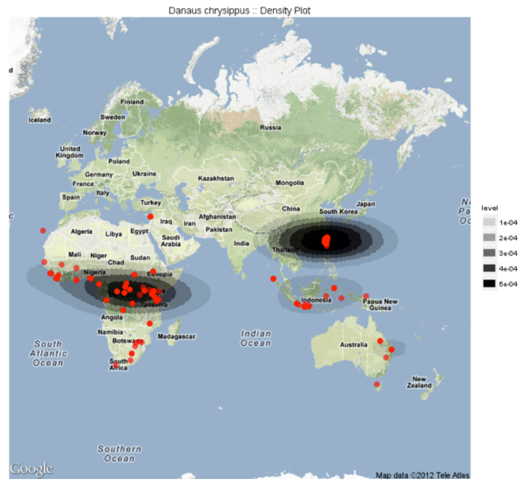
Analyseren data, plotten, reporting

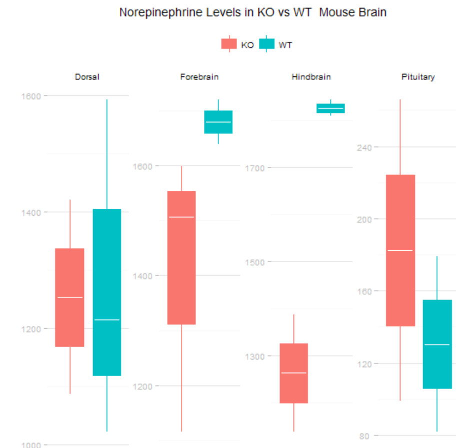
Voorbeelden:

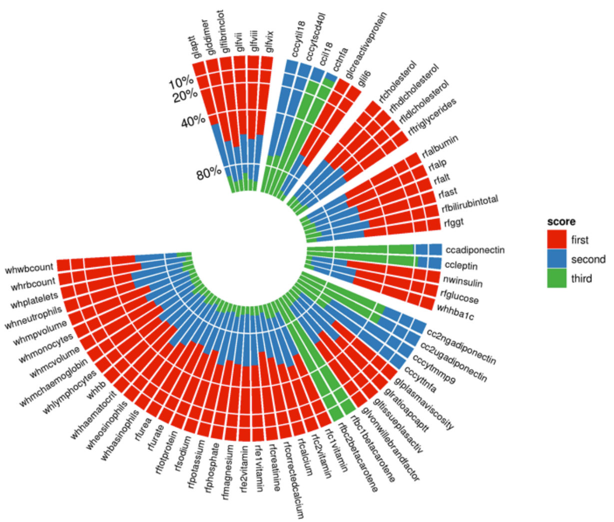
- Huizenprijsindex analyseren en voorspellen
- Voorspelling klantrespons, prioriteitsadvies
- Verbeteren advertentieplaatsing (AdWords)
- Stekelige geopolitieke vragen (Benetech, Saving the World with R)
- enz. enz.

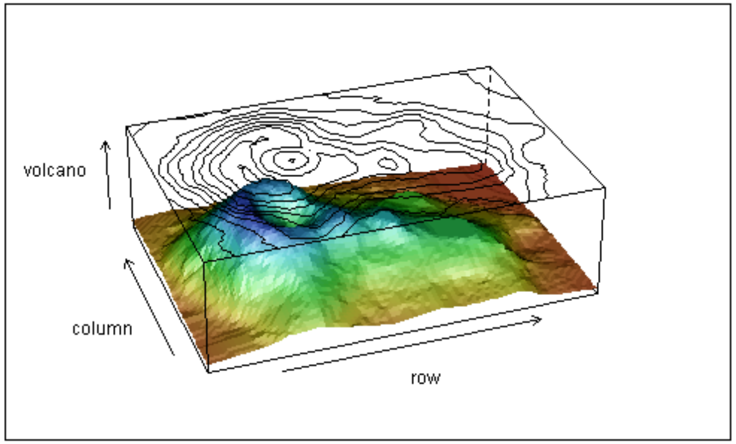
Alle grote bedrijven gebruiken het











R – Grip.QA

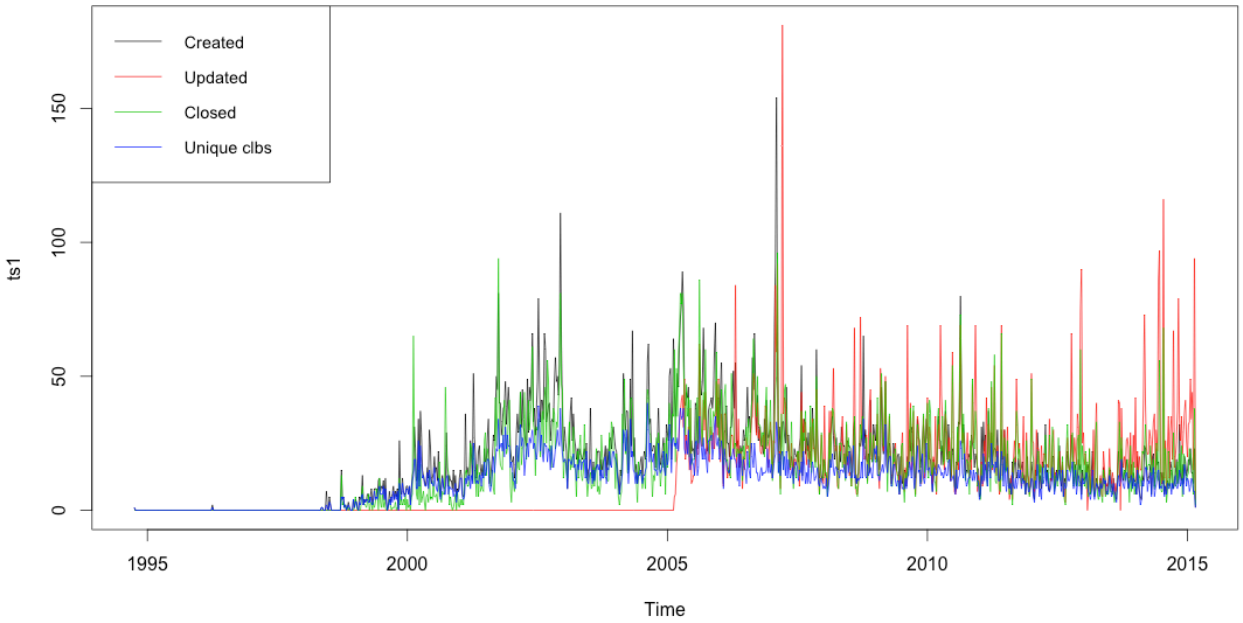
Grip.QA

Voorbeeld: Grip.QA

- Voorspellen kwaliteit software terwijl het gebouwd wordt
- Meet alles
- Ontdekken verbanden met doelvariabelen: kosten, snelheid, kwaliteit
- Voorspellen van:
  - verloop
  - verloop bij ingrijpen

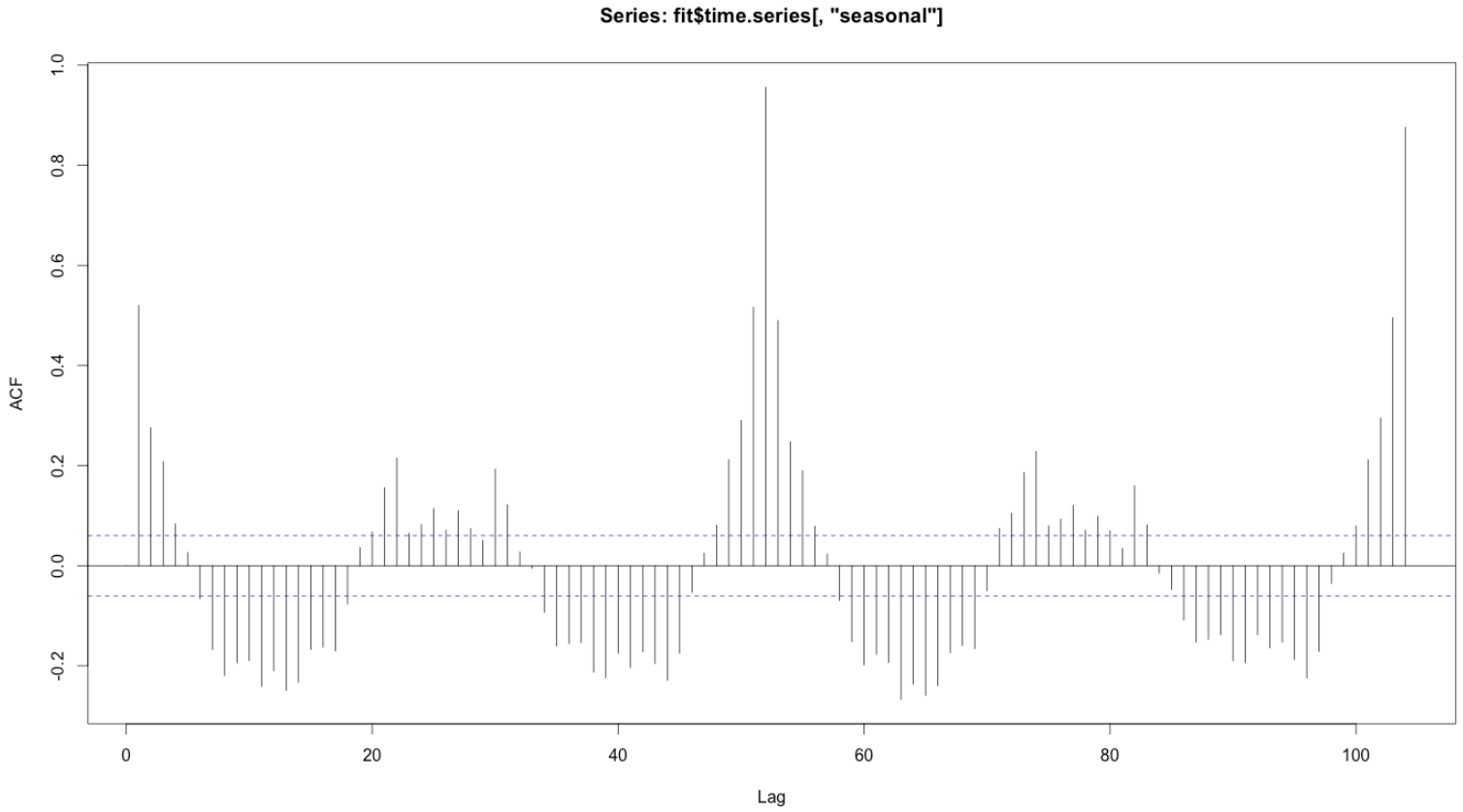
R – Grip.QA

Grip.QA



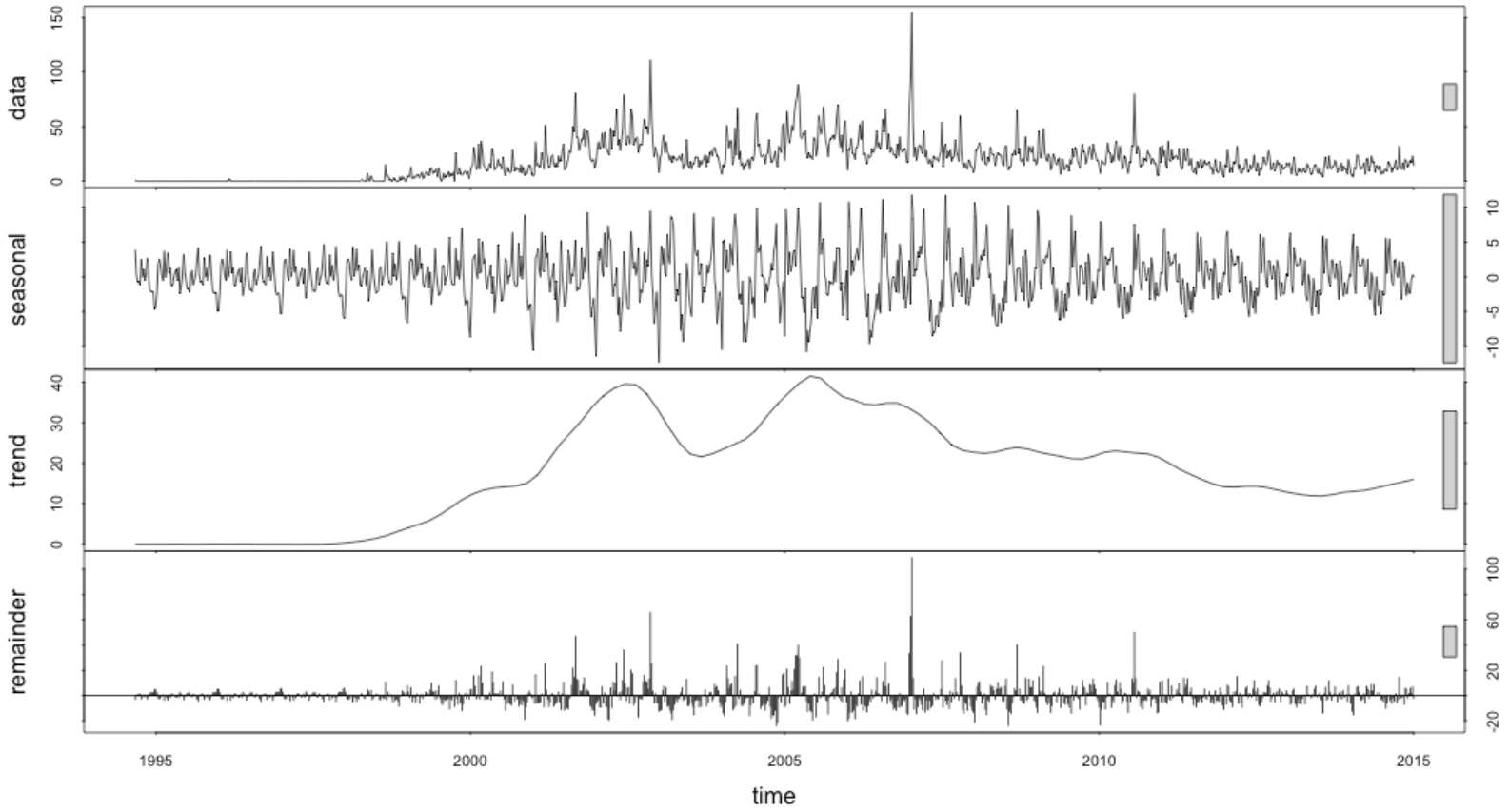
R – Grip.QA

Grip.QA



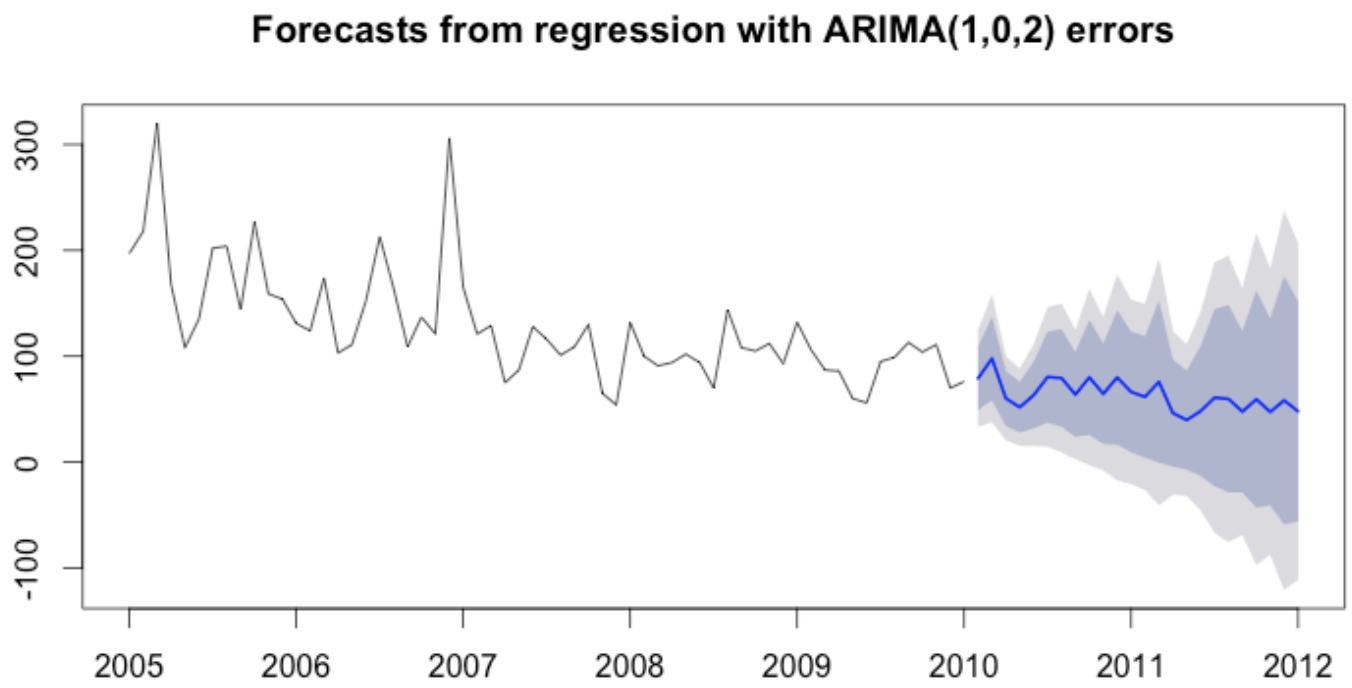
R – Grip.QA

Grip.QA



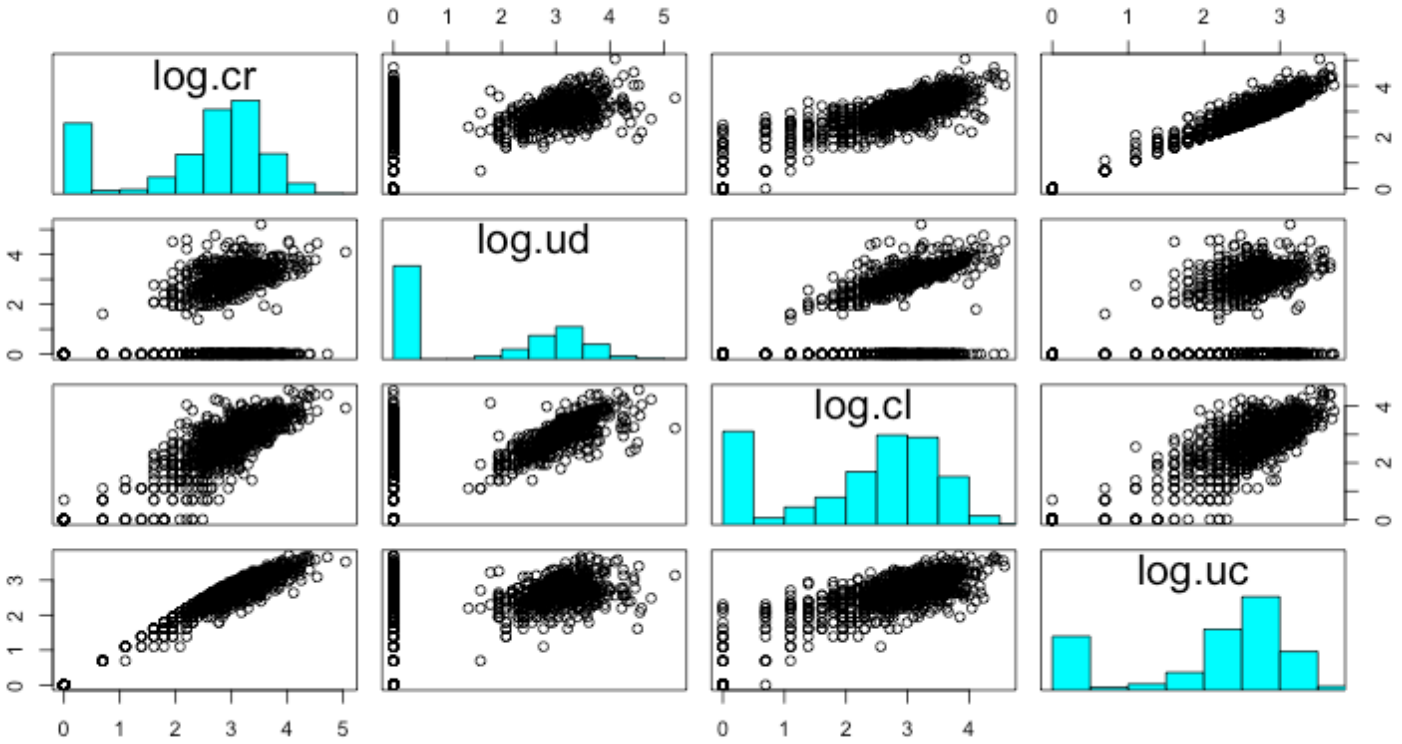
R – Grip.QA

Grip.QA



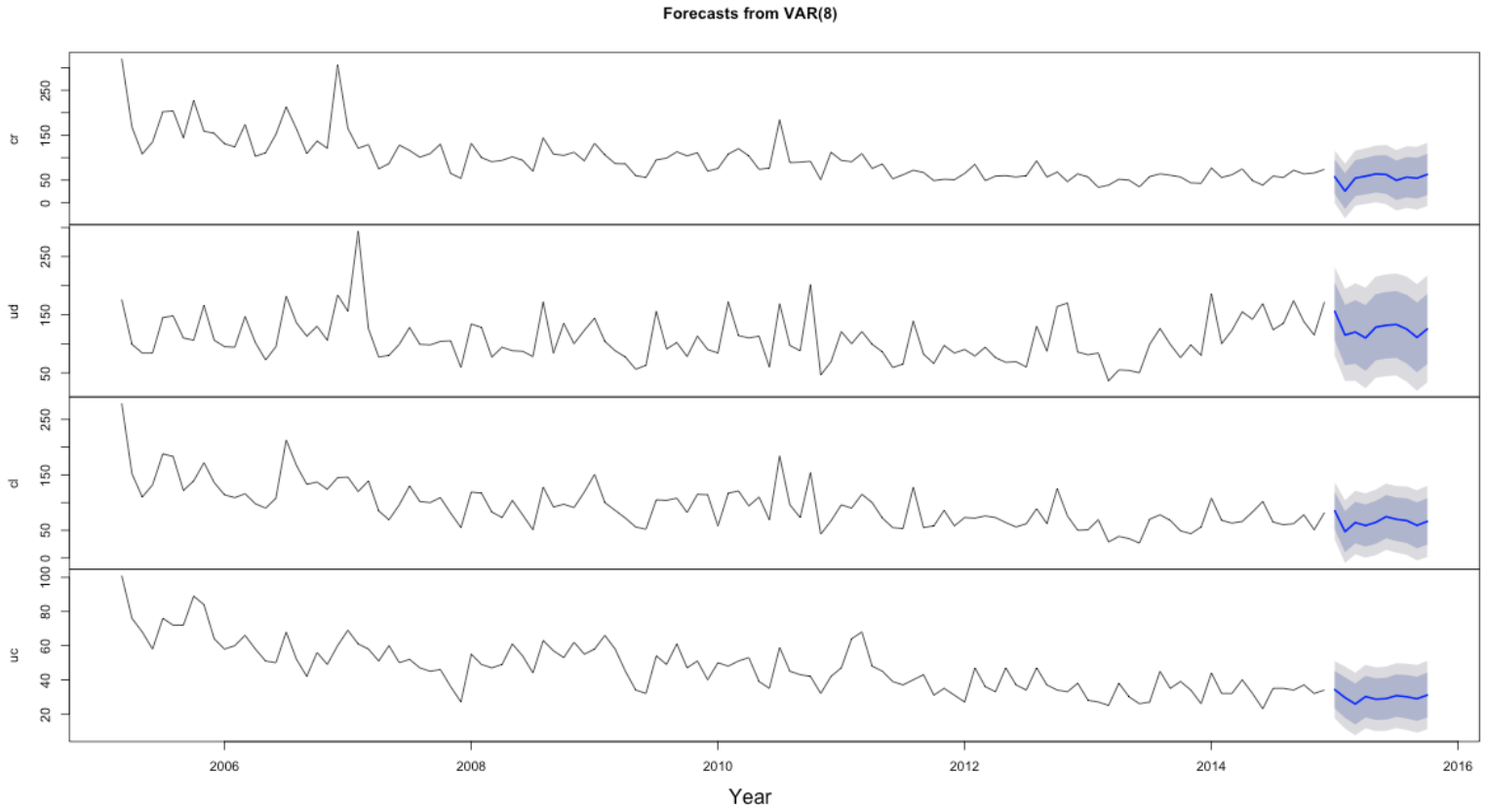
R – Grip.QA

Grip.QA



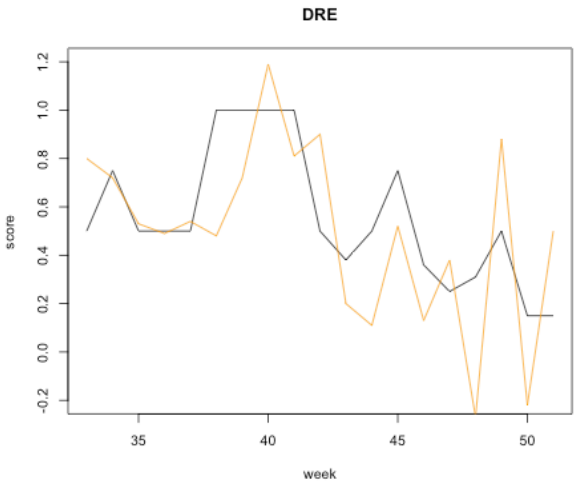
R – Grip.QA

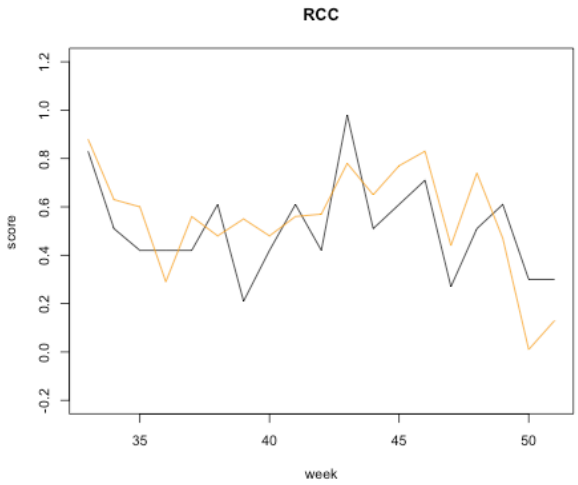
Grip.QA

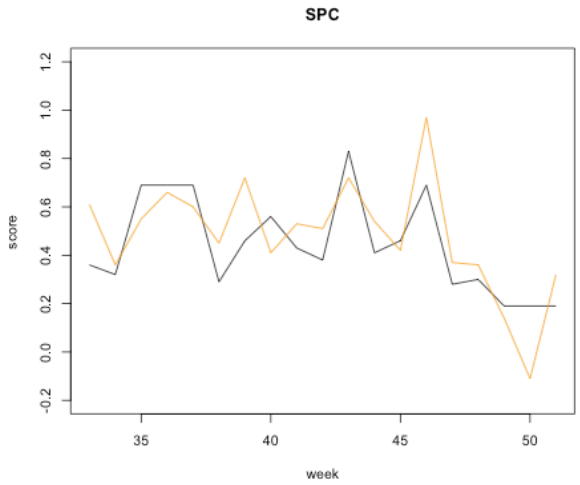


R – Grip.QA

Grip.QA







Waarom R?

- Populairste data science taal
- Snel ontwikkelen
  - Matlabachtige syntax
  - REPL
- Open source
- Libraries

Waarom niet?

- Performance
- Schaalt niet
- One-off

R – hoe te beginnen

- Installeer R
- Laad data
- Analyseer, plot, etc.

```
> returns <- read.csv("~/Downloads/regression-example-gnu-r.csv", header=TRUE)
>
> head(returns)
           USA        CANADA       GERMANY
1  0.0021000000  2.502450e-03  0.0026161513
2  0.0032930845  5.508304e-03  0.0030739634
3  0.0004973145 -3.635486e-06  0.0005597597
4  0.0046724326  4.169140e-03  0.0048291874
5  0.0015832179  3.591204e-03  0.0025519577
6 -0.0007903576  4.091737e-04 -0.0011018125
>
> returns.lm <- lm( returns$CANADA ~ returns$USA )
>
> summary(returns.lm)

Call:
lm(formula = returns$CANADA ~ returns$USA)

Residuals:
      Min        1Q    Median        3Q       Max
-0.045410 -0.001718  0.000031  0.001740  0.034456

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.921e-05  3.671e-05   1.613    0.107
returns$USA 8.383e-01  4.411e-03 190.062   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.003727 on 10332 degrees of freedom
Multiple R-squared:  0.7776,    Adjusted R-squared:  0.7776
F-statistic: 3.612e+04 on 1 and 10332 DF,  p-value: < 2.2e-16
```

- quiz -

Waar wordt R voor gebruikt?

Wat is R's zwakte?

Is R small of big data?

– hersengymnastiek –

Waar wordt R voor gebruikt?

Bewerken, analyseren, plotten data

Wat is R's zwakte?

Performance

Is R small of big data?

Small data