

Big Data Analytics

Thomas Götz

who am i

- Dipl. Physiker
- Seit 1999 im Bereich IT-Security tätig
- 1998: Erster 3D Hardware Treiber für X11 unter Linux (Matrox)

Was ist Big Data Analytics

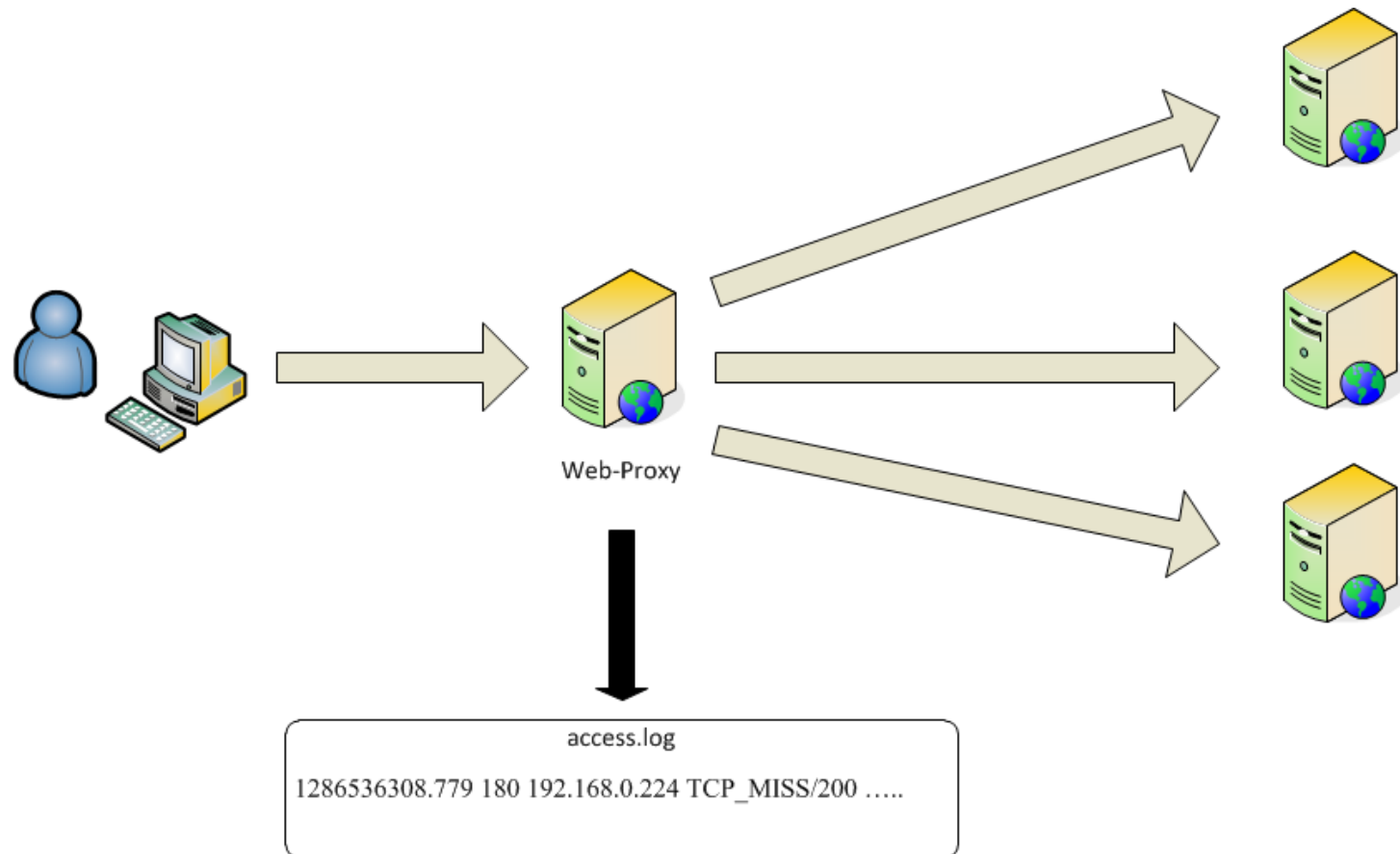
Definition:

Big data analytics is the process of examining large data sets containing a variety of data types ... to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information

BEISPIELANWENDUNG

Analyse einer Squid access.log

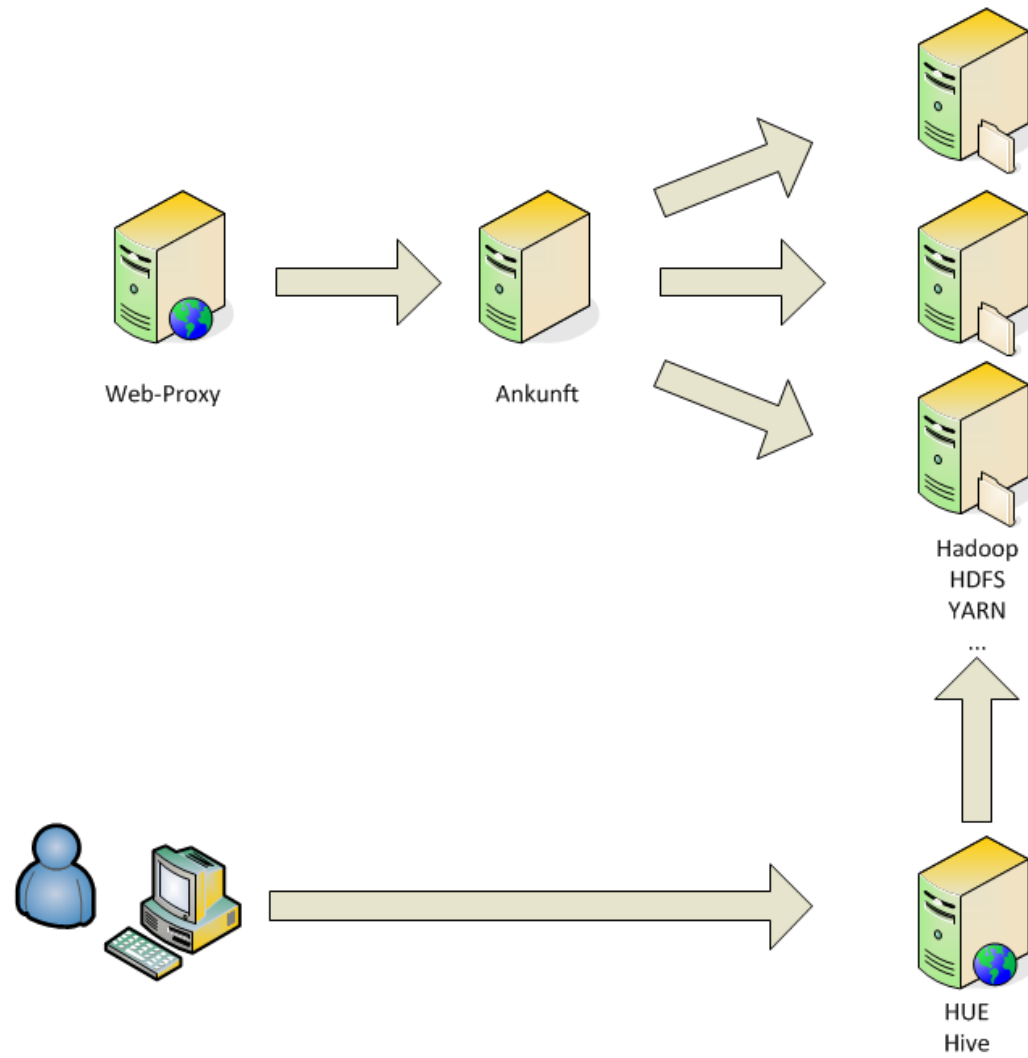
Squid Web-Proxy



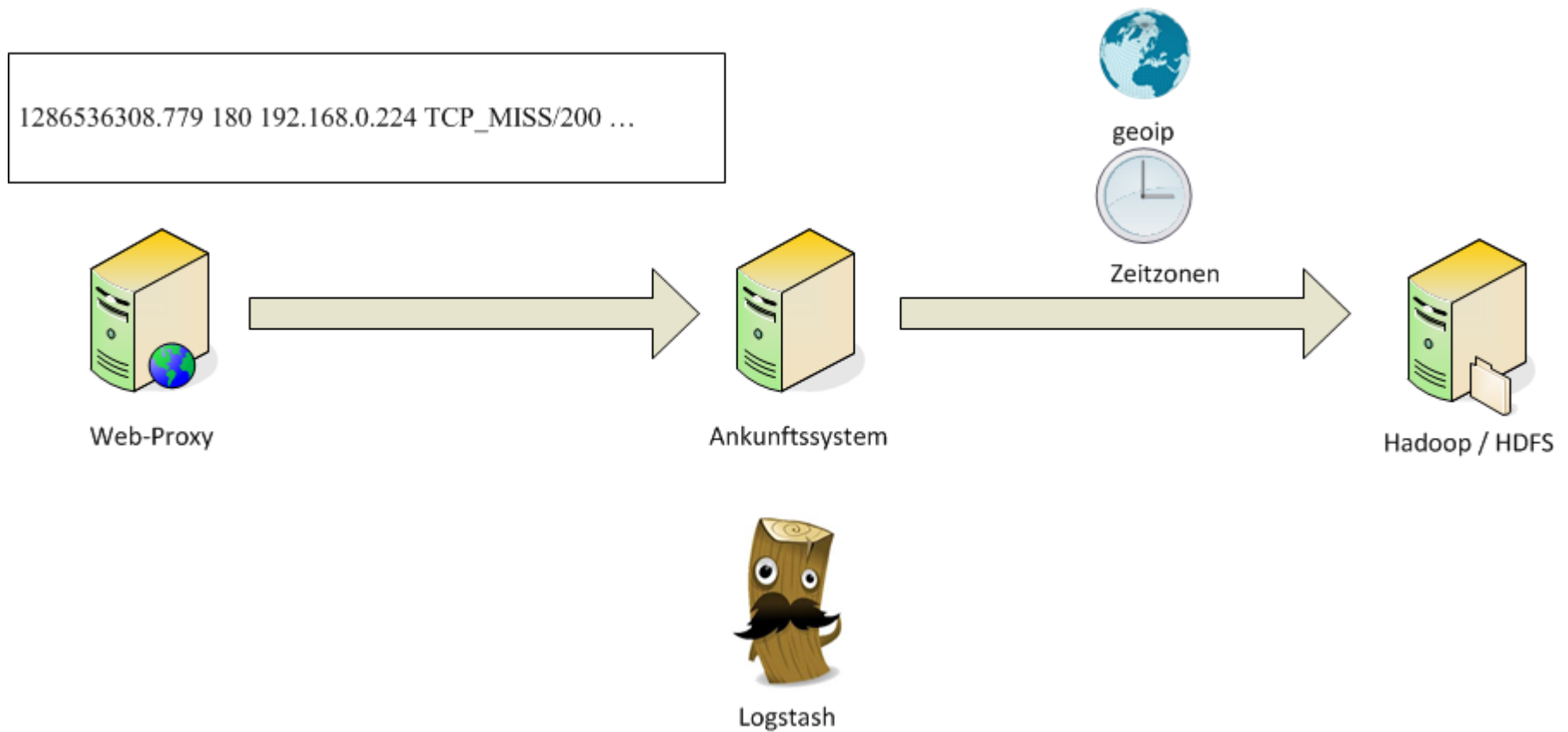
Squid access.log Format

```
logformat custom %>a %<A %ui %un %ts %>st  
%<st "%rm %>ru HTTP/%rv" %>Hs "%{Referer}  
>h" "%{User-Agent}>h" "%>h" %Ss:%Sh %<a  
%<p %<lp
```

Aufbau



Datenspeicherung im Detail



Logstash Configuration

```
filter {
  grok {
    match => [ "message" , '%{IP:clientip} %{DATA:targethost} %{DATA:user} %{DATA:auth} %{DATA:timestamp} %{NUMBER:reqbytes} %{NUMBER:respbytes} "(?:%{WORD:verb} %{NOTSPACE:request}(?: HTTP/%{NUMBER:httpversion})?|%{DATA:rawrequest})" %{NUMBER:response} "%{DATA:referrer}" "%{DATA:useragent}" "%{DATA:fullrequest}" %{WORD:cache}:%{WORD:hierarchy}' ]
    add_field => { "timestamp" => "%{timestamp}" }
  }

  date {
    match => [ "timestamp" , "UNIX" ]
  }

  geoip {
    source => "targethost"
    add_tag => [ "geoip" ]
  }
}
```

Logstash Configuration

```
output {  
  csv {  
    path => "/home/ankunft/incoming/access.log"  
    csv_options => {"col_sep" => " " "row_sep" => "  
" }  
    codec => plain  
    fields => ["clientip", "request", "timestamp[0]",  
"reqbytes", "respbytes", "verb", "httpversion" ,  
"response", "referer" , "geoip[ip]", "geoip[latitude]",  
"geoip[longitude]", "geoip[country_code3]" ,  
"geoip[continent_code]",  
"geoip[timezone]" , "fullrequest" ]  
    max_size => 1000  
  }  
}
```

logrotate

```
/home/ankunft/incoming/*.log {  
    size 10k  
    rotate 7  
    copytruncate  
    missingok  
    su ankunft users  
    postrotate /home/ankunft/bin/move-to-hdfs  
    endscript  
}
```

Daten ins HDFS

```
#!/bin/bash

INCOMING="/home/ankunft/incoming"
INCOMING_HDFS="/user/ankunft/incoming"

for i in ${INCOMING}/*.log.[0-9]*
do
    BN=`basename $i`
    echo "move to hdfs: $i"
    su -c "/usr/bin/hdfs dfs -put $i /user/ankunft/incoming" ankunft
    if [[ $? -eq 0 ]]; then
        rm $i
        /usr/bin/beeline -u "jdbc:hive2://localhost:10000/" -n ankunft
    << EOF
    LOAD DATA INPATH "$INCOMING_HDFS/$BN" INTO TABLE access_log;
    EOF
    fi
done
```

Schema im Hive

 access_log	 
 clientip (string)	
 url (string)	
 datetime (int)	
 reqbytes (bigint)	
 respbytes (bigint)	
 method (string)	
 httpversion (string)	
 response (int)	
 refererer (string)	
 geoip_ip (string)	
 latitude (float)	
 longitude (float)	
 country (string)	
 continent (string)	
 timezone (string)	
 fullrequest (string)	

USE-CASES

Analyse einer Squid access.log

Use-Cases

- Welche ungewöhnlichen Verbindungen gibt es in meinen HTTP Proxy Logs

Use-Cases

- Welche ungewöhnlichen Verbindungen gibt es in meinen HTTP Proxy Logs
- Was bedeutet ungewöhnlich?

Hypothesen oder was bedeutet „ungewöhnlich“

Ungewöhnlich sind:

- Server mit einem hohen Upload / Download Datenvolumenverhältnis
- Zugriffe auf Länder, bei denen sich die Anzahl der Requests stark ändert
- Server, bei denen nur auf wenige unterschiedlichen Pfade zugegriffen wird

USE-CASES MIT HIVE

Analyse einer Squid access.log

Hive Anfragen

Server mit einem hohen Upload / Download
Datenvolumenverhältnis

```
SELECT
    sum(t.reqbytes) / sum(t.respbytes) AS ratio,
    t.host
FROM
    ( SELECT reqbytes,respbytes,
        regexp_extract(url,'http://([^/]*)/.*',1) AS host
      FROM access_log ) t

GROUP BY t.host;
```

Hive Anfragen – Upload / Download

Aktuelle Abfragen			Abfrage	Protokoll	Spalten	Ergebnisse	Diagramm
	ratio	host					
18	6.279245283018868	36c3feec.mpstat.us					
9	1.5151599058562923	10.11.0.209:8008					
28	0.9566294919454771	a.analytics.yahoo.com					
14	0.8524590163934426	20562659p.rfihub.com					
10	0.8173207036535859	10.11.0.214					
25	0.515185601799775	5c4cf848f6454dc02ec8-c49fe7e7355d384845270f4a7a0a7aa1.r53.cf2.rackcdn.com					
5	0.3196861321661151	1.gravatar.com					
13	0.3156168207991624	2.gravatar.com					
6	0.31440329218106994	10.11.0.209:41562					
7	0.31440329218106994	10.11.0.209:49698					
8	0.31440329218106994	10.11.0.209:55275					
1	0.2909401478124102	0.gravatar.com					
23	0.24189765868286495	46.37.47.28					
0	0.21664810531726364						
22	0.18460281805186848	4368131.fls.doubleclick.net					
26	0.17511278984366804	80bola.com					
11	0.16805362921867775	2.bp.blogspot.com					
27	0.15921267835993458	a.adroll.com					
15	0.08979823813583404	209.152.160.51					
3	0.05893309136061894	1.bp.blogspot.com					
16	0.04656781889995453	3.bp.blogspot.com					
29	0.04384023837171089	a.disquscdn.com					
30	0.03112409657081347	a.tile.openstreetmap.org					
20	0.030579459097004918	4.bp.blogspot.com					
31	0.021905688779589427	a.tile.osm.org					
4	0.019609196923086106	1.f.ix.de					

Hive Anfragen

Zugriffe auf Länder, bei denen sich die Anzahl der Requests stark ändert

```
select country, stddev_pop(cnt)/avg(cnt)  
as dev , avg(cnt) as av FROM
```

```
(select country,day,count(1) as cnt from  
  ( select country,INT(datetime/86400) as  
day from          access_log where country !=  
' ' ) t group by country, day) u
```

```
group by country;
```

Ergebnis

◆	◆ country	◆ dev	◆ av
2	CAN	0	5
5	CZE	0	16
11	ISR	0	36
13	NOR	0	1
14	POL	0	1
15	ROU	0	1
16	RUS	0	1
17	SWE	0	3
4	CRI	0.4727809973919772	19.75
0	AUS	0.6	7.5
8	FRA	0.66111187669806	27.875
3	CHE	0.7144345083117603	8
7	EU	0.8184405248499284	95.18181818181819
9	GBR	0.8612083217861307	207.77777777777777
6	DEU	0.9745137059011629	1040.1818181818182
12	NLD	0.9809648095633724	92.88888888888889
1	AUT	1.054744109602881	29
18	USA	1.073216825080365	2183.818181818182
10	IRL	1.3377455343701634	101

Hive Anfragen

Server, bei denen nur auf wenige unterschiedlichen Pfade zugegriffen wird

Hive Anfragen

- Server, bei denen nur auf wenige unterschiedliche Pfade zugegriffen wird

Entropie

$$H(X) = - \sum_i P(x_i) \log_b P(x_i)$$

Hive User Defined Functions

- UDF: User Defined Function
 - Funktionen, die auf einzelne Elementen operieren
z.B. Zeitkonvertierungen, String-Operationen
- UDAF: User Defined Aggregation Function
 - Funktionen, die auf spalten operieren
z.B. Summenbildung

UDAF - entropy

entropy(String) → Double

...

...

```
public Double terminate() {  
    double entropy = 0.0;  
  
    for (MyInt my : state.values()) {  
        int i=my.count;  
        n+=i;  
        entropy -= i*Math.log(i);  
    }  
    entropy = Math.log(n) + entropy/n;  
  
    return entropy;  
}
```

...

...

Hive Anfragen

```
SELECT
    host,
    entropy(path) as entropy
FROM
    (SELECT
        parse_url(url, 'PATH') as path,
        parse_url(url, 'HOST') as host
    FROM access_log) t

WHERE isnotnull(host) and
      isnotnull(path)
GROUP BY host;
```

Entropie-Ergebnisse

◆	◆ host	▲ entropy
77	api.zanox.com	3.556887746287509
78	api.zanox.ws	3.556980740358072
0	0.gravatar.com	3.6690672970843856
1	059-ylz-577.mktoresp.com	3.6894539788655605
2	1.bp.blogspot.com	3.9112249685883853
76	api.smoot.apple.com	3.958540577476546
5	10.11.0.209	4.844782584659884
3	1.f.ix.de	4.999661382910399
47	ad2.adfarm1.adition.com	5.141914719148722
4	1.gravatar.com	5.142110189258755
48	ad4.adfarm1.adition.com	5.142898345953139
6	2.bp.blogspot.com	5.145341691757211
45	ad.yieldmanager.com	5.167506868075114
50	adfarm1.adition.com	5.173984385775848
46	ad1.adfarm1.adition.com	5.191728663110613
44	ad.yieldlab.net	5.234676548842
49	adadvisor.net	5.281794764135622
7	2.f.ix.de	5.738403352009112
10	20562659p.rfihub.com	5.899242585171697
12	3.bp.blogspot.com	5.899518444393779
9	20562657p.rfihub.com	5.905812839217957

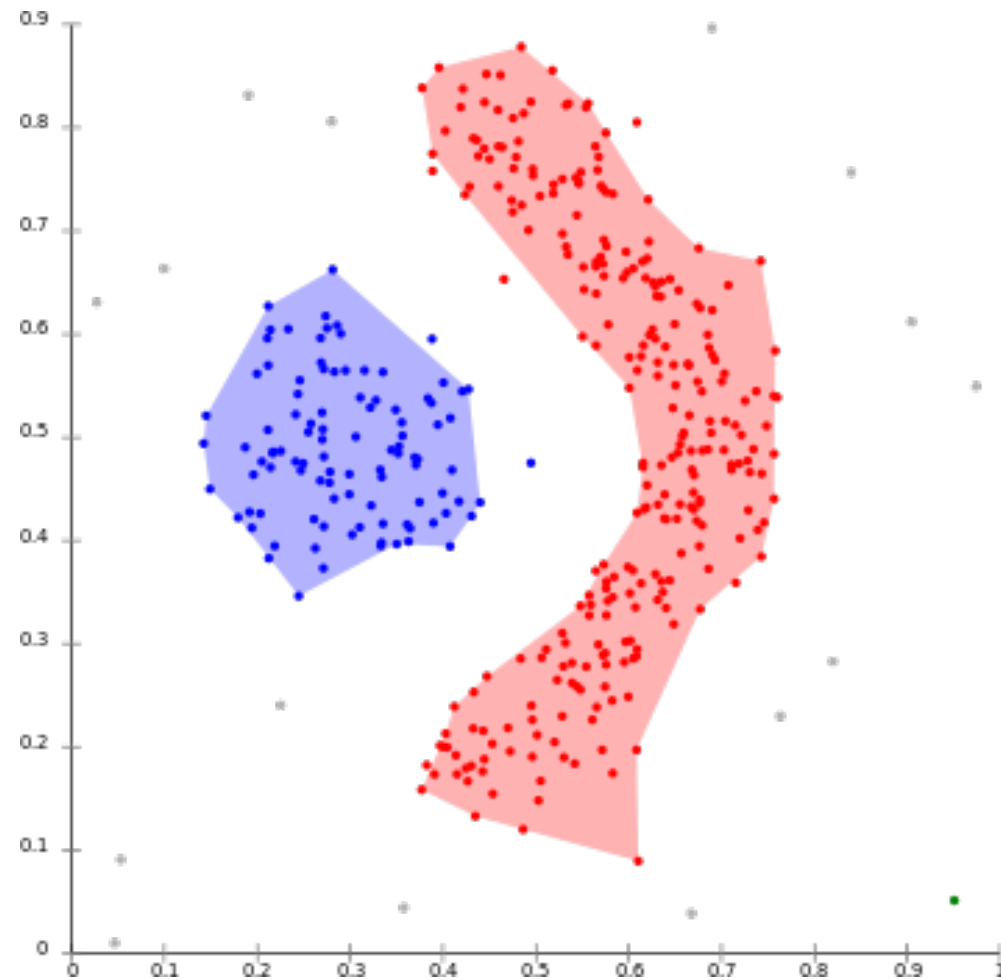
Data Mining

- Visualisierung
- Unüberwachtes Lernen
 - Clusteranalyse
- Überwachtes Lernen
 - Neuronale Netze
- Statistik
 - Statistische Tests
 - Bayesnetze

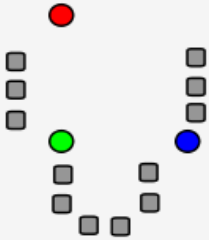
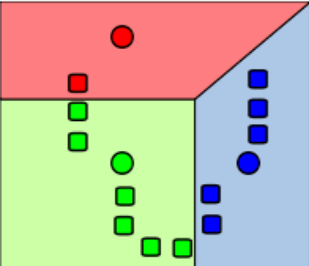
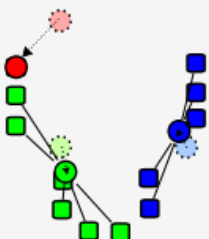
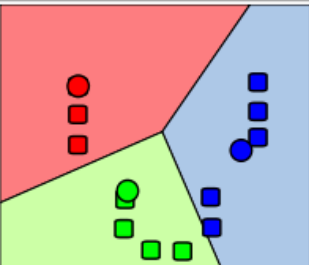
Datentypen

- Metrische Daten
 - Definiertes Abstandsmaß; z.B. Länge, Breite, Geschwindigkeit
- Ordinale Daten
 - Existierende Ordnung; z.B. T-Shirt Größen
 $S < M < L < XL \dots$
- Nominale Daten
 - Diskrete Daten ohne Ordnung; z.B. Farben

Unüberwachtes Lernen



K-Means

	<p>Drei Clusterzentren wurden zufällig gewählt.</p>
	<p>Die durch Rechtecke repräsentierten Objekte (Datenpunkte) werden jeweils dem Cluster mit dem nächsten Clusterzentrum zugeordnet.</p>
	<p>Die Zentren (jeweilige Schwerpunkte) der Cluster werden neu berechnet.</p>
	<p>Die Objekte werden neu verteilt und erneut dem Cluster zugewiesen, dessen Zentrum am nächsten ist.</p>

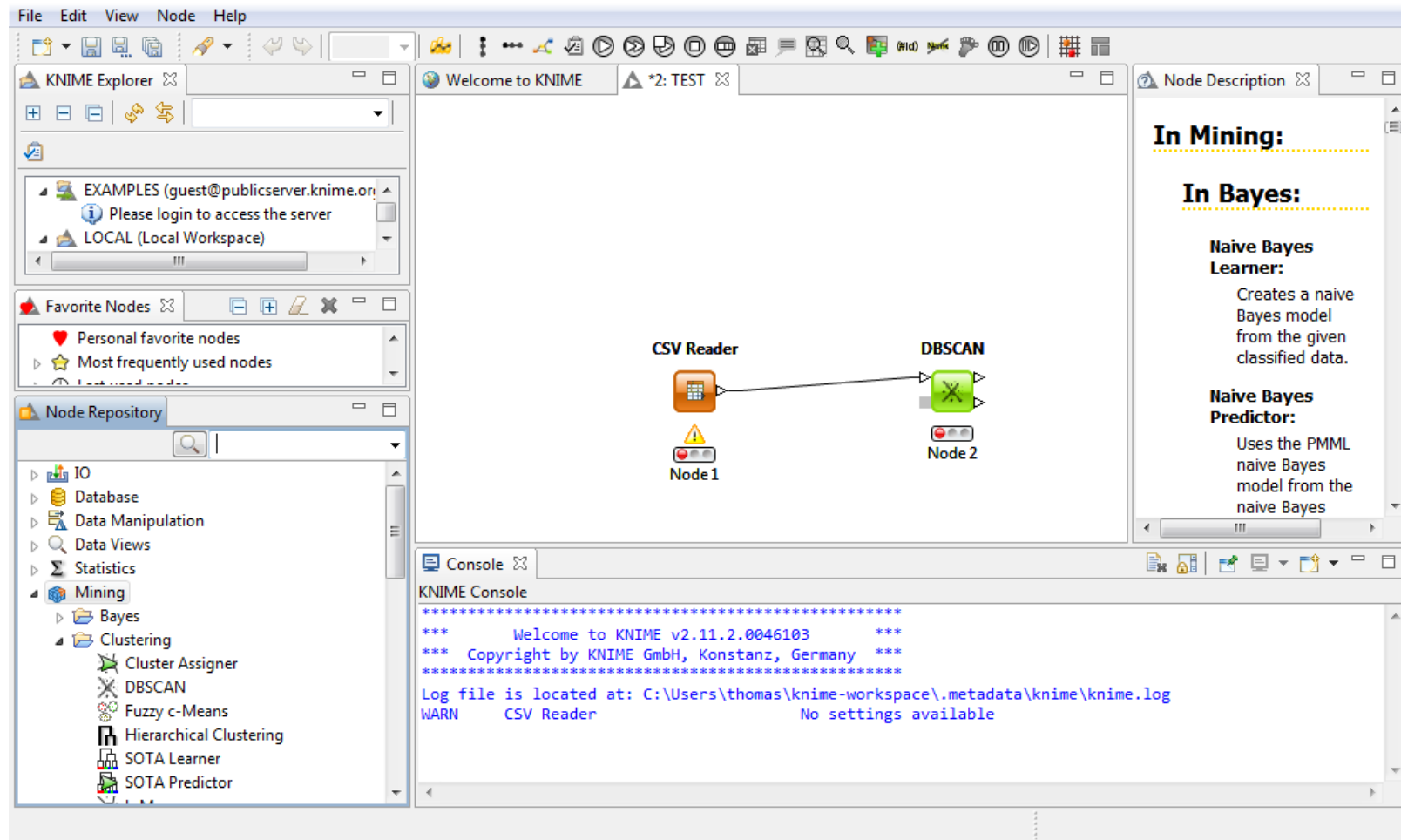
KNIME

KNIME

KNIME, der Konstanz Information Miner, ist eine [freie Software](#) für die interaktive [Datenanalyse](#). KNIME ermöglicht durch das modulare Pipelining-Konzept die Integration von zahlreichen Verfahren des [maschinellen Lernens](#) und des [Data-Mining](#). Die graphische Benutzeroberfläche ermöglicht das einfache und schnelle Aneinandersetzen von Modulen für die Datenvorverarbeitung ([ETL](#): Extraction, Transformation, Loading), der Modellierung und Analyse und der Visualisierung. KNIME ist seit etwa 2006 im Bereich der pharmazeutischen Forschung im Einsatz^[1]. KNIME wird aber auch in anderen Bereichen wie [Kundenpflege](#) (CRM), [Business Intelligence](#) und Finanzdatenanalyse eingesetzt.

Quelle: <http://de.wikipedia.org/wiki/KNIME>

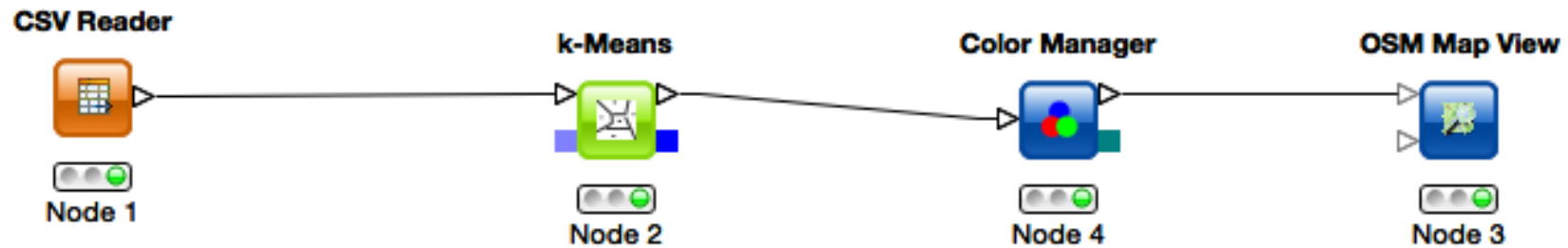
KNIME - GUI



Anbindung KNIME an Hive

- Es gibt einen JDBC Treiber für Hive
 - Einbindung diverser Klassen in den CLASSPATH von KNIME
`hive-jdbc.jar, hive-service.jar, hive-common.jar, hadoop-common.jar, libthrift-0.9.0.jar, httpclient-4.2.5.jar, httpcore-4.2.5.jar, commons-logging-1.1.3.jar, commons-codec-1.4.jar, slf4j-api-1.7.5.jar`
- Kommerziell: KNIME Big Data Extension
- Tabellen als CSV Datei herunterladen

Unüberwachtes Lernen



Row ID	D longit...	D lat
Row0	9	51
Row1	9	51
Row2	9	51
Row3	9	51
Row4	9	51
Row5	9	51
Row6	9	51
Row7	9	51
Row8	9	51
Row9	9	51
Row10	9	51
Row11	9	51
Row12	9	51
Row13	9	51
Row14	9	51
Row15	8	47

K-Means Properties

number of clusters: 10

max. number of iterations: 99

Exclude

Column(s): Search

Select

add >>

add all >>

Include

Column(s): Search

Select all search hits

D longitude

D latitude

Select one Column

Cluster

Nominal

cluster_7

cluster_2

cluster_0

cluster_1

cluster_3

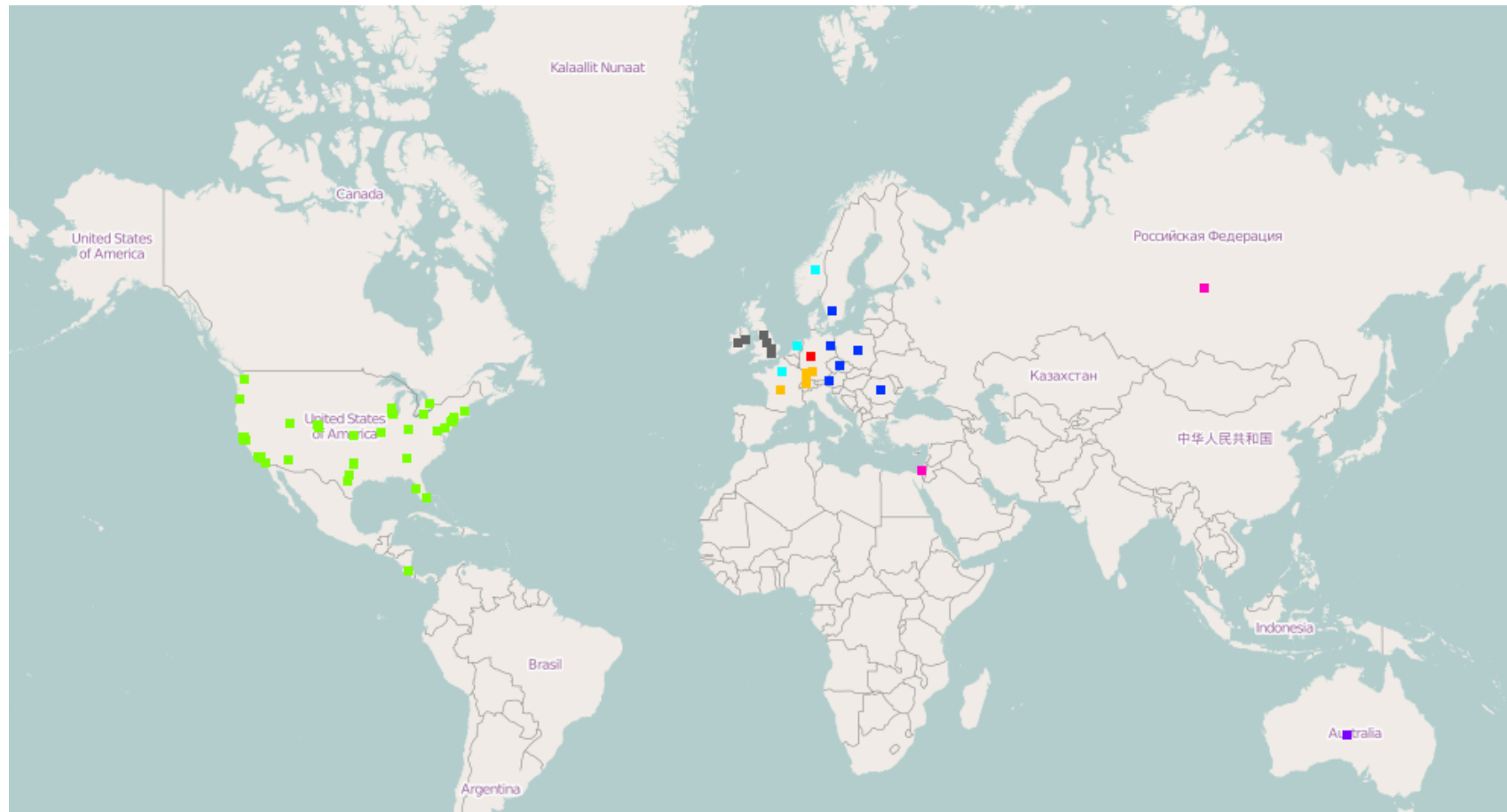
cluster_6

cluster_4

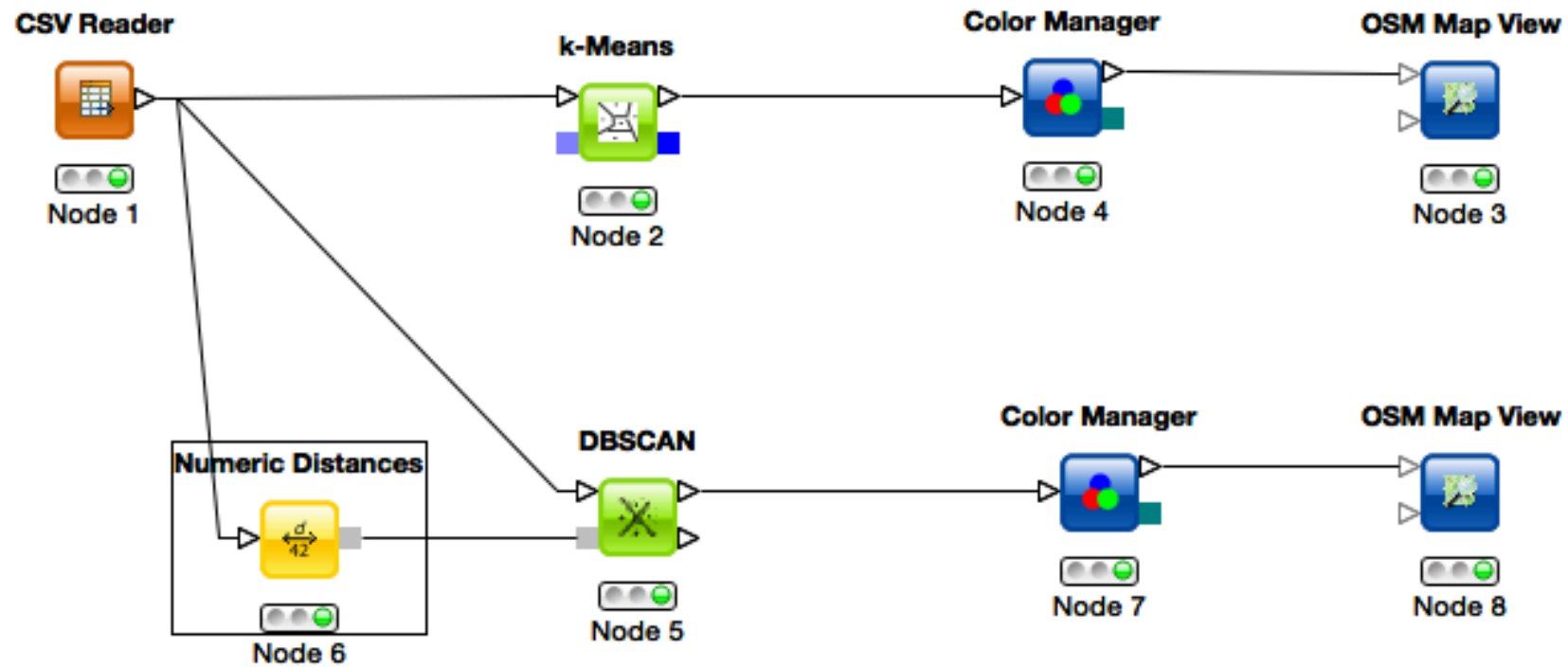
cluster_5

Preview

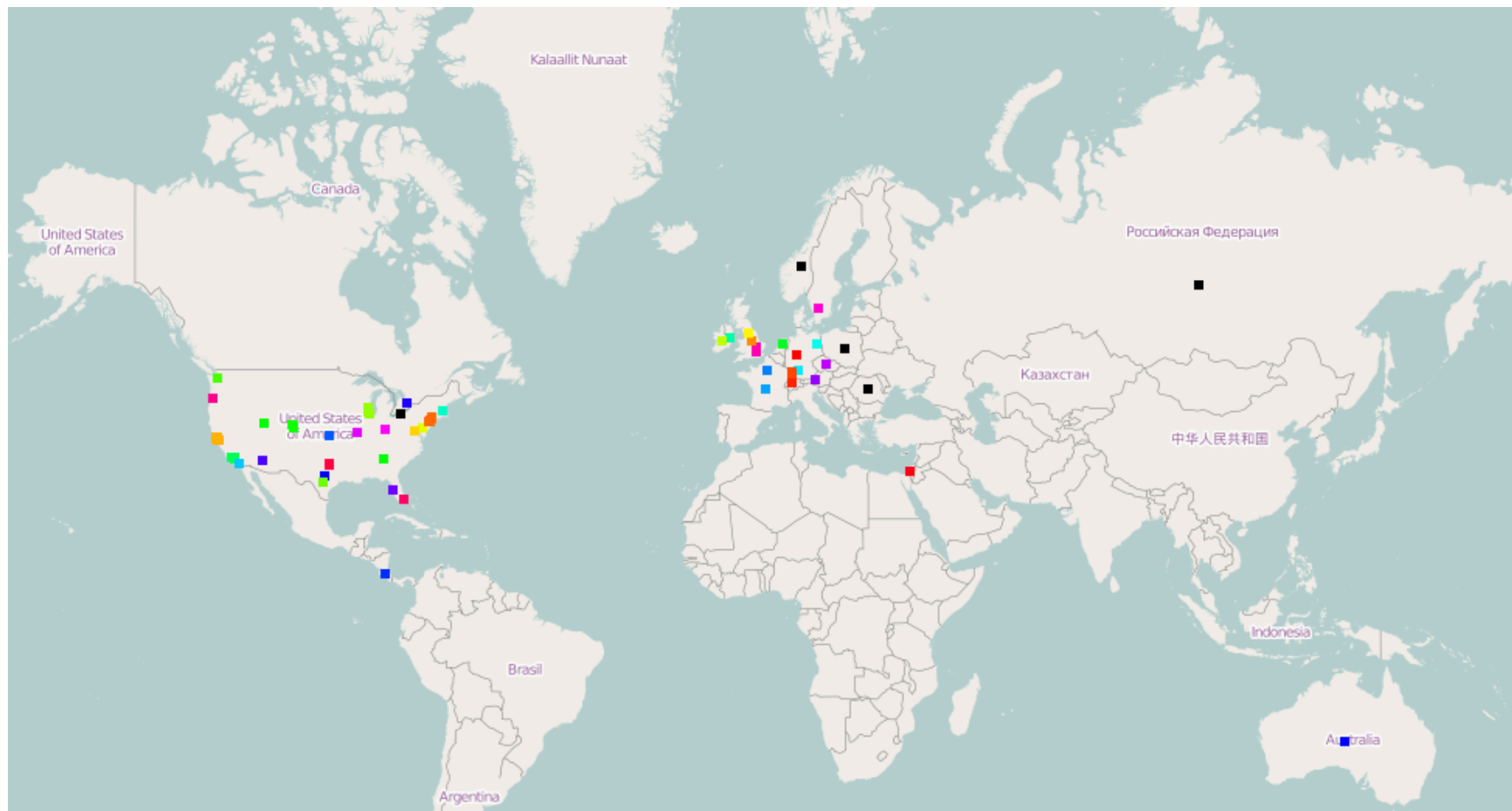
Unüberwachtes Lernen – k-Means



K-Means & DBSCAN



Unüberwachtes Lernen – DBSCAN



Überwachtes Lernen

- Histogramm über das Uploadvolumen, eine Einstufung, ob es sich um einen bestimmten Server handelt

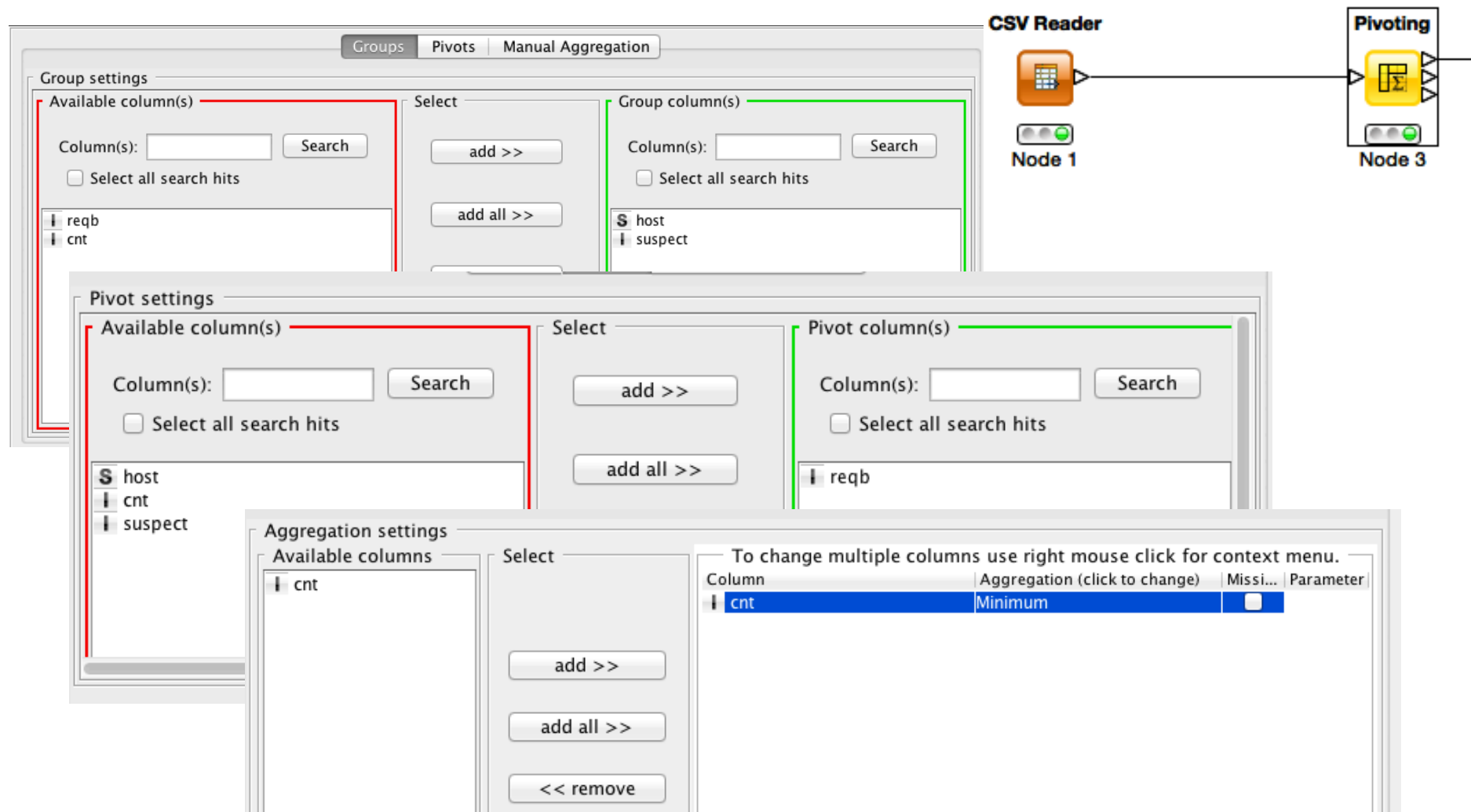
Überwachtes Lernen

```
select reqb, host,  
       count(1) as cnt, if(host='http.  
00.s.sophosx1.net', 1, 0) as suspect  
from  
  
( select INT(reqbytes/10) as  
reqb ,parse_url(url,'HOST') as host  
from access_log_2 ) t  
  
where isnotnull(host) group by reqb,  
host ;
```

Überwachtes Lernen

◆	◆ reqb	◆ host	◆ cnt	◆ suspect
0	7	softwareupdate.vmware.com	8	0
1	16	iphone-ld.apple.com	2	0
2	17	cl2.apple.com	3	0
3	17	cl4.apple.com	1	0
4	18	cl2.apple.com	25	0
5	18	dci.sophosupd.com	47	0
6	18	http.00.s.sophosxl.net	514	1
7	19	cl2.apple.com	9	0
8	19	http.00.s.sophosxl.net	1611	1
9	20	http.00.s.sophosxl.net	586	1
10	20	www.swr3.de	2	0
11	21	d1.sophosupd.com	141	0
12	21	http.00.s.sophosxl.net	511	1
13	21	pgp.uni-mainz.de	1	0
14	21	s.mzstatic.com	1	0
15	21	sd.symcb.com	2	0
16	22	crl.apple.com	2	0
17	22	d1.sophosupd.com	141	0

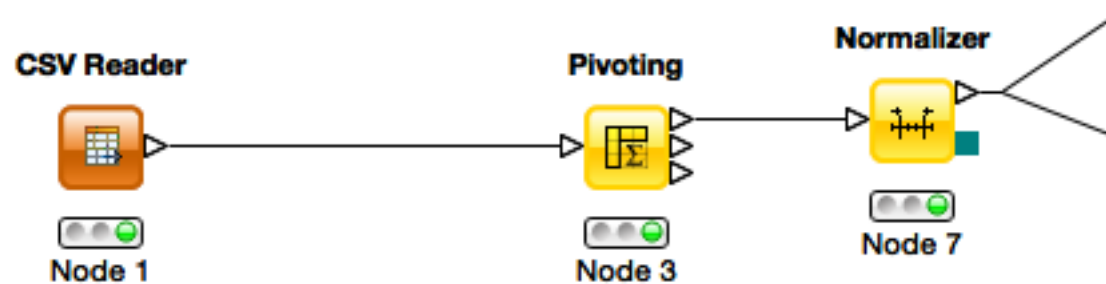
Überwachtes Lernen



Überwachtes Lernen

[illegible]

Überwachtes Lernen



☒ Manual Selection ☐ Wildcard/Regex Selection

Exclude

Column(s):

☐ Select all search hits

☒ suspect

☒ Enforce exclusion

Select

Include

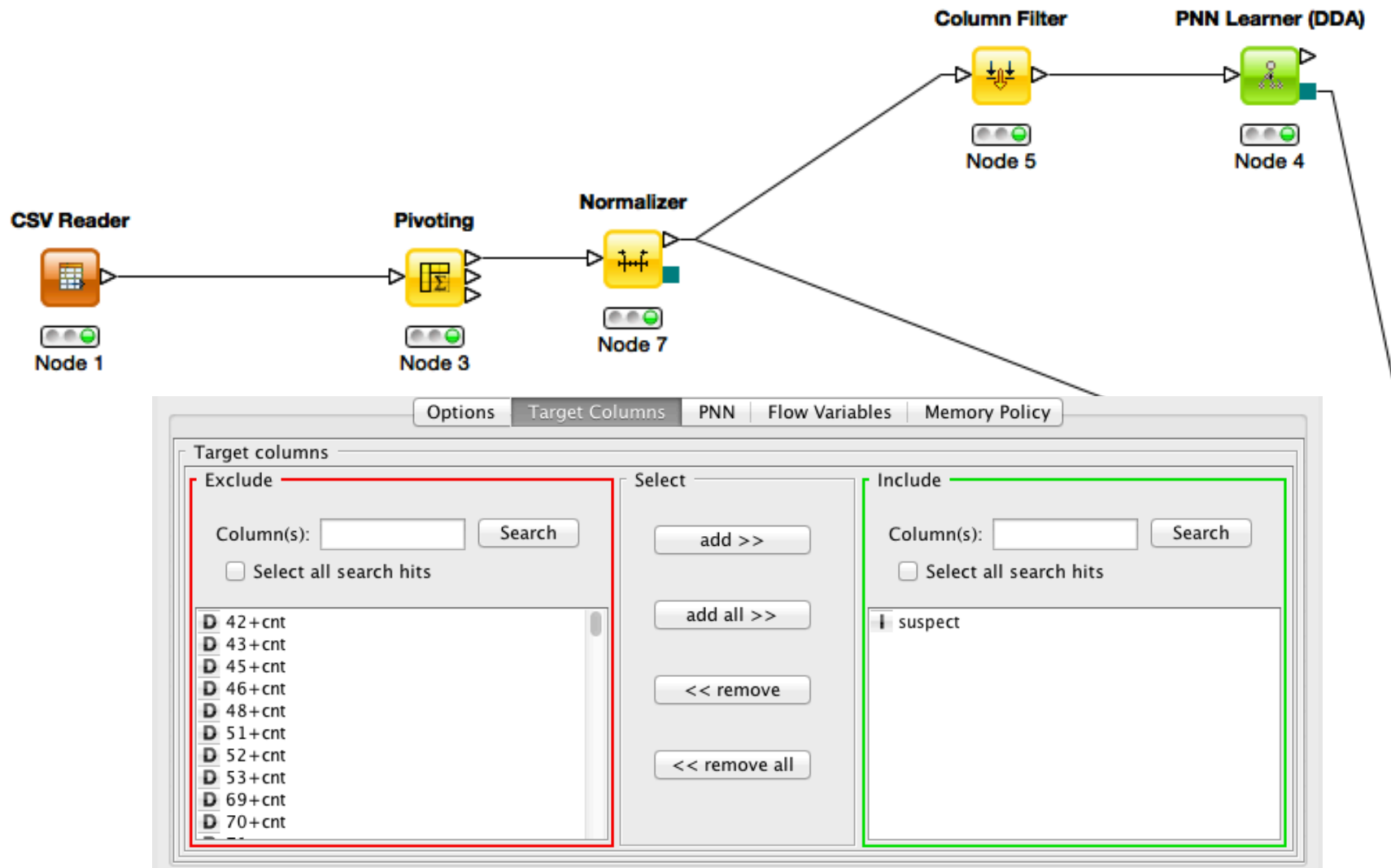
Column(s):

☐ Select all search hits

42+cnt
43+cnt
45+cnt
46+cnt
48+cnt
51+cnt
52+cnt
53+cnt

☐ Enforce inclusion

Überwachtes Lernen



Überwachtes Lernen

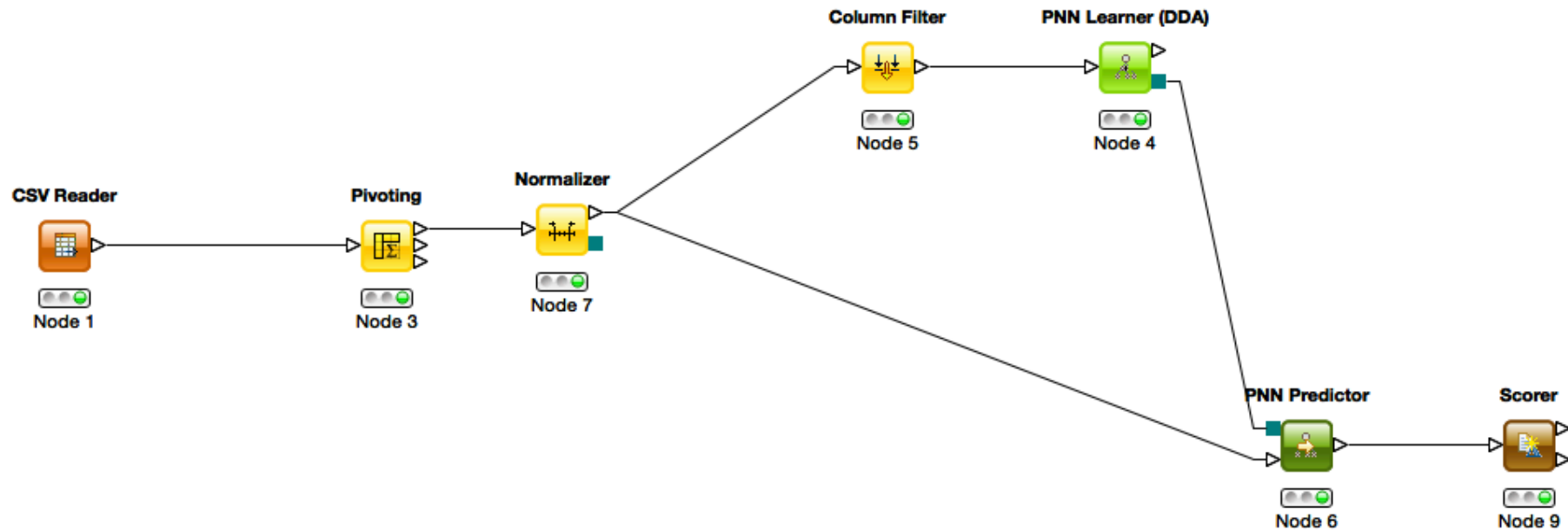


Table "spec_name" - Rows: 2		
Row ID	0	1
0	834	0
1	1	0

Überwachtes Lernen



R

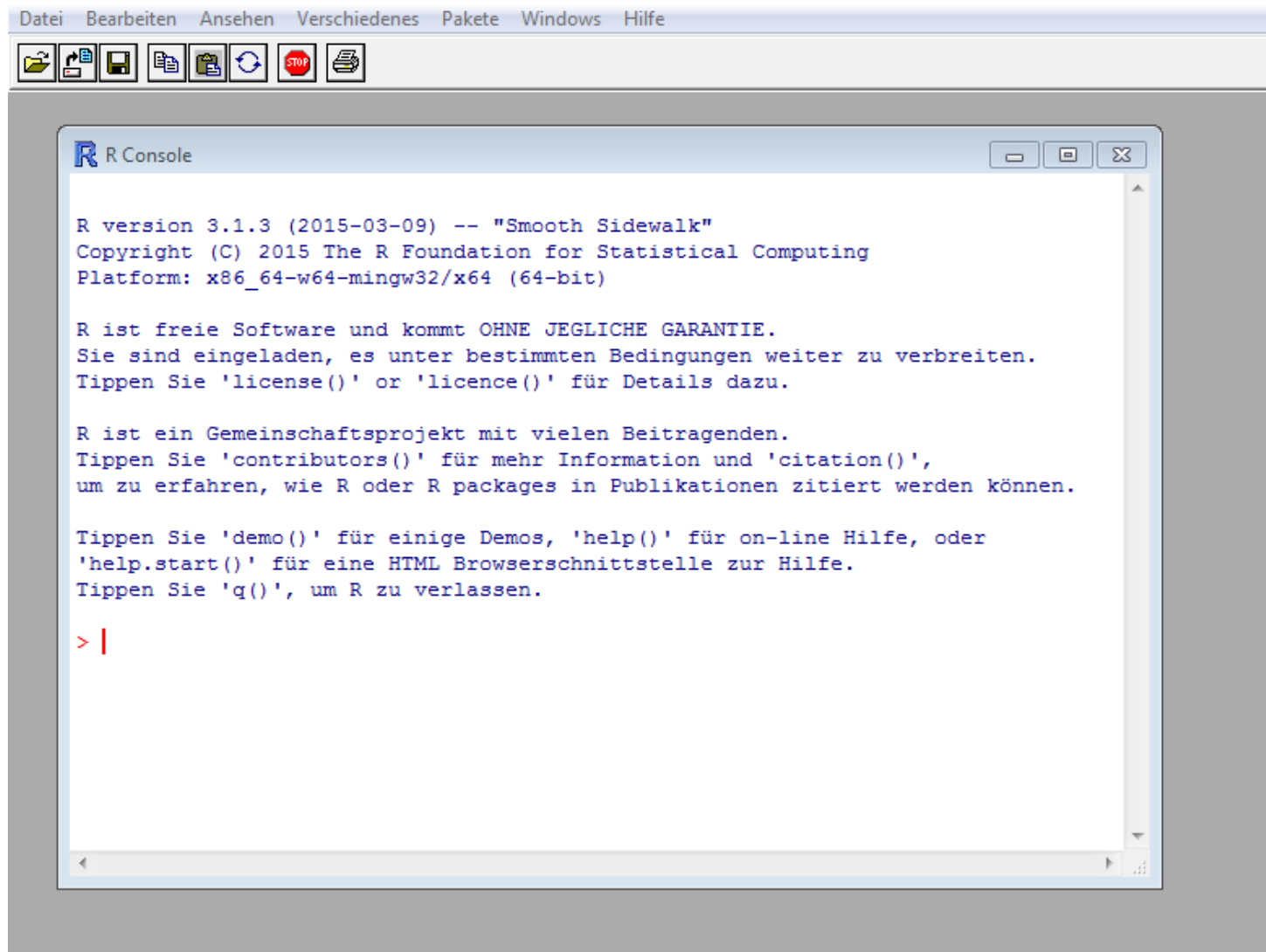
R (Programmiersprache)

R ist eine [freie Programmiersprache](#) für [statistisches Rechnen](#) und [statistische Grafiken](#). Sie ist in Anlehnung an die Programmiersprache [S](#) entstanden und weitgehend mit dieser kompatibel. Außerdem orientierten sich die Entwickler an der Programmiersprache [Scheme](#).

R ist Teil des [GNU-Projekts](#) und auf vielen [Plattformen](#) verfügbar. R gilt zunehmend als die Standardsprache für statistische Problemstellungen sowohl im kommerziellen als auch im wissenschaftlichen Bereich (obwohl vor allem im kommerziellen Bereich [SAS](#) ebenfalls sehr populär ist).^{[1][2]} Im aktuellen [TIOBE-Index](#) (Stand: Januar 2015) belegt R Platz 18.^[3] Beschäftigte mit guten R-Kenntnissen, die an der Dice Tech Salary Survey (2013) teilnahmen (insgesamt 17236 – vorwiegend US-amerikanische – Beschäftigte aus der Technologiebranche), hatten ein höheres Durchschnittseinkommen als Beschäftigte mit anderen IT-Fähigkeiten.^[4]

Quelle: [http://de.wikipedia.org/wiki/R_\(Programmiersprache\)](http://de.wikipedia.org/wiki/R_(Programmiersprache))

R - GUI



Unterschied KNIME / R

Funktionalität	KNIME (OSS)	R
Skripte	✗	✓
Berechnungen im Cluster	✗	✓
Frontend	✓	✓
Usability	++	0

RHive

```
hsum <- function(prev, sal) {  
  if (is.null(prev)) sal else prev + sal  
}  
hsum.partial <- function(agg_sal) {  
  agg_sal  
}  
hsum.merge <- function(prev, agg_sal) {  
  if (is.null(prev)) agg_sal else prev + agg_sal  
}  
hsum.terminate <- function(agg_sal) {  
  agg_sal  
}
```

RHive

Funktion exportieren:

```
rhive.assign('hsum', hsum)
rhive.assign('hsum.partial',
             hsum.partial)
rhive.assign('hsum.merge', hsum.merge)
rhive.assign('hsum.terminate',
             hsum.terminate)
rhive.exportAll('hsum')
```

Funktion nutzen:

```
rhive.query("select RA('hsum',sal) from
emp group by empno")
```

DEMO

Analyse einer Squid access.log

VIELEN DANK