Text analysis 1

Lukáš Lehotský

“text analysis is just a fancy and convoluted way how to obtain independent or dependent variable”

Inaki Sagarzazu

Concepts

Bag of words

Bag of words

- The quick brown fox jumps over the lazy dog

Word	Occurrence
brown	1
dog	1
fox	1
jumps	1
lazy	1
over	1
quick	1
the	2

Co-occurrence

Co-occurrence

- The quick **brown fox** jumps over the lazy dog. **Brown dog** sleeps well.

Word	Sentence 1	Sentence 2
brown	1	1
dog	1	1
fox	1	
jumps	1	
lazy	1	
over	1	
quick	1	
sleeps		1
the	2	
well		1

Co-locations and N-grams

Co-locations/n-grams

- Established phrases – usually occur together and form a meaning

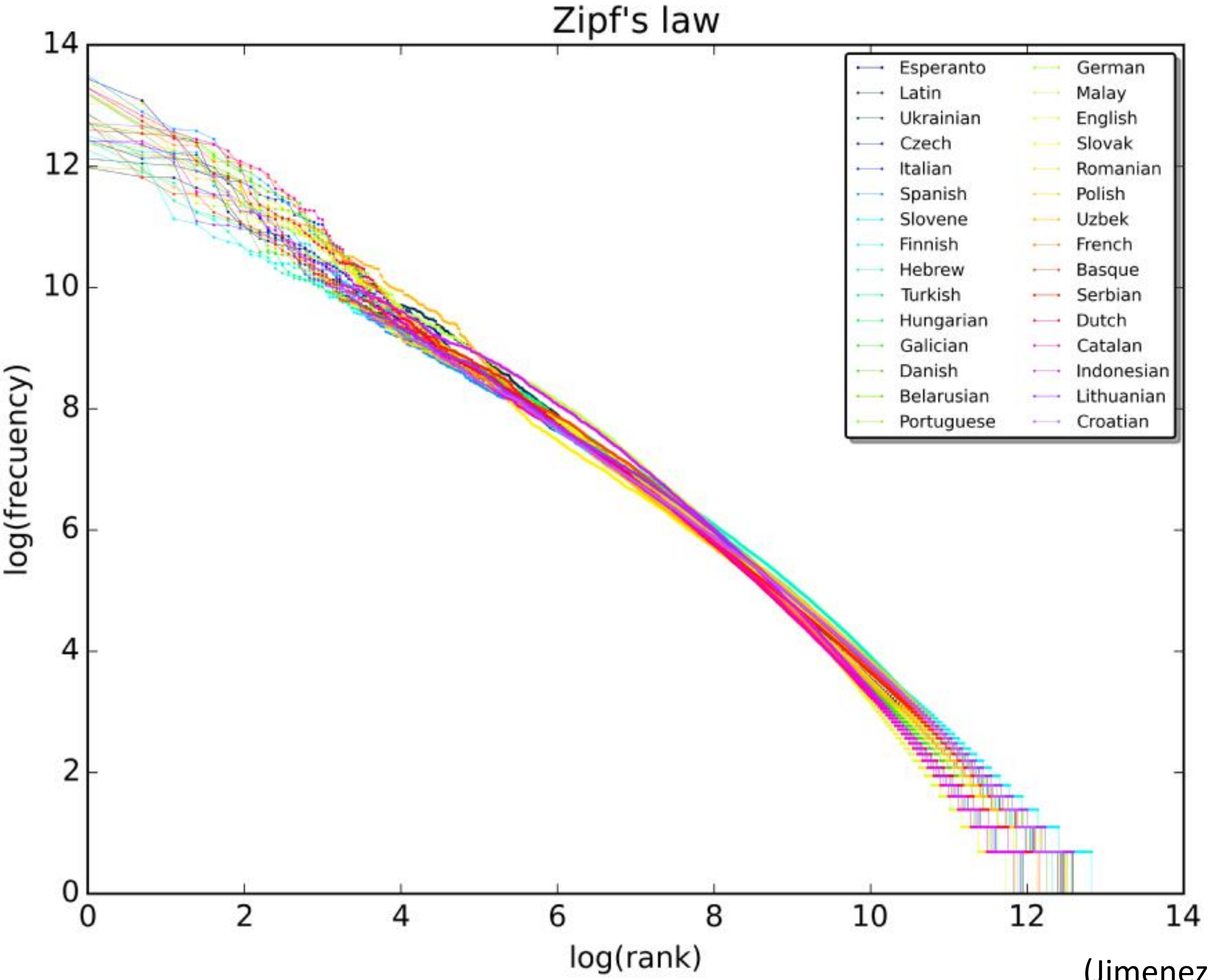
Ministry of the Environment

European Union

prime minister

toilet paper

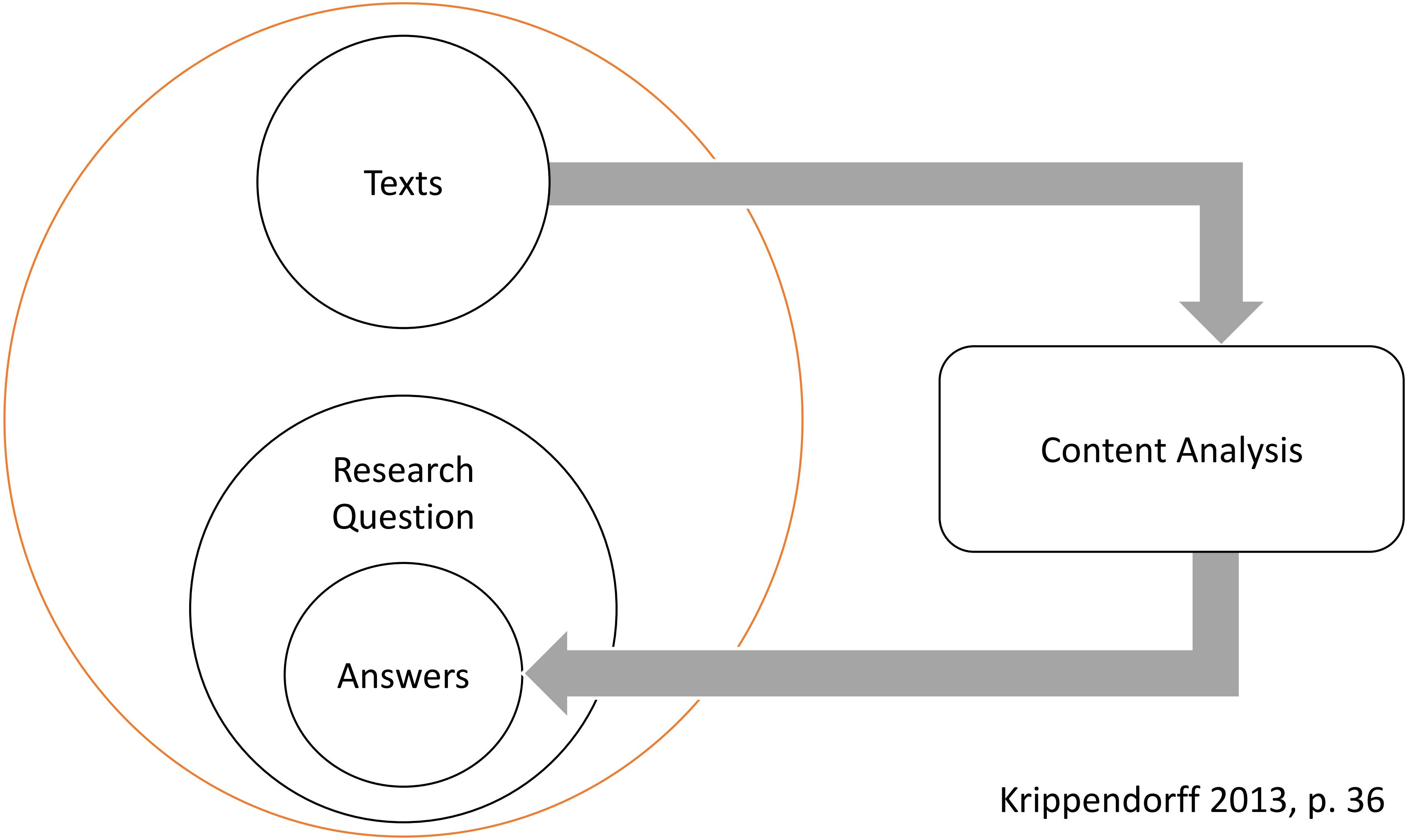
Zipf law



(Jimenez, 2015)

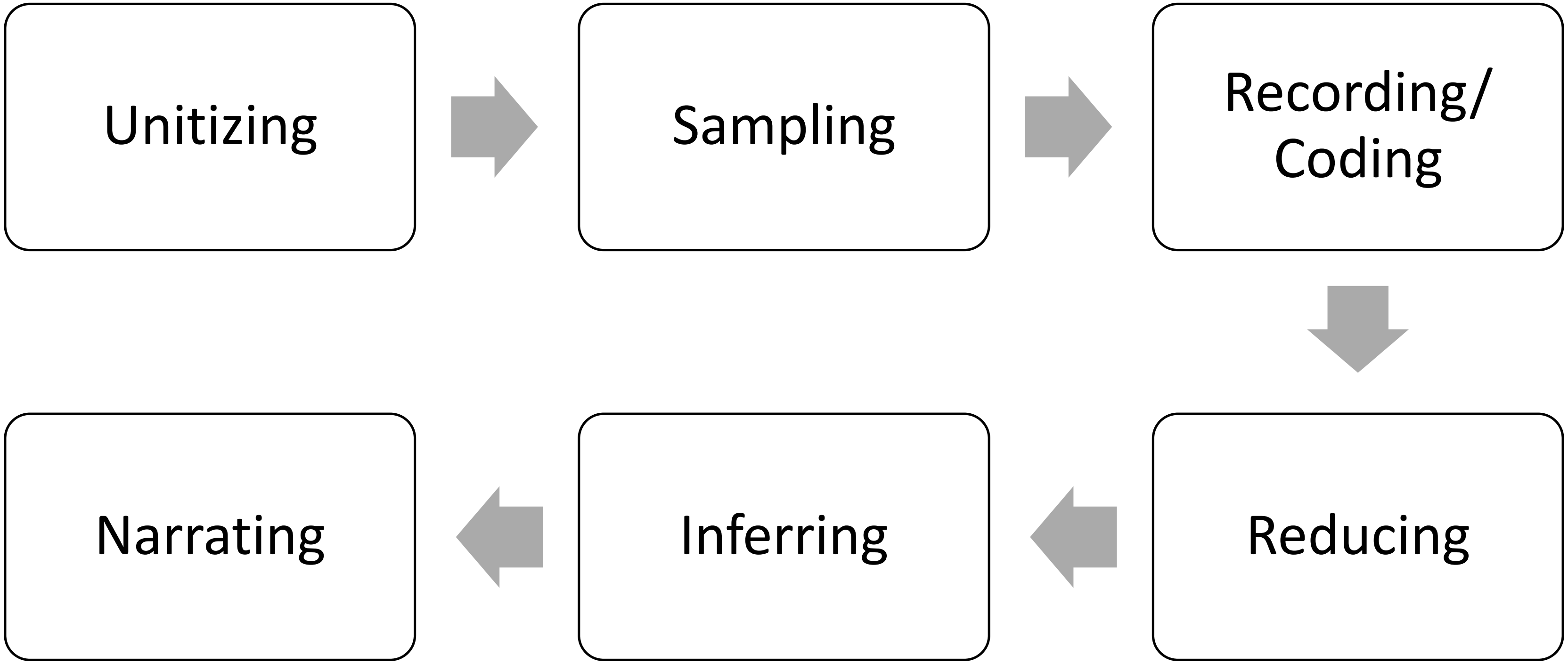
Manifest vs. latent content

Design of CA research



Krippendorff 2013, p. 36

Design of CA research


Unitizing

Sampling

Recording/
Coding

Reducing

Inferring

Narrating

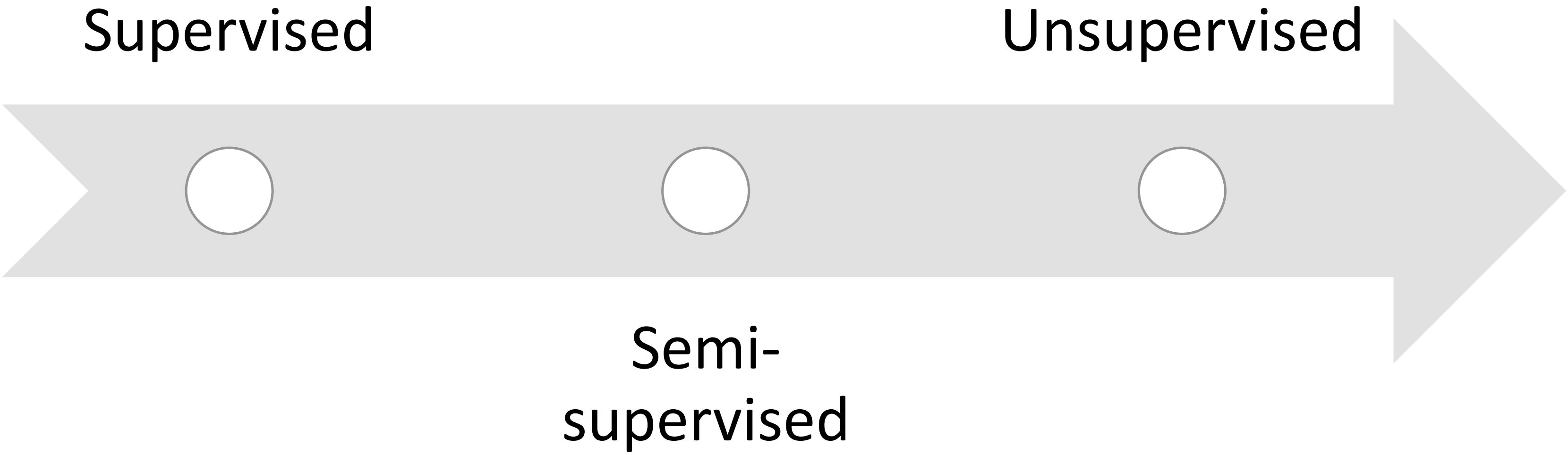
Krippendorff 2013, p. 86

Basic terminology

- Corpus
  - Body of all text pieces available for the content analysis
- Term
  - Text token, usually word
- Term-document matrix
  - Matrix which records occurrence of terms in documents

Methods

Methods of TA



Supervised

Unsupervised

Semi-supervised

Methods of TA

- Supervised methods
    - Manual coding
- Semi-supervised
    - Dictionary-based methods
        - Deductively given dictionary
        - Dictionary obtained from data
            - Automatically
            - Manually
- Unsupervised
    - Frequencies
    - Topic modeling
    - …

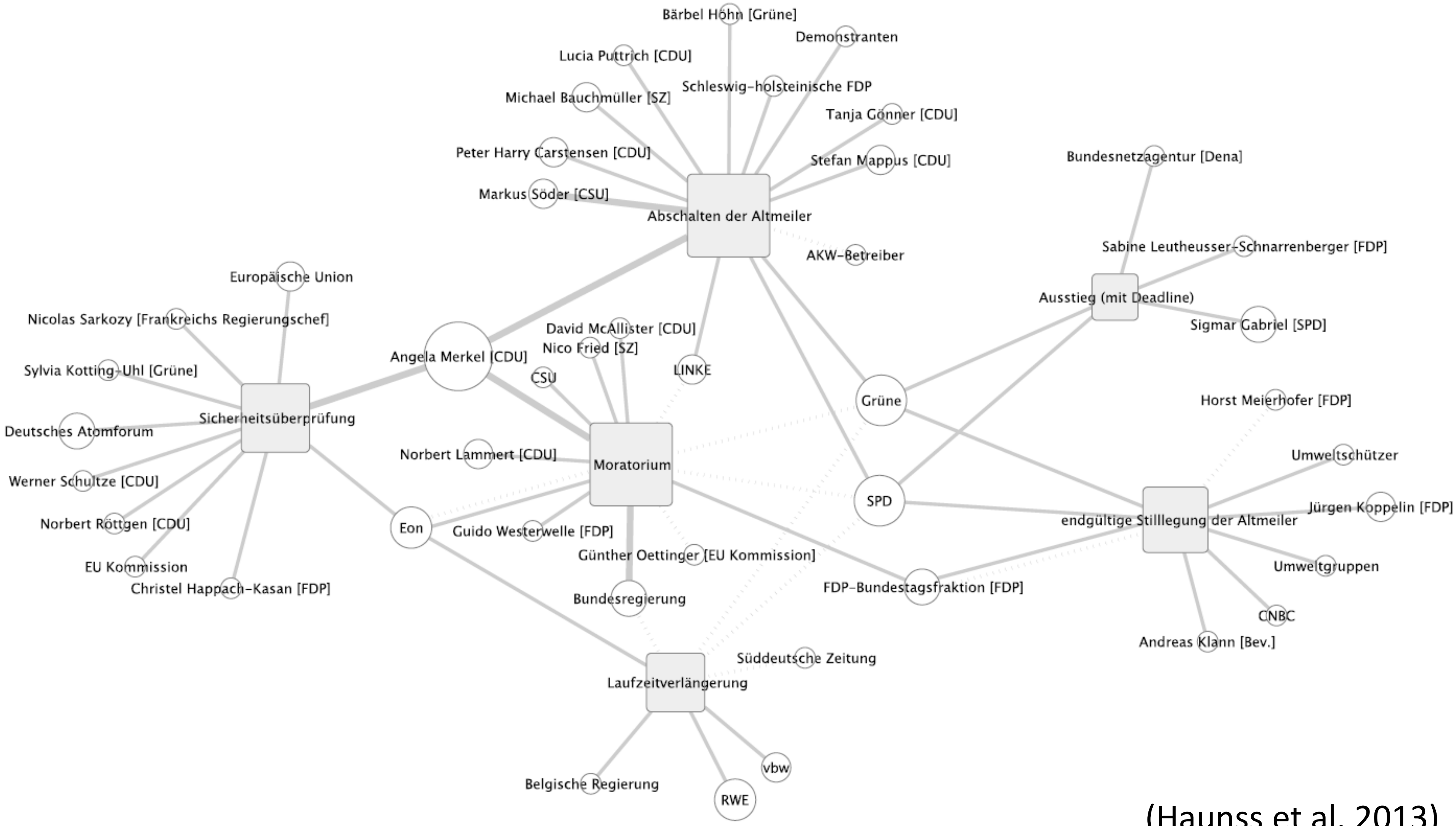
Fully supervised – manual coding

- Manual coding of text units
- Inductive vs. deductive coding
  - Inductive – data-driven
    - Categories not known
    - Open coding – categories emerge in iterative text reading
    - Axial coding – abstraction from open coding into categories
  - Deductive – theory-driven
    - Categories known a-priori
    - Existing code-book applied over data

Fully supervised

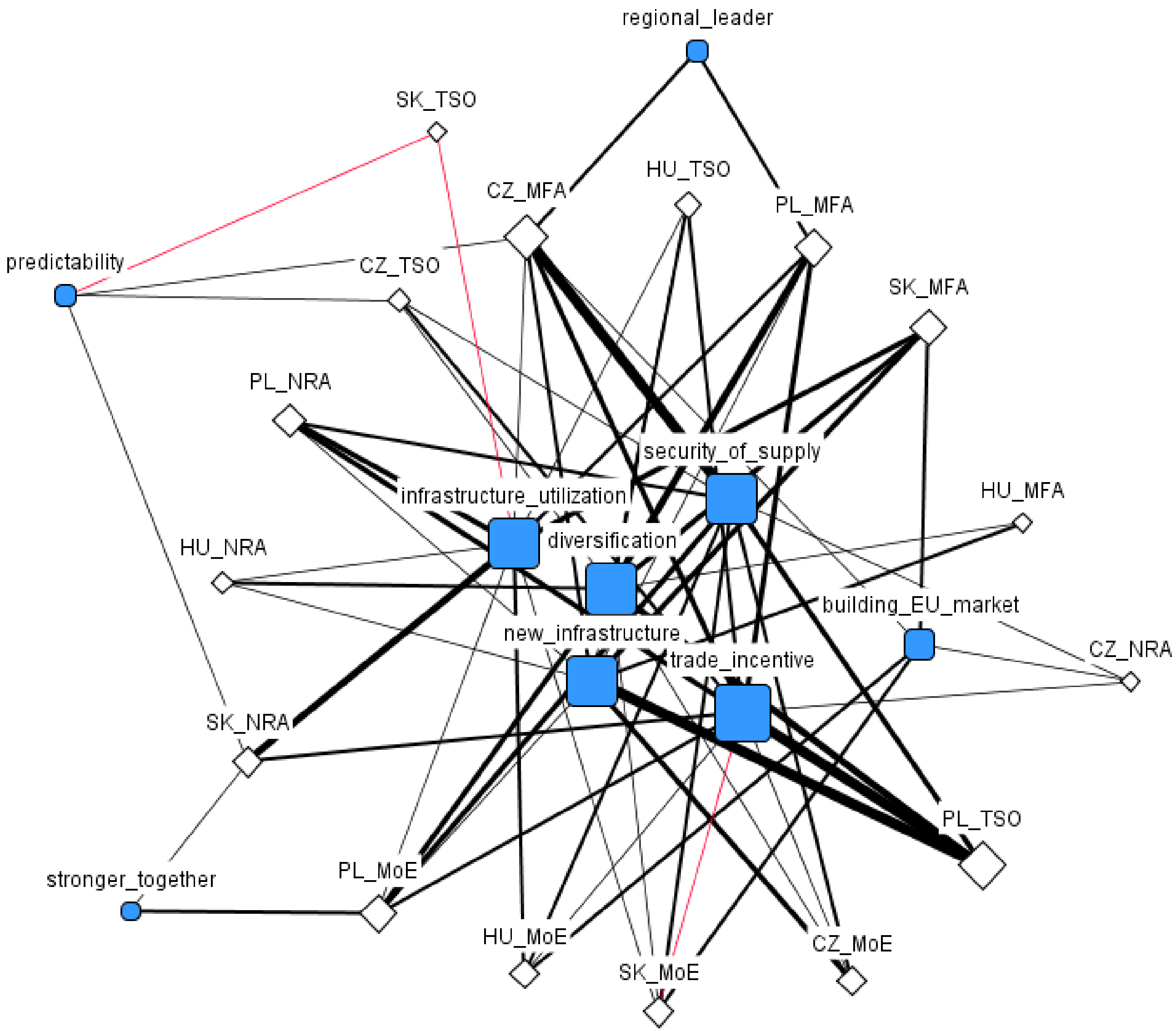
- Coding is input for further analysis
  - Frequencies of codes
  - Temporal development
  - Standard statistical methods
  - Socio-semantic networks
- Discourse network analysis
  - Socio-semantic networks of actors and meanings (codes) they use

Fully supervised - DNA



(Haunss et al. 2013)

Fully supervised - DNA



Issues with manual coding

- Questions of validity and reliability
- Reliability of human coders needs to be measured and accounted for
    - Intra-coder reliability (variation by same coder)
    - Inter-coded reliability (variation by different coder)
- Ways how to measure, e.g. Krippendorff α
- Ways how to account for
    - Only overlap
    - Resolution of differences

Semi-supervised

- Dictionary-based automated coding
    - Words in dictionary are discovered across the corpus
    - Coding process is done automatically
- Construction of dictionaries
    - Given pre-defined dictionary
        - WordStat, LIWC, …
    - Constructed from data
        - Theoretically-informed
        - Automatically generated
            - WordFish
            - WordScores

Semi-supervised

- Existing dictionaries
    - WordStat (Laver & Garry 2000)
        - Estimation of policy positions from political texts
        - 415 words, 19 categories
    - LIWC (Linguistic Inquiry Word Count)
    - Sentiment dictionaries
    - General Inquirer
- Logic of this approach is to crawl over texts, discover tokens in dictionary and score texts
    - Scoring whole corpus
    - Scoring individual texts

Semi-supervised

- Dictionary from data
  - Sample of texts with known properties
  - Other texts related – e.g. legal/conceptual documents
- Dictionaries built by researchers
  - Long process
  - High validity – researchers know texts
  - Lower reliability – same reasons as manual coding

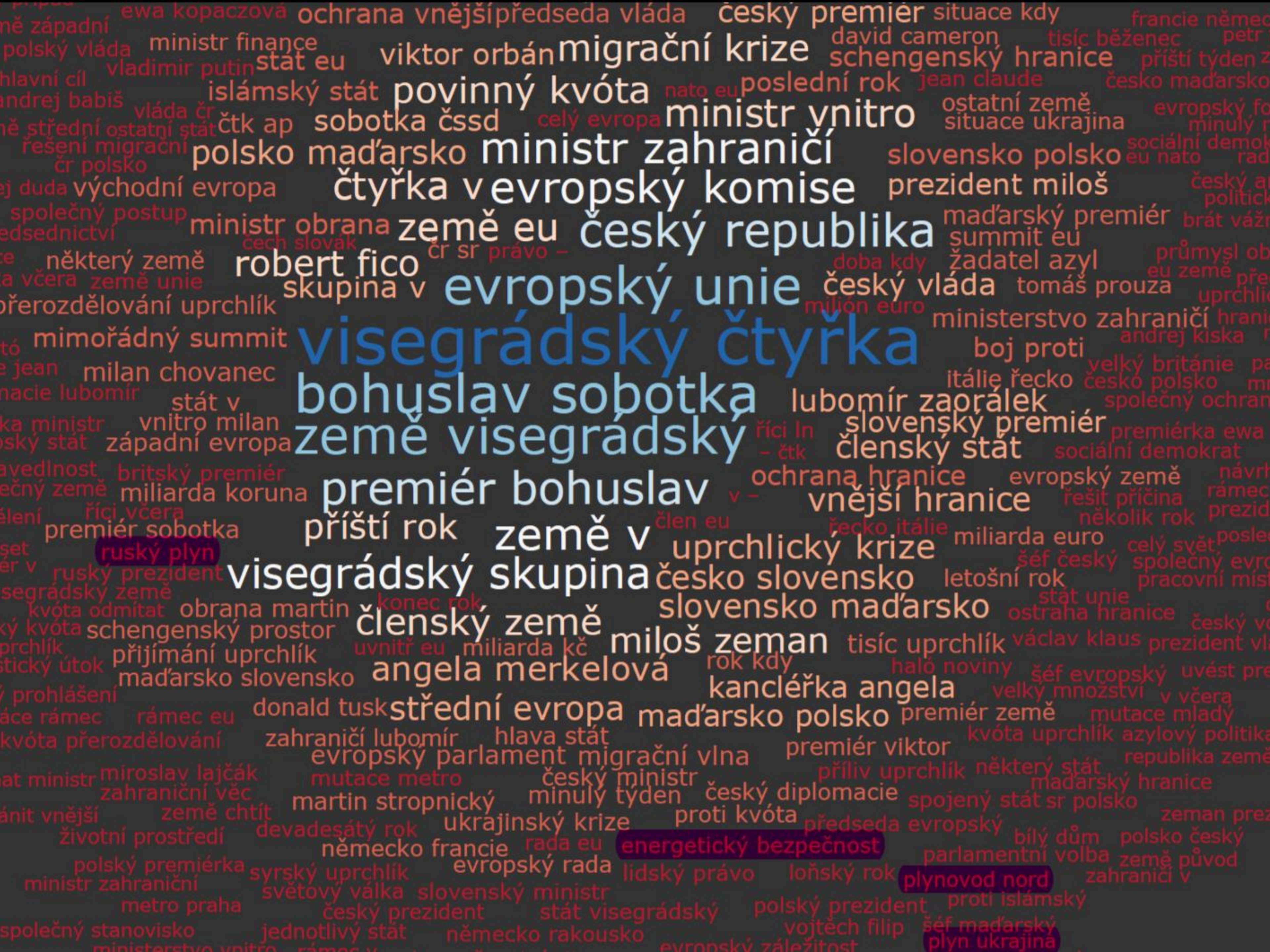
Semi-supervised

- Automated dictionary constructing
  - Laver, Benoit and Garry 2003
  - Two populations of texts
    - Texts with known properties – training set
    - Texts with unknown properties – target set
  - Logic of the process
    - Assign values of the category to known texts (training sample)
    - Let computer find words in the training sample and assign individual scores to words from texts
    - Code unknown texts with existing dictionary
  - High reliability, but questionable validity

Unsupervised

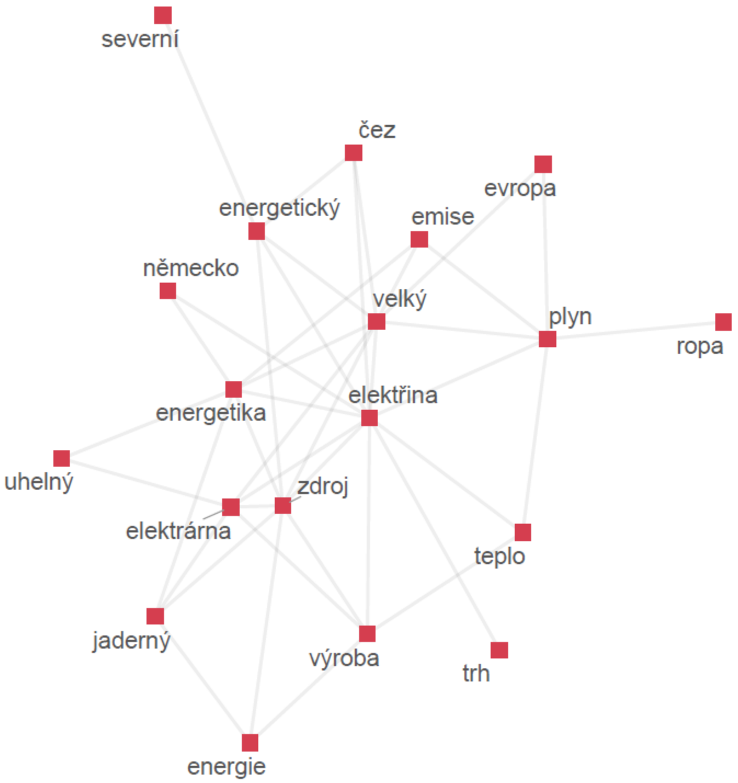
- Most naïve – word frequencies
  - Just a crude exploratory hint of what is in text
- Clustering and multidimensional scaling of words
  - Based on co-occurrence of words
- Unsupervised categorization on term-document matrix
  - Topic modeling
- Co-occurrence term networks

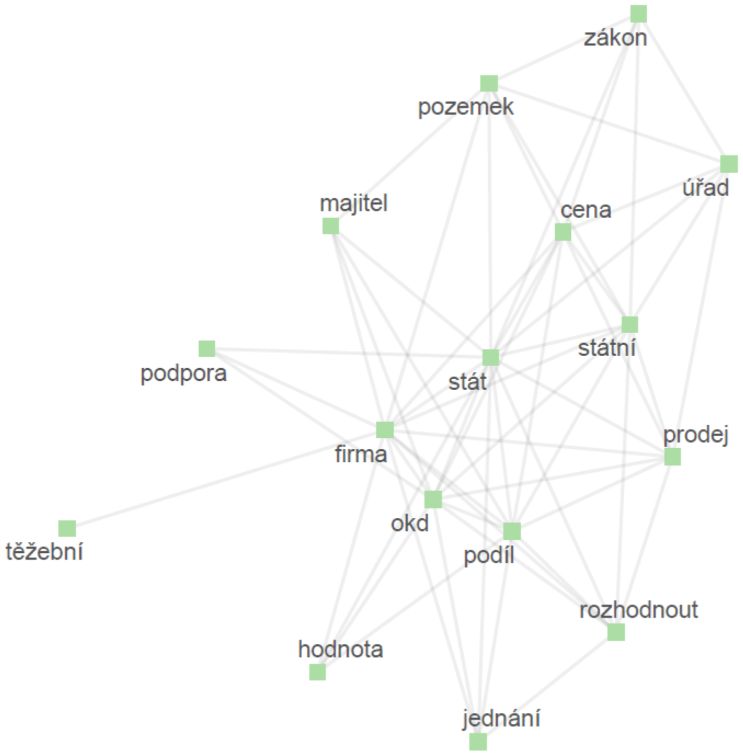
země
český
rok
stát
premiér
evropský
evropa
vláda
uprchlík
hranice
maďarsko
visegrádský
kvóta
společný
sobotka
německo
člověk
polsko
česko
včera
chtít
ukrajina
slovensko
ministr
říci
unie
muset
prezident
velký
energetický
plyn
rusko
zeman
právo
konference
přijmout
obrana
případ
poslední
zástupce
poslanec
prostředek
itálie
ukázat
včetně
systém
parlament
vést
potřeba
festival
existovat
pomáhat
martin
blízký
konec
schengen
fungovat
jediný
znamenat
vlna
úřad
národní
důležitý
solidarita
viktor
bezpečnost
práce
příští
stále
orbán
takzvaný
mimo
dítě
dohodnout
diskuse
kdyby
rozhodnout
mezinárodní
souvislost
hovořit
fórum
debata
unijní
něco
uprchlický
vztah
myslit
volba
řada
především
často
významný
nejen
zákon
platit
brzy
trh
některý
czech
občan
cesta
zahraniční
prostor
návštěva
vůči
kvůli
sám
jednat
vysoký
rozdíl
brát
čas
autor
dohoda
skupina
postoj
téma
shodnout
kontrola
zaorálek
pomoc
patřit
vojenský
visegrád
republika
summit
maďarský
současný
postup
hlava
čekat
procento
čssd
návrh
silný
politik
ministerstvo
divadlo
spíše
totiž
nějaký
řešení
vědět
politika
ekonomický
přístup
klíčový
milión
řešit
ruský
tlak
jít
místo
zahraničí
cíl
členský
výsledek
ovšem
bývalý
život
chovanec
miliarda
čtyřka
fond
krize
přitom
útok
peníze
lze
zájem
prý
slovenský
oba
usa
doba
schengenský
důvod
pan
tam
migrační
názor
pozice
tady
merkelová
spolupráce
věc
takový
říkat
čech
západní
navíc
kolega
možnost
schůzka
informace
kancléřka
ukrajinský
situace
hlavní
region
začít
síla
moc
například
jiří
zahrada
připravit
dostat
proti
udělat
andrej
služba
čtk
šéf
otázka
podpora
teď
putin
program
praha
dělat
město
americký
velmi
student
jednotlivý
stejně
euro
vidět
dojít
organizace
jednání
uvést
dodat
svět
slovák
zároveň
oblast
datetimestamp
dobře
nato
povinný
zatím
jasný
cena
zůstat
angela
východní
problém
běženec
zdůraznit
německý
akce
řecko
opatření
daleko
snažit
krok
tisíc
dát
politický
vnitro
možný
projekt
západ
policista
odmítat
rámec
polský
migrant
společnost
většina
malý
afrika
podpořit
měsíc
tiskový
nyní
turecko
část
předseda
fico
prohlásit
základ
británie
miloš
druhý
bohuslav
komise
celý
mluvit
změna
naopak
rakouský
jan
několik
kdy
člen
dobrý
brusel
vnější
válka
francie
veřejný
minulý
zejména
žádný
ochrana
jaký
rád
rakousko
podporovat
demokracie
kyjev
pozice
odmítnout
sociální
přijít
zdroj
málo
armáda
nikdo
centrum
mluvčí
letošní
boj
rada
právě
týden
dále
robert



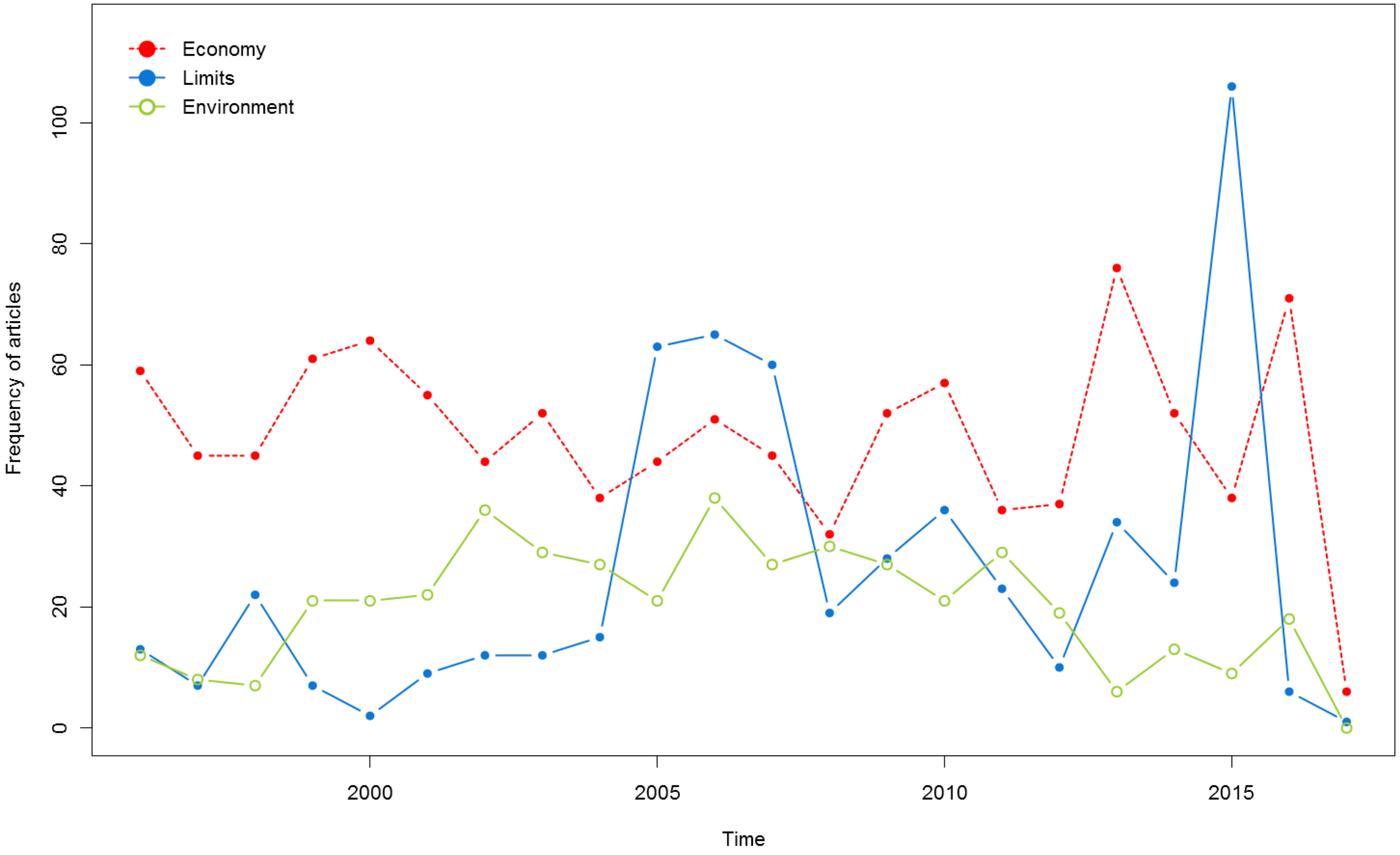
visegrádský čtyřka
evropský unie
bohuslav sobotka
země visegrádský
český republika
evropský komise
ministr zahraničí
premiér bohuslav
visegrádský skupina
ministr vnitro
povinný kvóta
migrační krize
čtyřka v
země eu
robert fico
skupina v
země v
příští rok
členský země
miloš zeman
angela merkelová
střední evropa
polsko maďarsko
sobotka čssd
islámský stát
viktor orbán
stát eu
ministr finance
ewa kopaczová
ochrana vnější
předseda vláda
český premiér
situace kdy
david cameron
schengenský hranice
poslední rok
jean claude
ostatní země
situace ukrajina
slovensko polsko
prezident miloš
maďarský premiér
summit eu
žadatel azyl
český vláda
tomáš prouza
ministerstvo zahraničí
boj proti
itálie řecko
lubomír zaorálek
slovenský premiér
členský stát
ochrana hranice
vnější hranice
uprchlický krize
česko slovensko
slovensko maďarsko
tisíc uprchlík
kancléřka angela
maďarsko polsko
premiér viktor
český diplomacie
proti kvóta
evropský parlament
migrační vlna
český ministr
minulý týden
martin stropnický
ukrajinský krize
německo francie
evropský rada
donald tusk
maďarsko slovensko
přijímání uprchlík
schengenský prostor
obrana martin
milan chovanec
mimořádný summit
přerozdělování uprchlík
východní evropa
ministr obrana
západní evropa
vnitro milan
stát v
miliarda koruna
premiér sobotka
ruský plyn
ruský prezident
některý země
země unie
společný postup
čtk ap
vláda čr
vladimir putin
hlavní cíl
andrej babiš
polský vláda
řešení migrační
čr polsko
ostatní stát
celý evropa
nato eu
čech slovák
čr sr
právo
doba kdy
milión euro
andrej kiska
velký británie
česko polsko
společný ochrana
premiérka ewa
sociální demokrat
evropský země
řešit příčina
několik rok
miliarda euro
šéf český
celý svět
společný evropský
pracovní místo
letošní rok
stát unie
ostraha hranice
václav klaus
šéf evropský
halo noviny
velký množství
premiér země
mutace mladý
kvóta uprchlík
azylový politika
republika země
některý stát
maďarský hranice
spojený stát
sr polsko
příliv uprchlík
předseda evropský
bílý dům
polsko český
parlamentní volba
země původ
zahraničí v
plynovod nord
proti islámský
polský prezident
vojtěch filip
šéf maďarský
plyn ukrajina
energetický bezpečnost
lidský právo
loňský rok
rada eu
rok kdy
uvnitř eu
miliarda kč
konec rok
člen eu
řecko itálie
říci ln
čtk
hlava stát
zahraničí lubomír
mutace metro
miroslav lajčák
zahraniční věc
země chtít
devadesátý rok
životní prostředí
polský premiérka
ministr zahraniční
syrský uprchlík
světový válka
slovenský ministr
metro praha
český prezident
stát visegrádský
společný stanovisko
jednotlivý stát
německo rakousko
evropský záležitost
rámec eu
kvóta přerozdělování
britský premiér
říci včera
kvóta odmítat
visegrádský země
francie německo
tisíc běženec
příští týden
česko maďarsko
evropský fond
minulý rok
sociální demokrat
eu nato
český
politický
brát vážně
průmysl
eu země
uprchlík
hranice
v včera
zeman prezident
prezident
uvést

Unsupervised – co-occurrence net





Unsupervised – topic modeling



How to get TDM?

Data pre-processing

- Any text analysis must be preceded by data pre-processing
  - Dropping sparse terms – has a word which occurs in 1.5M corpus once, any value?
  - Dropping most frequent terms – does most profound word of interest any informative value?
  - Dropping “stopwords” – a, the, …
  - Dropping numerals, punctuation, …
  - Dropping time and place information
  - …
- No general rules on how to do that – rule of thumb

Data pre-processing

- Stemming/lemmatization
    - Disposal of grammatical features of text
        - Dictionary-based
        - Rules-based
    - Both introduce some error into the corpus
- Lemmatization
    - Identification of lemmas (lexemes) of the words - transformation to lemmas
- Stemming
    - Stripping the word of prefixes or suffixes, leaving only word stems

Lemmatization and stemming

“This was the most tranquil presidential address.
President’s approach was very relaxed.”

- Lemmatization

“This be the most tranquil presidential address.
President approach be very relax.”

- Stemming

“This be the most tranquil presidenti address.

Presid approach be veri relax.”

Corpus generation

- Decision on document unitizing
- Decision over sampling
  - Does 5M texts provide more information than 15k?
  - Random vs. non-random sampling
- Inclusion of metadata – allow for filtering later
  - Author
  - Time and date
  - Source (e.g. media/newspaper)
  - …

Term-document matrix

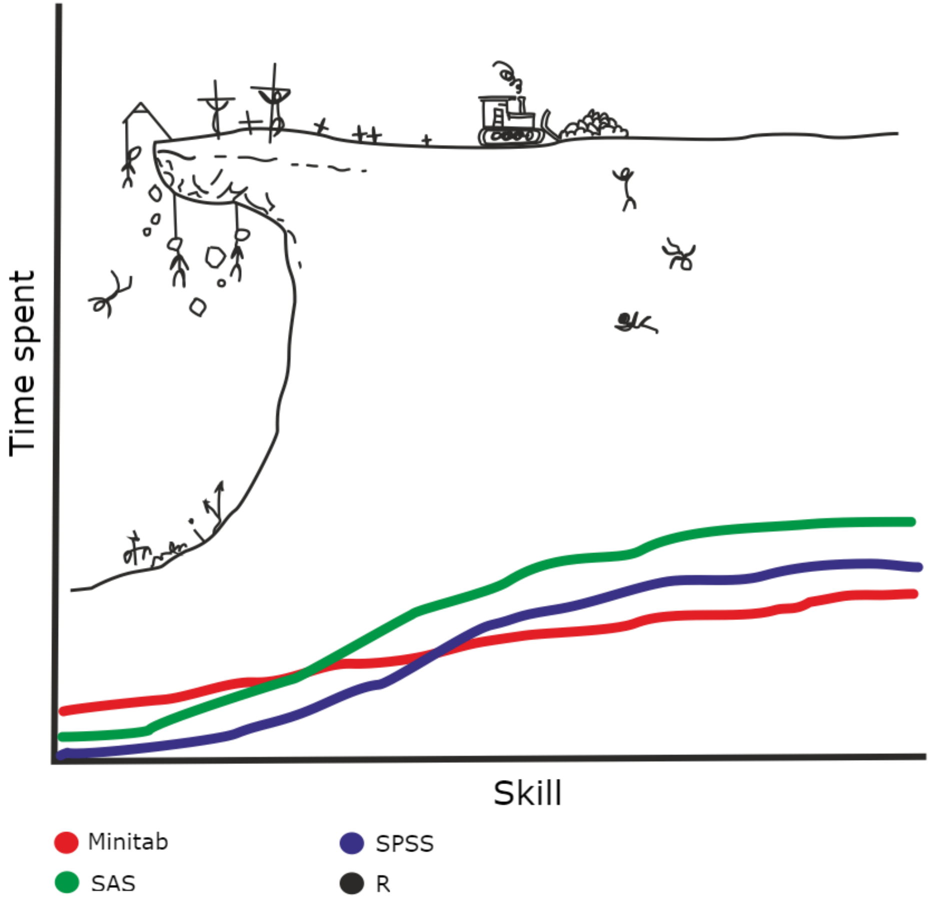
- Matrix – most methods based on this
    - 1st dim – Tokens
    - 2nd dim – Documents/units
    - Cells – frequency of tokens in documents
        - Boolean – Present vs. Not present (1/0)
        - Weighted
            - Absolute frequency (how many times word occur in document)
            - TF-IDF
- Grows large easily
    - 500 documents * easily 4k unique tokens = 2M cells
- At the same time, very sparse
    - Most of cells are empty – contain 0

Term-document matrix

	2003-2004-cz	**2004-2005-pl**	**2005-2006-hu**	**2006-2007-sk**	**2007-2008-cz**	**Sum**
agriculture	3	6	2	5	3	**19**
aim	4	2	7	12	6	**31**
area	11	8	8	28	26	**81**
base	1	2	2	2	5	**12**
border	5	9	9	3	3	**29**
central	2	3	6	3	5	**19**
cohesion	3	1	7	4	4	**19**
commission	2	7	3	2	4	**18**
common	10	9	17	8	17	**61**
community	2	2	3	3	6	**16**
concern	9	13	12	18	6	**58**

Programming in R

Learning curves of popular stats programs



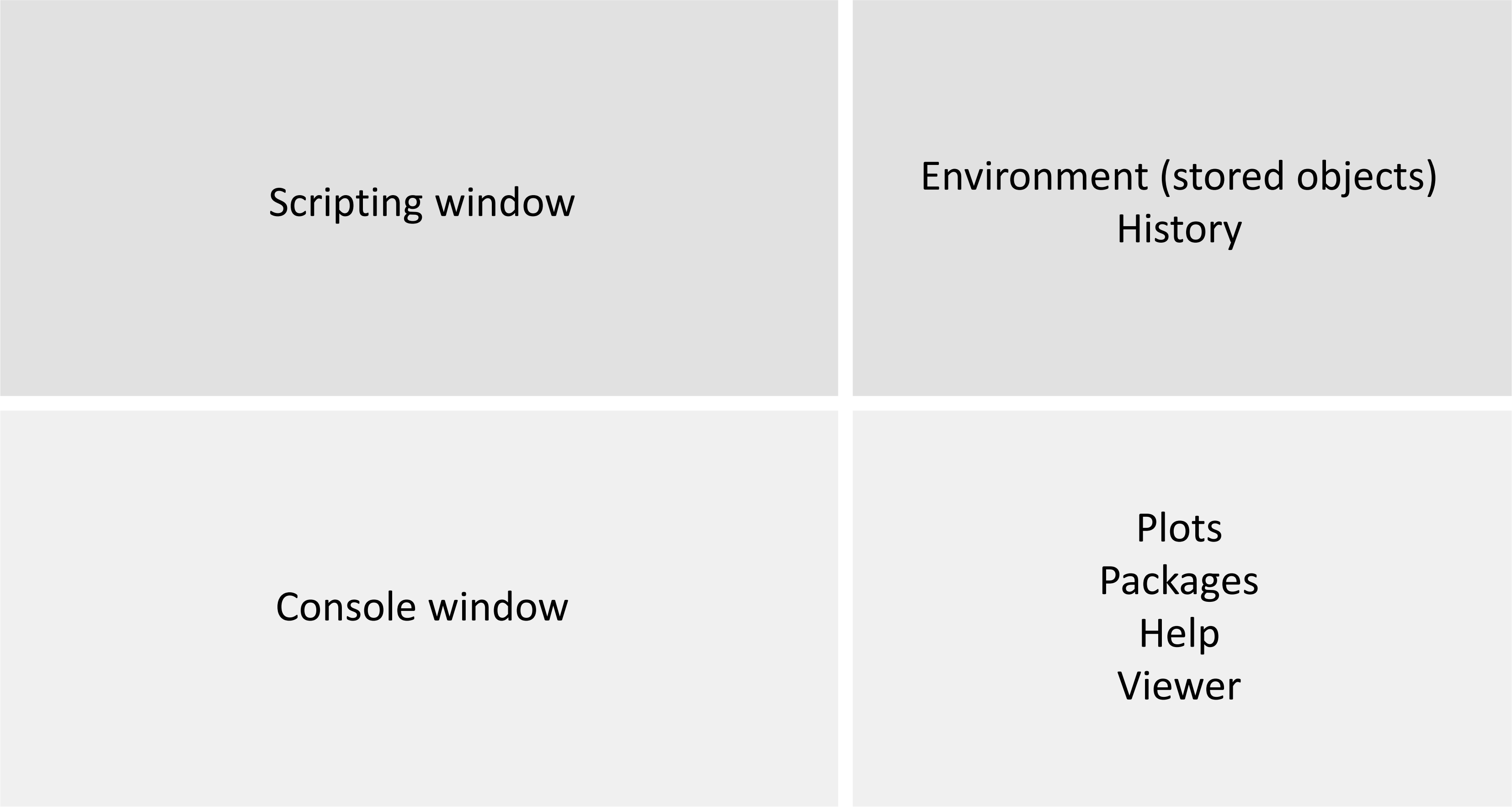
R community / resources

- there is huge number of free resources
- R package / library manuals
- R site: http://cran.r-project.org
- community forums:
  - http://stackoverflow.com
  - http://www.statmethods.net
  - http://www.r-bloggers.com
- Youtube videos: https://www.youtube.com/watch?v=qHfSTRNg6jE
- googling (often fastest)

R as language – focus on logic

- Any programming language is just very **condensed and formalized** speech
  - Just like mathematical notation
- Understand and formulate the **process**
- If you think about the **procedure** of what needs to be done, scripting becomes matter of knowing right expressions

R studio layout

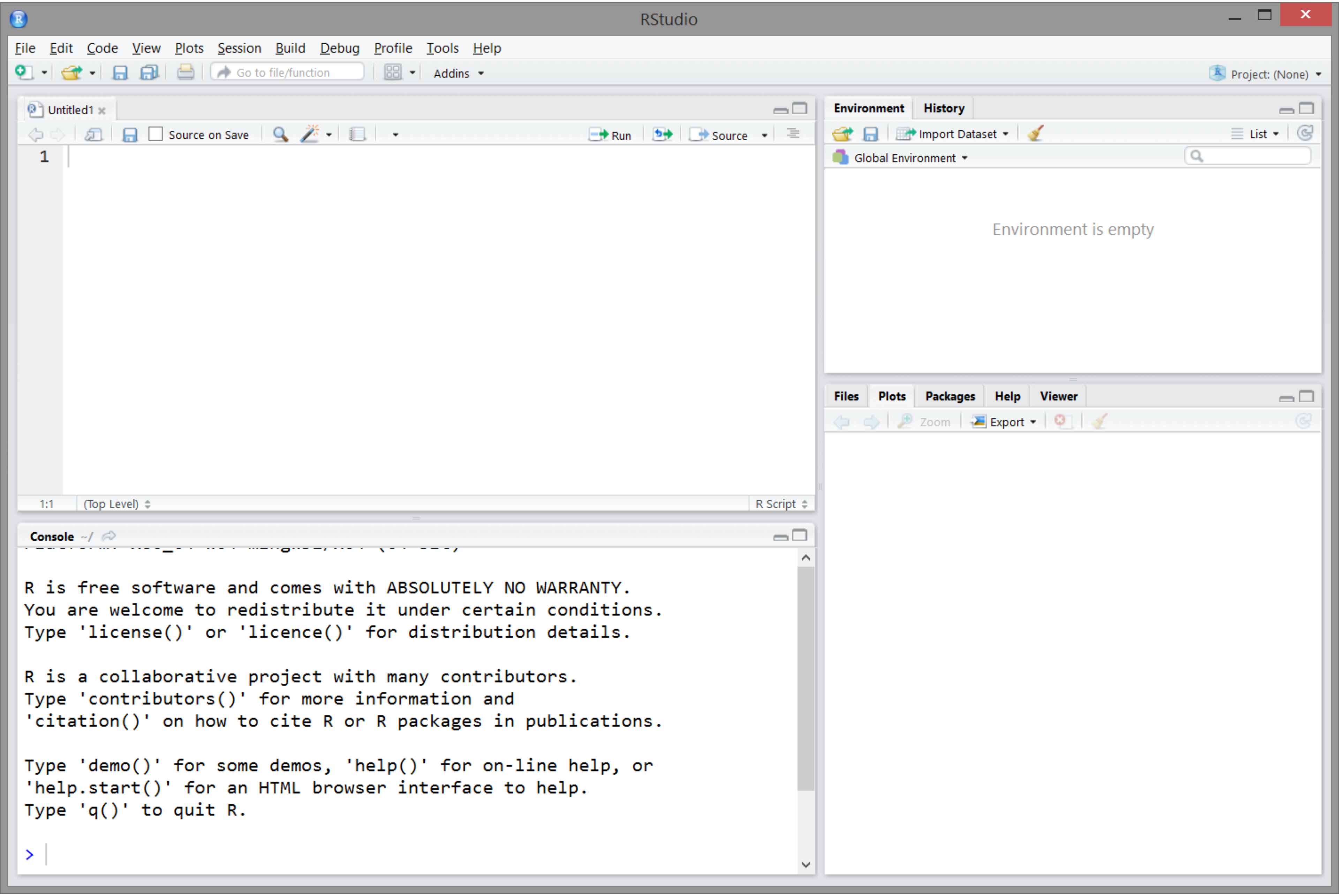


Scripting window

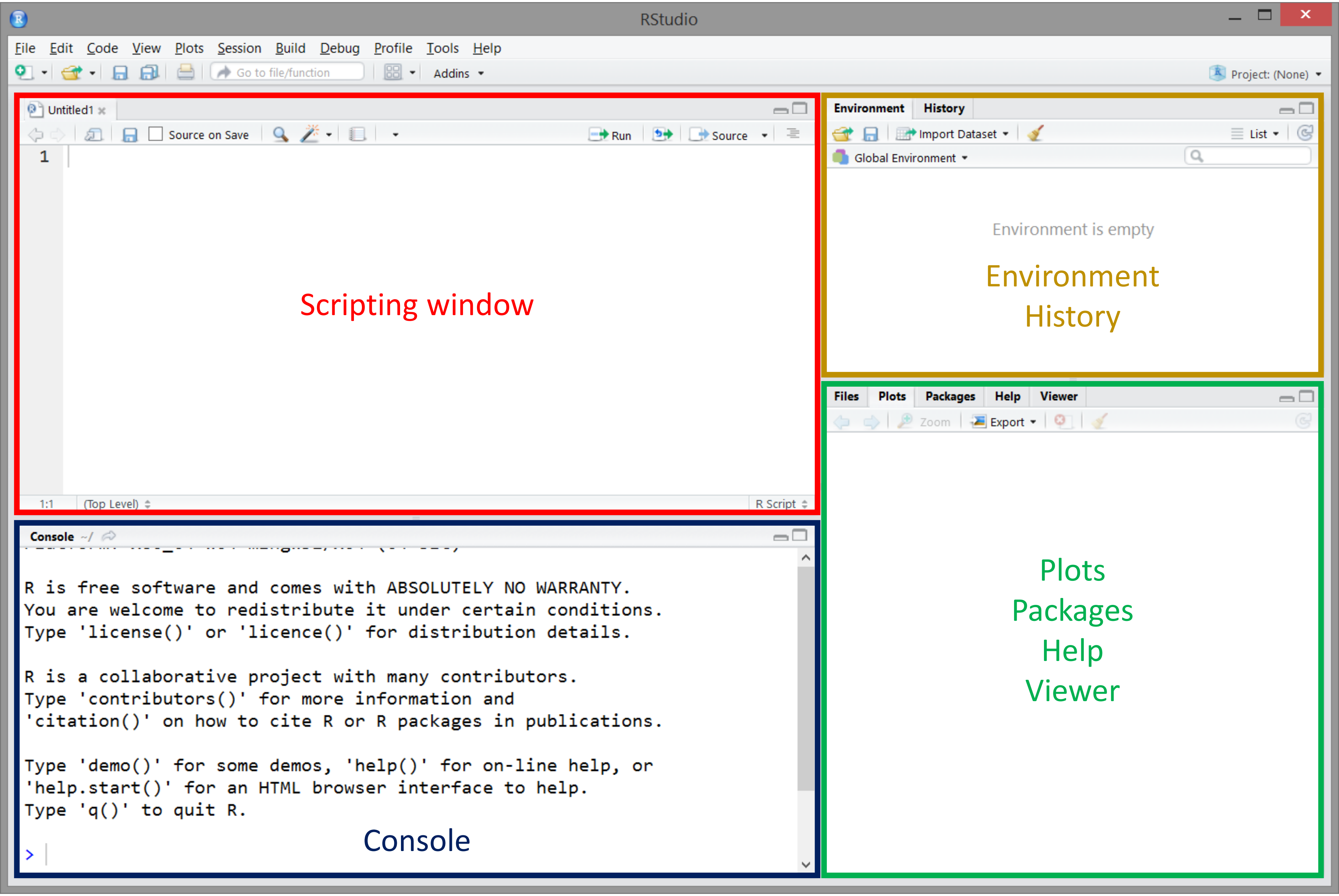
Environment (stored objects)
History

Console window

Plots
Packages
Help
Viewer



```
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```



```
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Terminology used

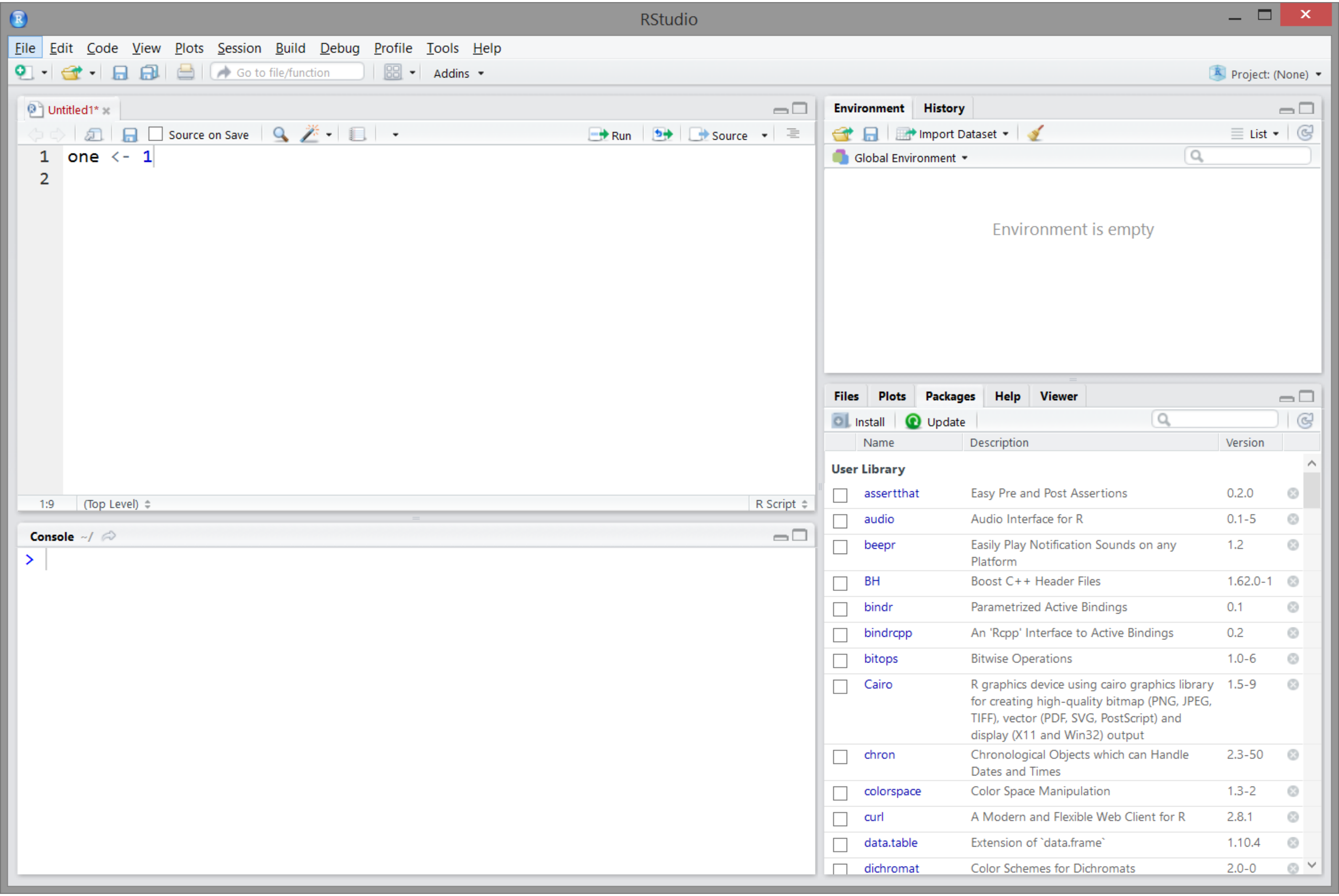
- Data, (data) element
  - **unit of information** (e.g. 1, 2, “word”, TRUE, FALSE)
- Data class
  - describes **properties of data elements** (numeric, character, logical, etc.)
- Object
  - a **“container”** that stores and organizes data in the R environment
- Object type
  - describes **properties of objects** (vector, matrix, list, data frame, etc.)
- Function
  - transforms inputs into outputs based on certain rules (methods/procedures)
  - arguments of the function specify the inputs and applied rules

Object

- object: instance of a certain data class that can be manipulated according set of procedures (methods)

```
one <- 1
```

Object

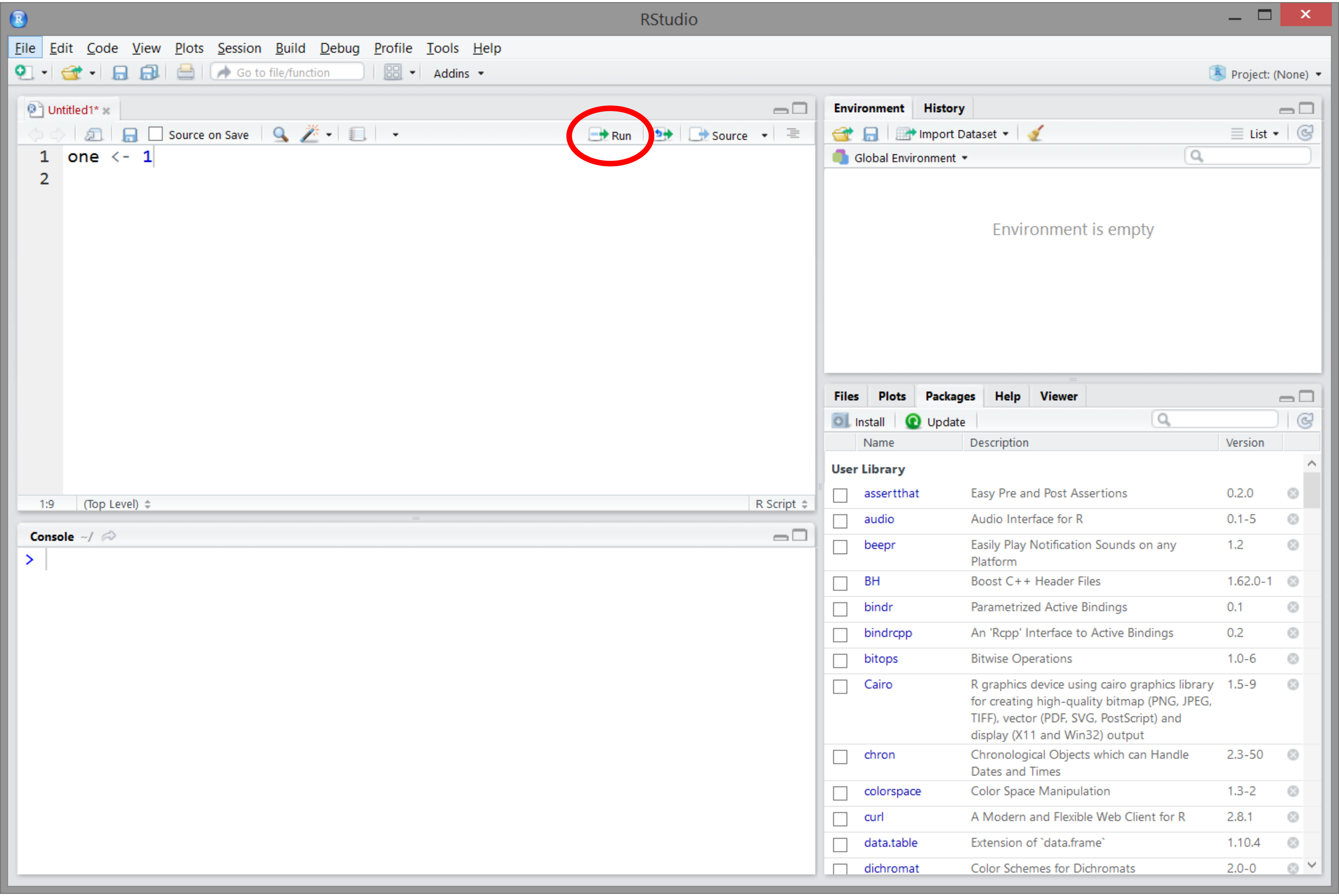


```
one <- 1
```

```
>
```

Environment is empty

	Name	Description	Version
	User Library		
□	assertthat	Easy Pre and Post Assertions	0.2.0
□	audio	Audio Interface for R	0.1-5
□	beepr	Easily Play Notification Sounds on any Platform	1.2
□	BH	Boost C++ Header Files	1.62.0-1
□	bindr	Parametrized Active Bindings	0.1
□	bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2
□	bitops	Bitwise Operations	1.0-6
□	Cairo	R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output	1.5-9
□	chron	Chronological Objects which can Handle Dates and Times	2.3-50
□	colorspace	Color Space Manipulation	1.3-2
□	curl	A Modern and Flexible Web Client for R	2.8.1
□	data.table	Extension of `data.frame`	1.10.4
□	dichromat	Color Schemes for Dichromats	2.0-0

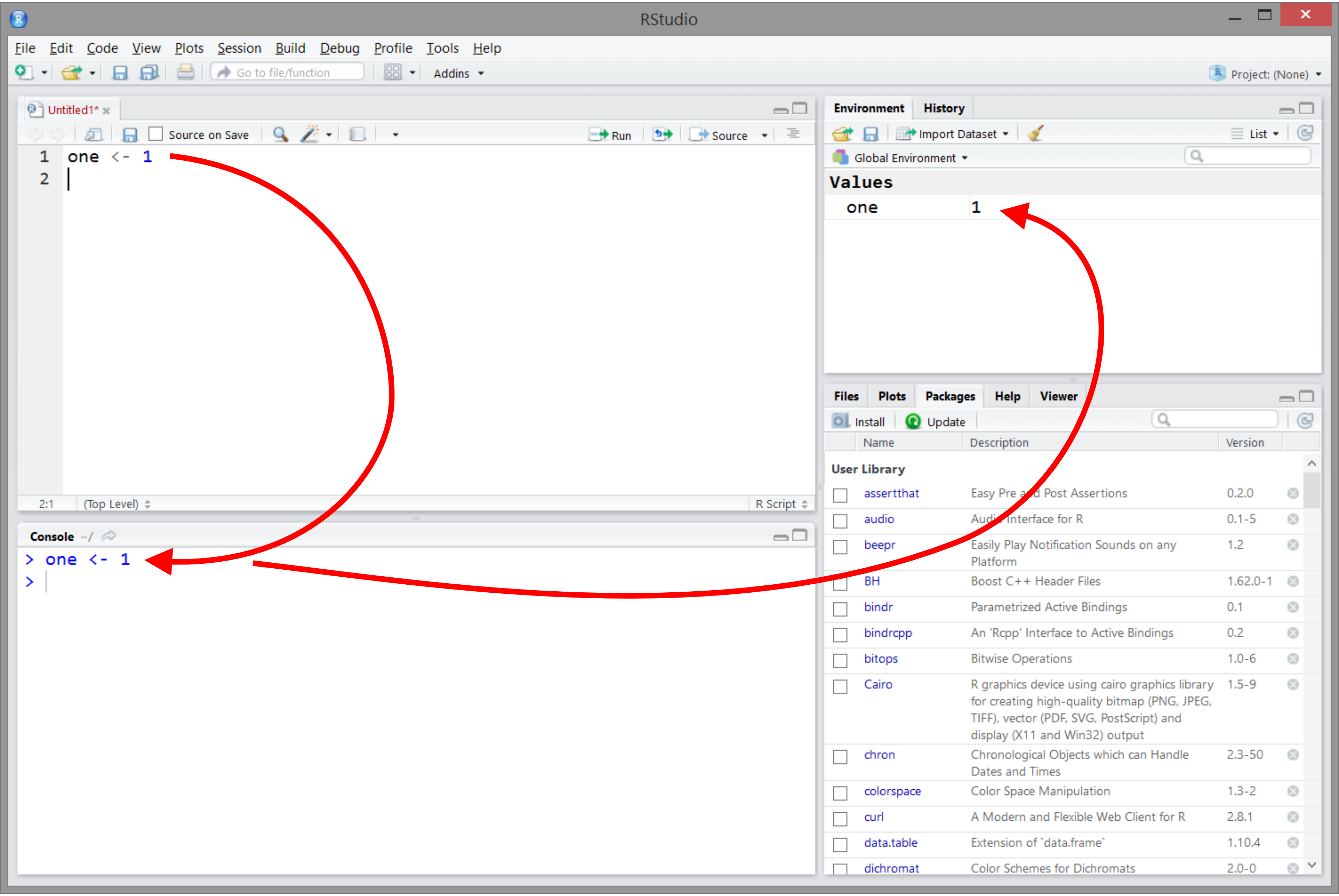


```
one <- 1
```

```
>
```

Environment is empty

Name	Description	Version
User Library		
□ assertthat	Easy Pre and Post Assertions	0.2.0
□ audio	Audio Interface for R	0.1-5
□ beepr	Easily Play Notification Sounds on any Platform	1.2
□ BH	Boost C++ Header Files	1.62.0-1
□ bindr	Parametrized Active Bindings	0.1
□ bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2
□ bitops	Bitwise Operations	1.0-6
□ Cairo	R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output	1.5-9
□ chron	Chronological Objects which can Handle Dates and Times	2.3-50
□ colorspace	Color Space Manipulation	1.3-2
□ curl	A Modern and Flexible Web Client for R	2.8.1
□ data.table	Extension of `data.frame`	1.10.4
□ dichromat	Color Schemes for Dichromats	2.0-0



```
one <- 1
```

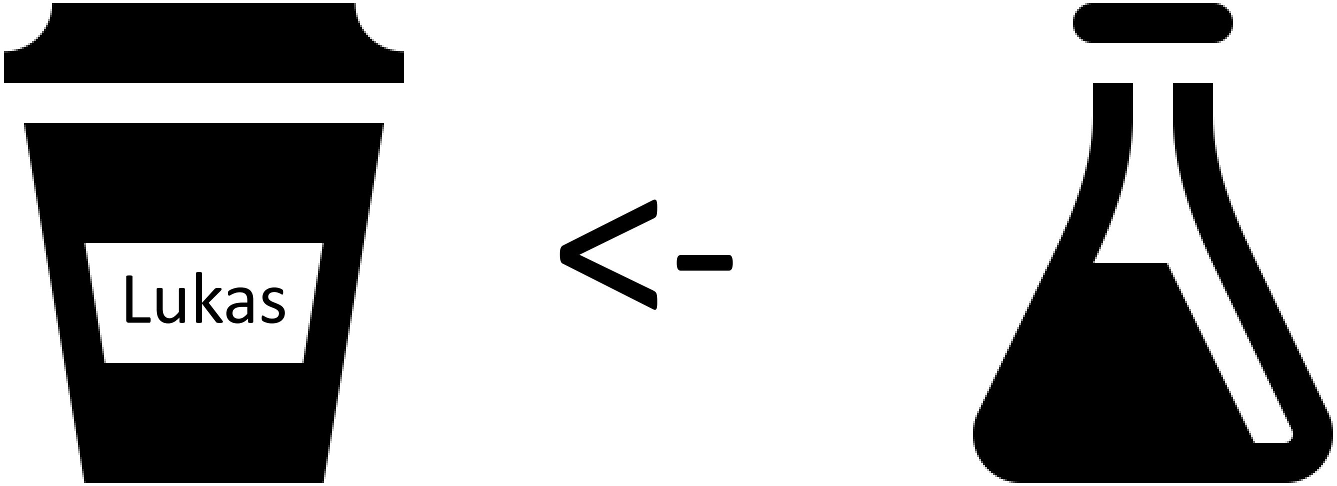
```
> one <- 1
>
```

Values

one	1

	Name	Description	Version
	User Library		
□	assertthat	Easy Pre and Post Assertions	0.2.0
□	audio	Audio Interface for R	0.1-5
□	beepr	Easily Play Notification Sounds on any Platform	1.2
□	BH	Boost C++ Header Files	1.62.0-1
□	bindr	Parametrized Active Bindings	0.1
□	bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2
□	bitops	Bitwise Operations	1.0-6
□	Cairo	R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output	1.5-9
□	chron	Chronological Objects which can Handle Dates and Times	2.3-50
□	colorspace	Color Space Manipulation	1.3-2
□	curl	A Modern and Flexible Web Client for R	2.8.1
□	data.table	Extension of `data.frame`	1.10.4
□	dichromat	Color Schemes for Dichromats	2.0-0

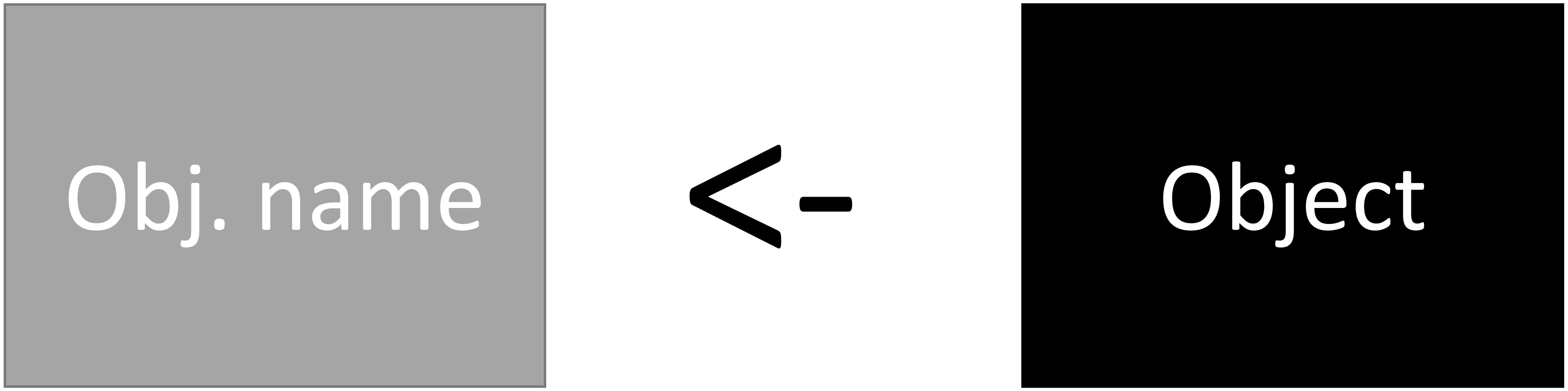
Creating/storing objects



Lukas

<-

Creating/storing objects



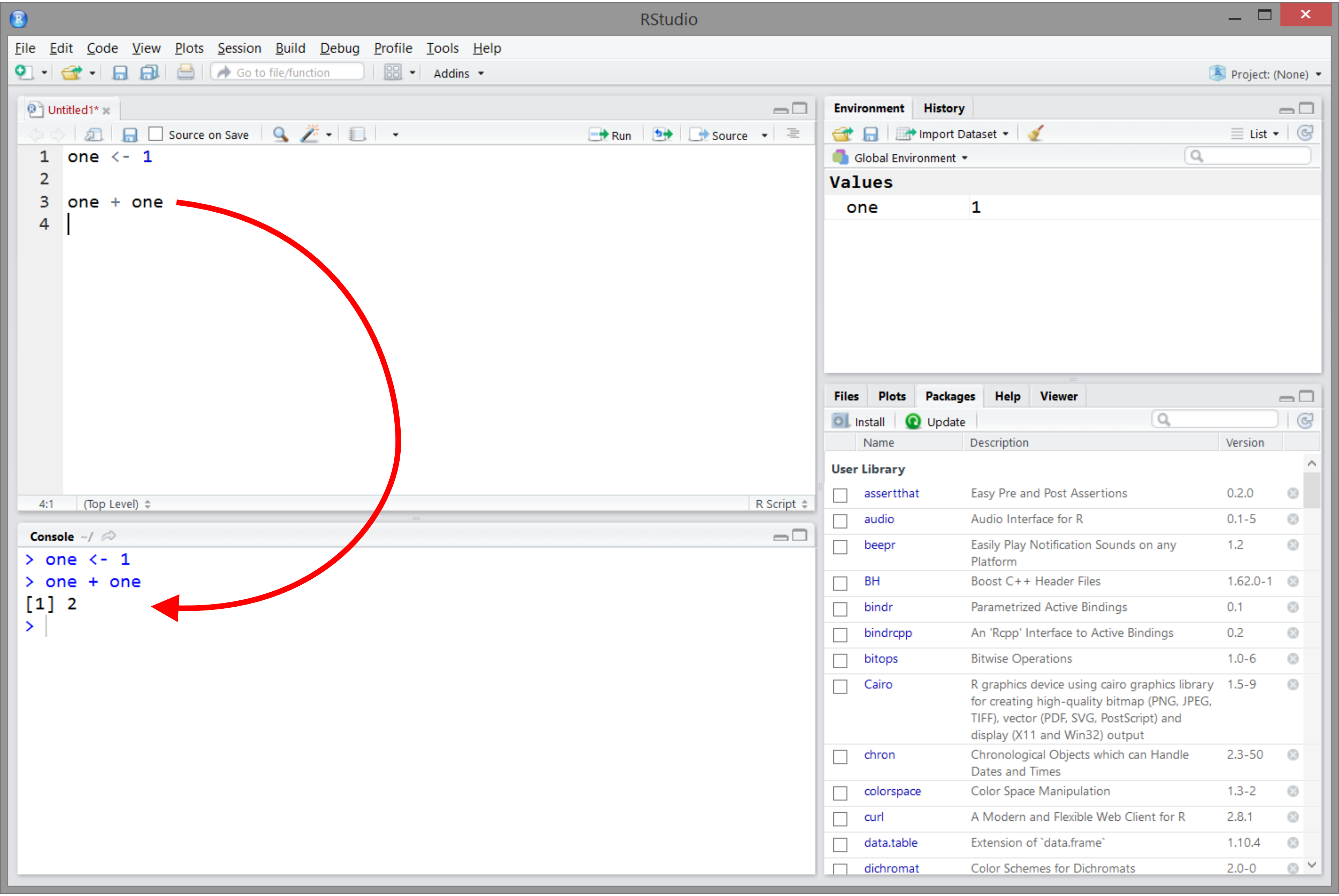
Obj. name <- Object

Object

- Once objects exist, operations over objects may be applied

```
one <- 1

one + one
```

```
> one <- 1
> one + one
[1] 2
```



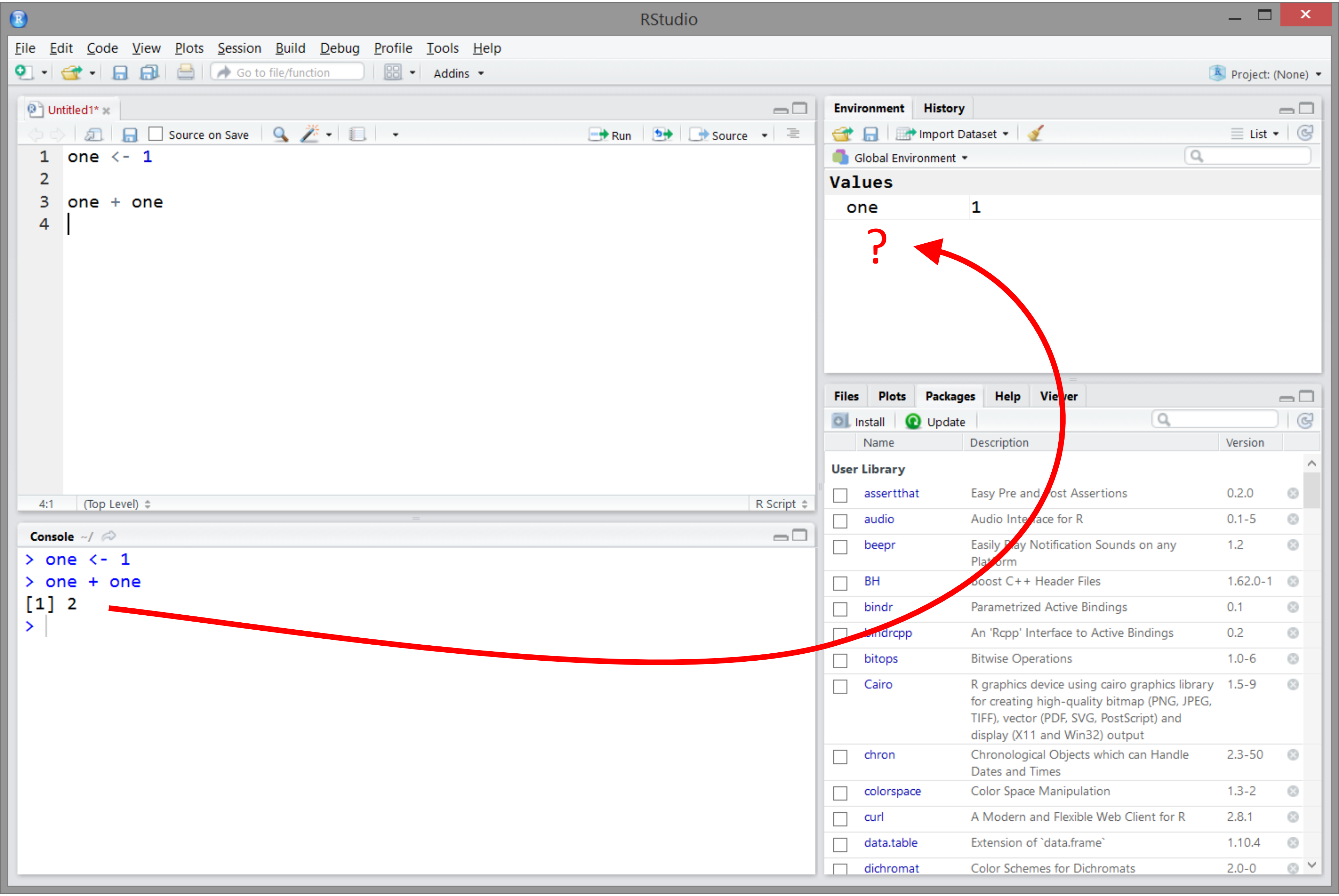
```
one <- 1

one + one
```

```
> one <- 1
> one + one
[1] 2
> 
```

Values

one	1

Name	Description	Version
User Library		
□ assertthat	Easy Pre and Post Assertions	0.2.0
□ audio	Audio Interface for R	0.1-5
□ beepr	Easily Play Notification Sounds on any Platform	1.2
□ BH	Boost C++ Header Files	1.62.0-1
□ bindr	Parametrized Active Bindings	0.1
□ bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2
□ bitops	Bitwise Operations	1.0-6
□ Cairo	R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output	1.5-9
□ chron	Chronological Objects which can Handle Dates and Times	2.3-50
□ colorspace	Color Space Manipulation	1.3-2
□ curl	A Modern and Flexible Web Client for R	2.8.1
□ data.table	Extension of `data.frame`	1.10.4
□ dichromat	Color Schemes for Dichromats	2.0-0



```
one <- 1

one + one
```

```
> one <- 1
> one + one
[1] 2
>
```

Values

one	1

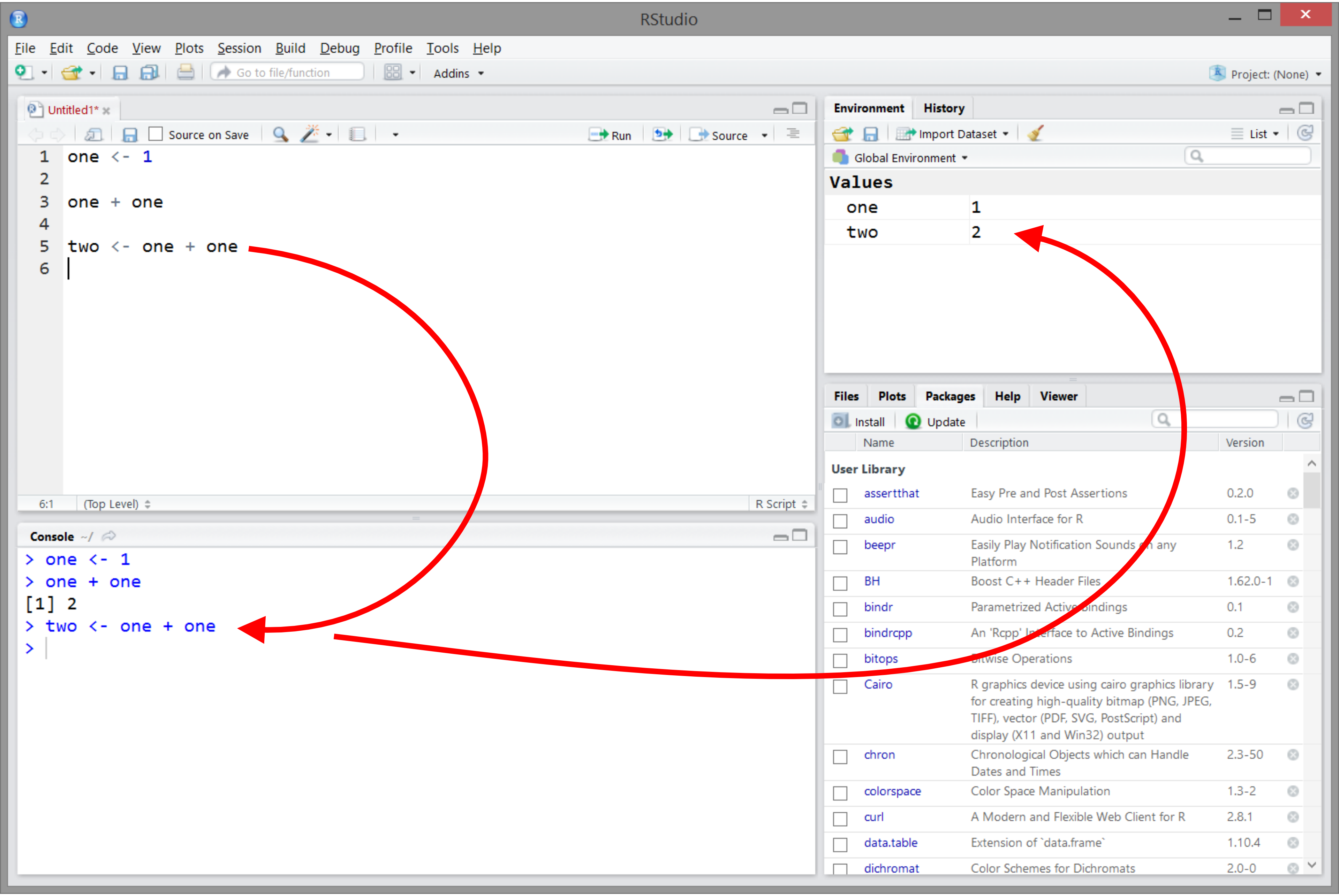
	Name	Description	Version
	User Library		
□	assertthat	Easy Pre and Post Assertions	0.2.0
□	audio	Audio Interface for R	0.1-5
□	beepr	Easily Play Notification Sounds on any Platform	1.2
□	BH	Boost C++ Header Files	1.62.0-1
□	bindr	Parametrized Active Bindings	0.1
□	bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2
□	bitops	Bitwise Operations	1.0-6
□	Cairo	R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output	1.5-9
□	chron	Chronological Objects which can Handle Dates and Times	2.3-50
□	colorspace	Color Space Manipulation	1.3-2
□	curl	A Modern and Flexible Web Client for R	2.8.1
□	data.table	Extension of `data.frame`	1.10.4
□	dichromat	Color Schemes for Dichromats	2.0-0

What is an object?

- **Anything** may become an object

- Temporary objects
  - Only **appear** in console
  - Their values must be stored in order to use them in operations

- Stored objects
  - Must be **defined** by user
  - **Remain the same** unless overwritten
  - Must be **removed** by user as well



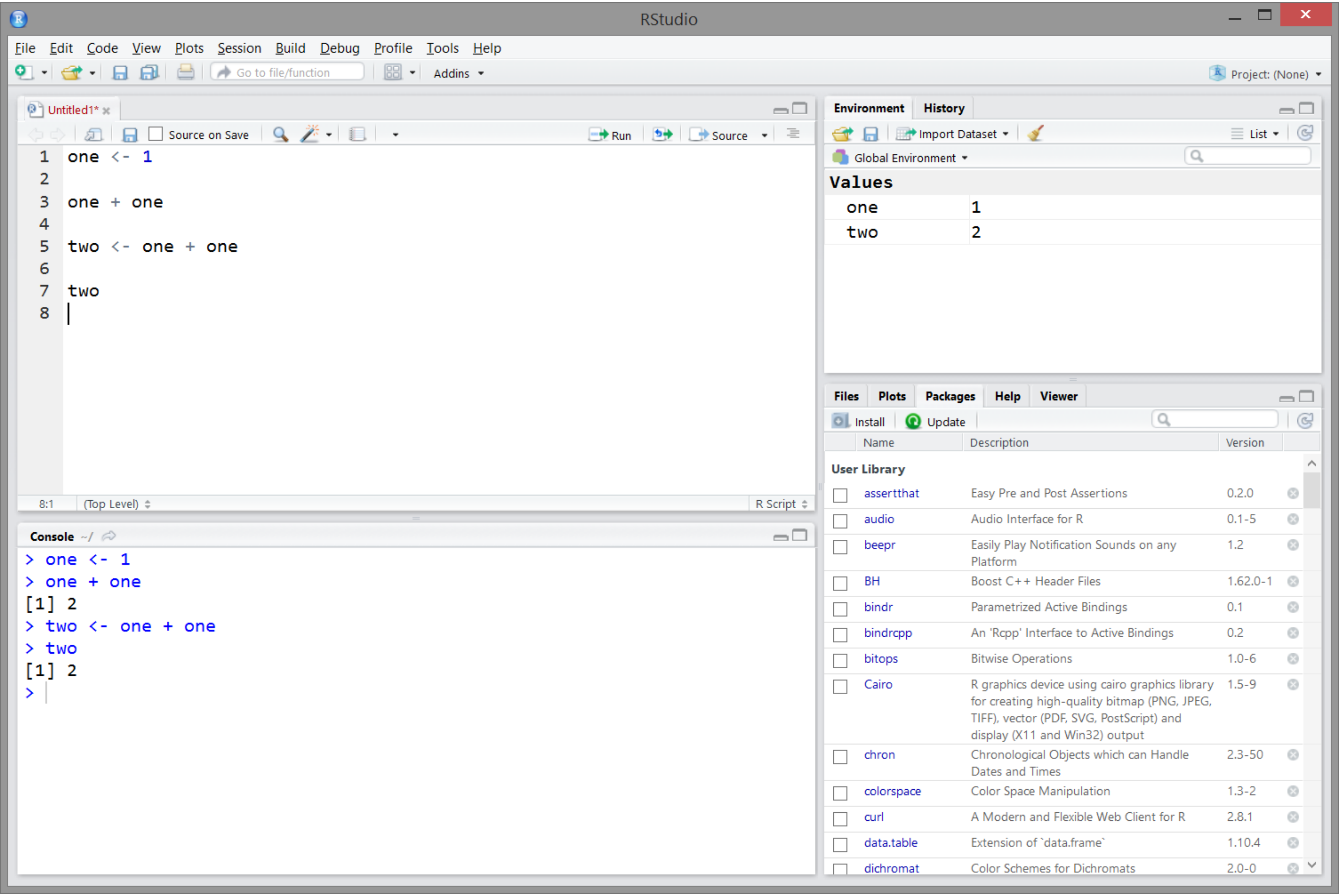
```
one <- 1

one + one

two <- one + one

```

```
> one <- 1
> one + one
[1] 2
> two <- one + one
> 
```

Values

one	1
two	2

	Name	Description	Version
	User Library		
□	assertthat	Easy Pre and Post Assertions	0.2.0
□	audio	Audio Interface for R	0.1-5
□	beepr	Easily Play Notification Sounds on any Platform	1.2
□	BH	Boost C++ Header Files	1.62.0-1
□	bindr	Parametrized Active Bindings	0.1
□	bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2
□	bitops	Bitwise Operations	1.0-6
□	Cairo	R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output	1.5-9
□	chron	Chronological Objects which can Handle Dates and Times	2.3-50
□	colorspace	Color Space Manipulation	1.3-2
□	curl	A Modern and Flexible Web Client for R	2.8.1
□	data.table	Extension of `data.frame`	1.10.4
□	dichromat	Color Schemes for Dichromats	2.0-0



```
one <- 1

one + one

two <- one + one

two

```

```
> one <- 1
> one + one
[1] 2
> two <- one + one
> two
[1] 2
> 
```

Values

one	1
two	2

	Name	Description	Version
	User Library		
□	assertthat	Easy Pre and Post Assertions	0.2.0
□	audio	Audio Interface for R	0.1-5
□	beepr	Easily Play Notification Sounds on any Platform	1.2
□	BH	Boost C++ Header Files	1.62.0-1
□	bindr	Parametrized Active Bindings	0.1
□	bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2
□	bitops	Bitwise Operations	1.0-6
□	Cairo	R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output	1.5-9
□	chron	Chronological Objects which can Handle Dates and Times	2.3-50
□	colorspace	Color Space Manipulation	1.3-2
□	curl	A Modern and Flexible Web Client for R	2.8.1
□	data.table	Extension of `data.frame`	1.10.4
□	dichromat	Color Schemes for Dichromats	2.0-0

Data classes – prop. of elements

- Numeric
  - continuous numeric data
  - -1, 0.5, 10.49
- Integer
  - discrete numeric data
  - -1, 0, 1
- Character
  - string values
  - "anythingWithinQuotes"
- Logical
  - output of logical operation – TRUE/FALSE
  - 5 > 10

Data classes

```
> as.numeric(10.49)
[1] 10.49
>
> as.integer(10.49)
[1] 10
>
> as.character(-1)
[1] "-1"
>
> as.numeric("anythingwithinquotes")
[1] NA Warning message: NAs introduced by coercion
>
> 5 > 10
[1] FALSE
>
> as.character(5 > 10)
[1] "FALSE"
```

Object types – prop. of objects

Object types – prop. of objects

- vector
  - sequence (1-dimensional) of elements of same data class
- matrix
  - 2-dimensional rectangular collection of elements of same data class
  - array: n-dimensional matrix
- list
  - vector that can contain elements of different data classes
- data frame
  - list of vectors of equal length
  - table data

Vector

```
> c(2,3,5)
[1] 2 3 5
>
> c("aa", "bb", "cc", "dd", "ee")
[1] "aa" "bb" "cc" "dd" "ee“
>
> c(TRUE, FALSE, TRUE, FALSE, FALSE)
[1] TRUE FALSE TRUE FALSE FALSE
>
```

Matrix

```
> m <- matrix(data = c(1,2,3,4,5,6,7,8,9,10,11,12),
+                  nrow = 3,
+                  ncol = 4)
>
> m
      [,1]  [,2]  [,3]  [,4]
[1,]  1     4     7     10
[2,]  2     5     8     11
[3,]  3     6     9     12
>
```

List

```
> n <- c(2, 3, 5)
> s <- c("aa", "bb", "cc", "dd", "ee")
> x <- list(n, s, b, 3) # x contains copy of n, s
> x
[[1]]
[1] 2 3 5

[[2]]
[1] "aa" "bb" "cc" "dd" "ee"

[[3]]
[1] TRUE FALSE TRUE FALSE FALSE

[[4]]
[1] 3
```

Data frame

```
> teams <- c("PHI","NYM","FLA","ATL","WSN")
> wins <- c(92,89,94,72,59)
> losses <- c(70,73,77,90,102)
>
> data <- data.frame(teams,wins,losses)
>
> data
      teams wins  losses
1     PHI   92    70
2     NYM   89    73
3     FLA   94    77
4     ATL   72    90
5     WSN   59    102
>
```

R functions

- `word()` indicates function

```
> sqrt(9)
[1] 3
```

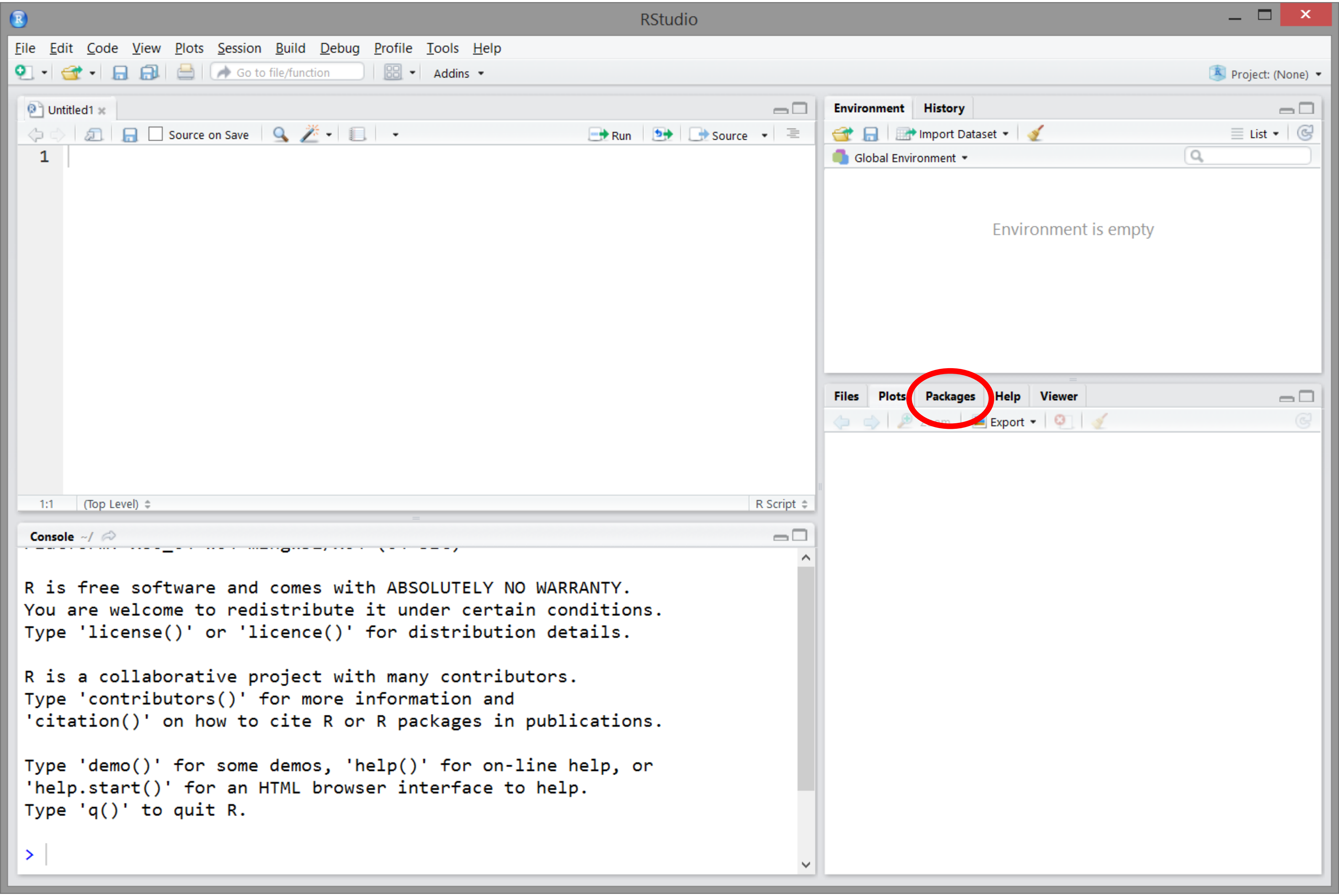
- `function(argument_1, argument_2, …)`

```
> sample(x = 0:100, size = 10, rep = FALSE)
[1] 48 50 37 94 42 39 21 19 63 95
```

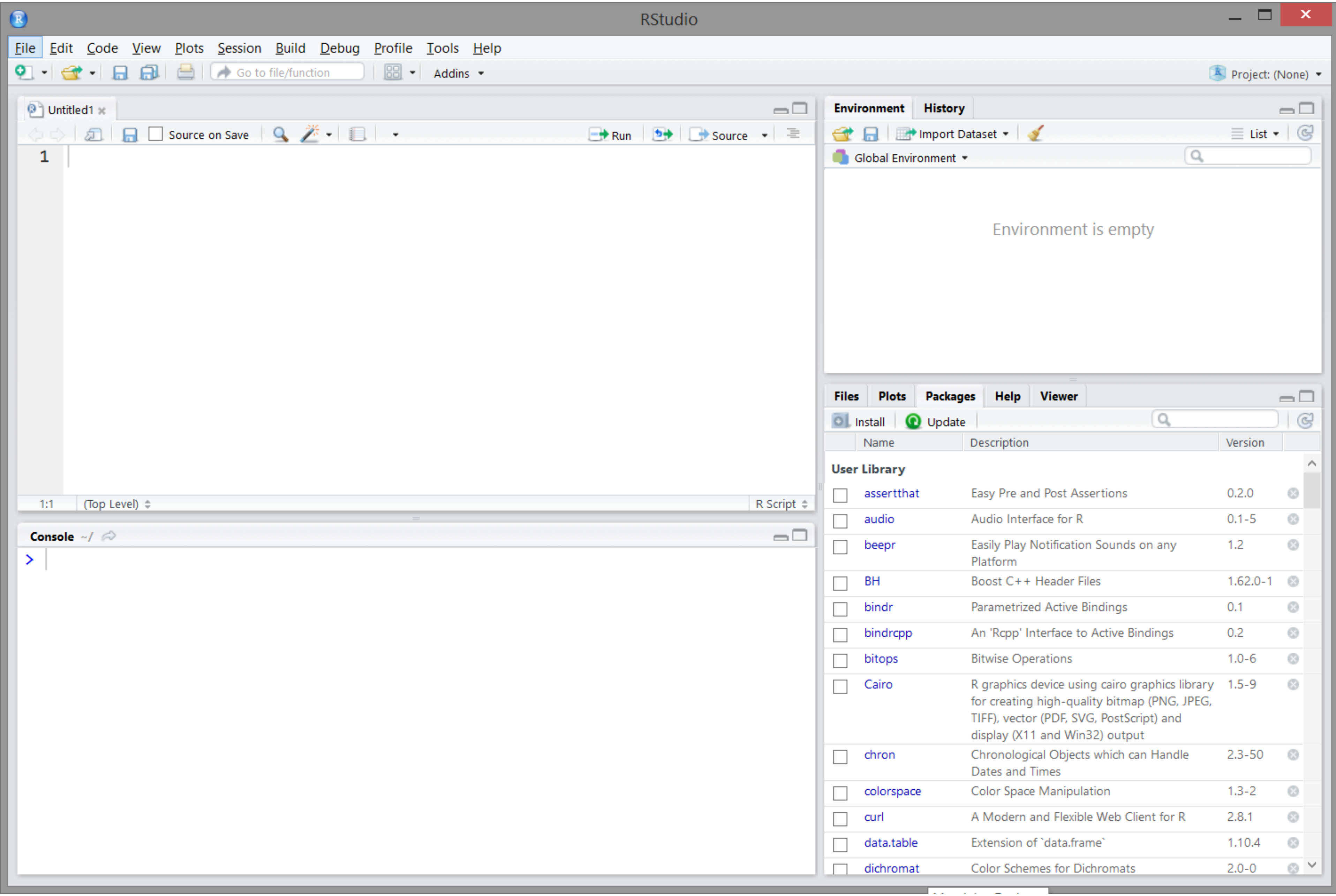
- basic functions (part of the basic R package)
- package functions (part of the particular package)
- user functions (user-defined functions)

R libraries

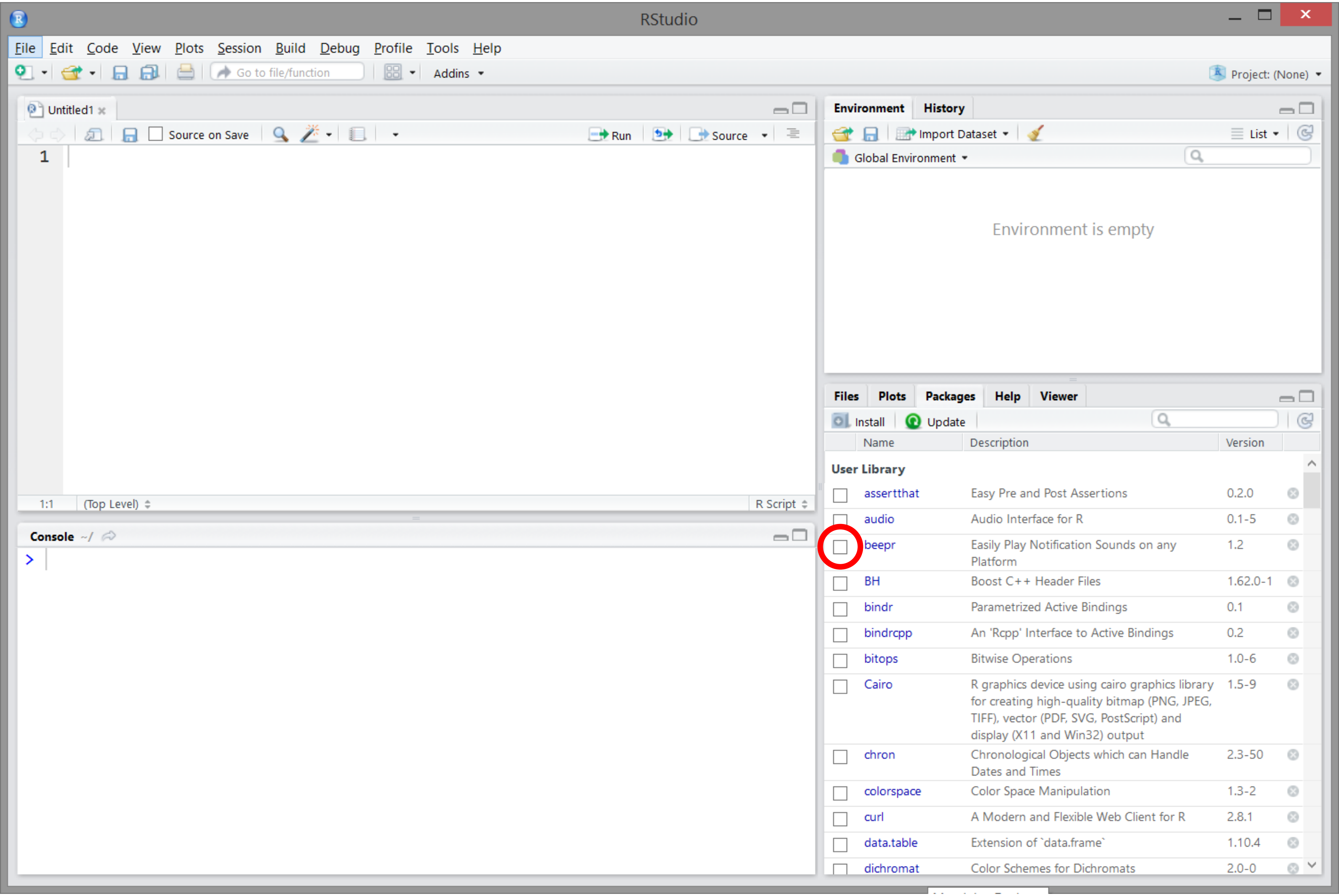
- Libraries allow to load pre-defined functions according to problem at hand
- Load, install and unload either using R Studio or using functions in script
- Libraries download and install automatically



```
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```



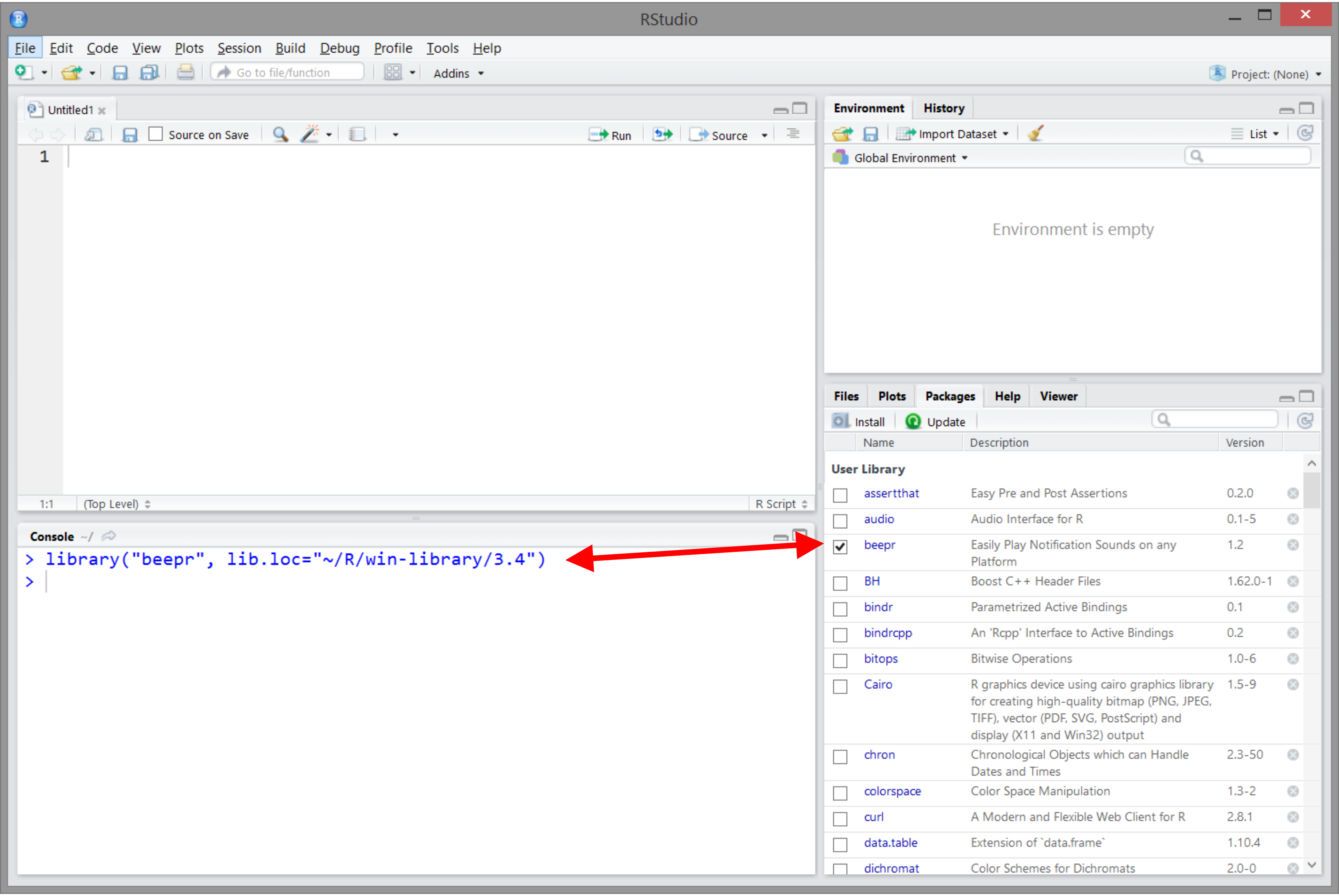
	Name	Description	Version
	User Library		
□	assertthat	Easy Pre and Post Assertions	0.2.0
□	audio	Audio Interface for R	0.1-5
□	beepr	Easily Play Notification Sounds on any Platform	1.2
□	BH	Boost C++ Header Files	1.62.0-1
□	bindr	Parametrized Active Bindings	0.1
□	bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2
□	bitops	Bitwise Operations	1.0-6
□	Cairo	R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output	1.5-9
□	chron	Chronological Objects which can Handle Dates and Times	2.3-50
□	colorspace	Color Space Manipulation	1.3-2
□	curl	A Modern and Flexible Web Client for R	2.8.1
□	data.table	Extension of `data.frame`	1.10.4
□	dichromat	Color Schemes for Dichromats	2.0-0



Environment is empty

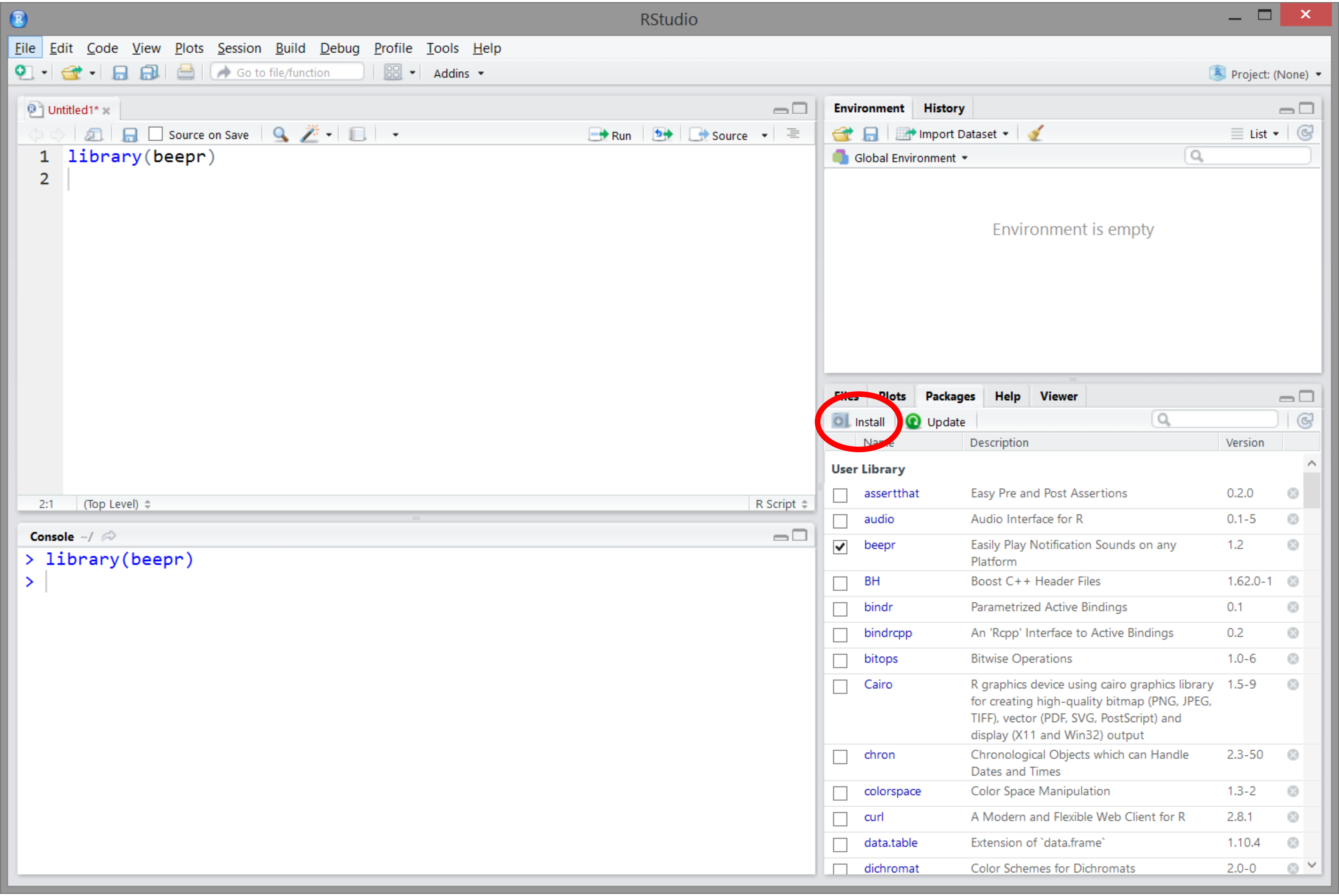
User Library

	Name	Description	Version
☐	assertthat	Easy Pre and Post Assertions	0.2.0
☐	audio	Audio Interface for R	0.1-5
☐	beepr	Easily Play Notification Sounds on any Platform	1.2
☐	BH	Boost C++ Header Files	1.62.0-1
☐	bindr	Parametrized Active Bindings	0.1
☐	bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2
☐	bitops	Bitwise Operations	1.0-6
☐	Cairo	R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output	1.5-9
☐	chron	Chronological Objects which can Handle Dates and Times	2.3-50
☐	colorspace	Color Space Manipulation	1.3-2
☐	curl	A Modern and Flexible Web Client for R	2.8.1
☐	data.table	Extension of `data.frame`	1.10.4
☐	dichromat	Color Schemes for Dichromats	2.0-0



```
> library("beepr", lib.loc="~/R/win-library/3.4")
> 
```

	Name	Description	Version
	User Library		
□	assertthat	Easy Pre and Post Assertions	0.2.0
□	audio	Audio Interface for R	0.1-5
☑	beepr	Easily Play Notification Sounds on any Platform	1.2
□	BH	Boost C++ Header Files	1.62.0-1
□	bindr	Parametrized Active Bindings	0.1
□	bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2
□	bitops	Bitwise Operations	1.0-6
□	Cairo	R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output	1.5-9
□	chron	Chronological Objects which can Handle Dates and Times	2.3-50
□	colorspace	Color Space Manipulation	1.3-2
□	curl	A Modern and Flexible Web Client for R	2.8.1
□	data.table	Extension of `data.frame`	1.10.4
□	dichromat	Color Schemes for Dichromats	2.0-0

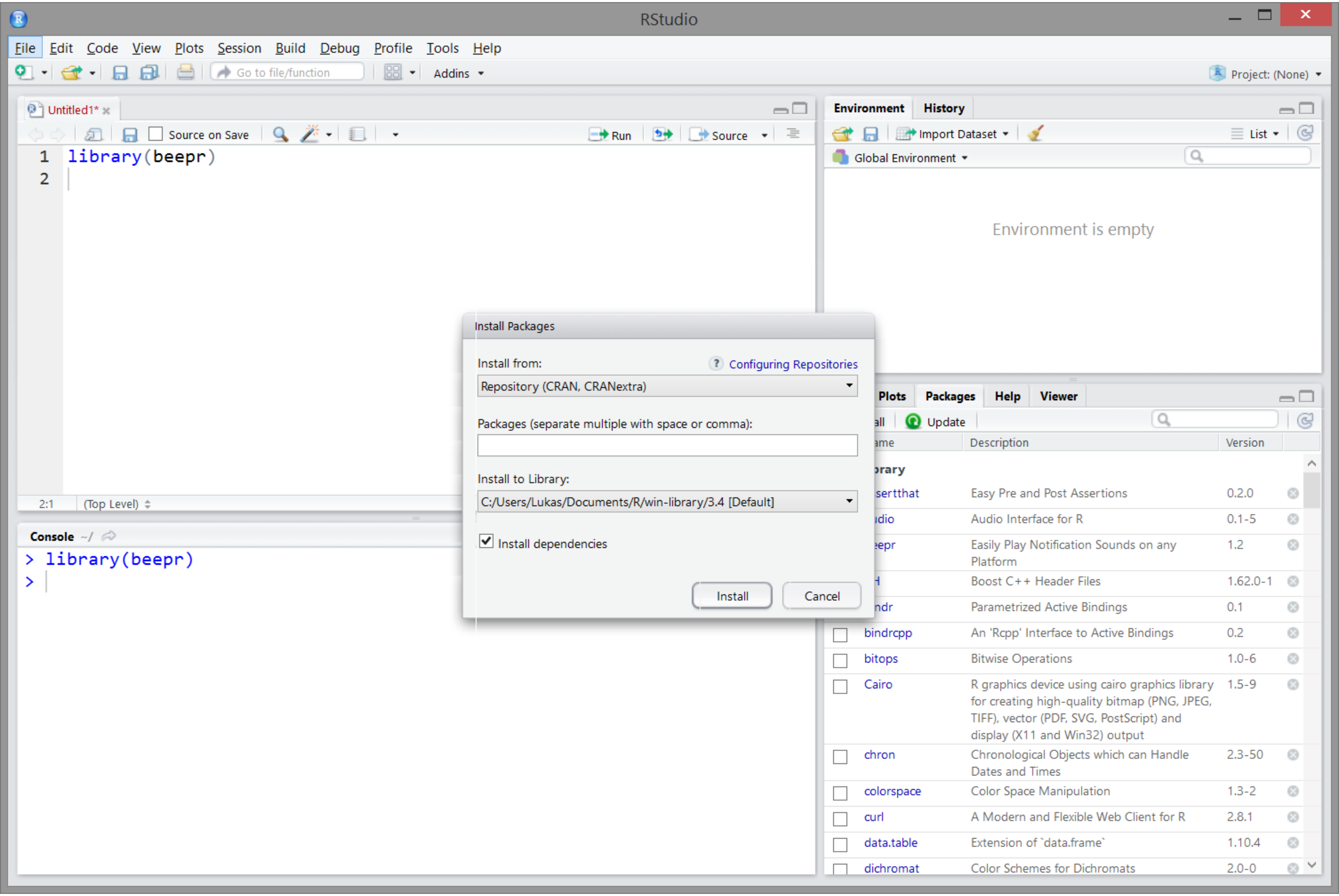


```
library(beepr)
```

```
> library(beepr)
> 
```

Environment is empty

	Name	Description	Version
	User Library		
□	assertthat	Easy Pre and Post Assertions	0.2.0
□	audio	Audio Interface for R	0.1-5
☑	beepr	Easily Play Notification Sounds on any Platform	1.2
□	BH	Boost C++ Header Files	1.62.0-1
□	bindr	Parametrized Active Bindings	0.1
□	bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2
□	bitops	Bitwise Operations	1.0-6
□	Cairo	R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output	1.5-9
□	chron	Chronological Objects which can Handle Dates and Times	2.3-50
□	colorspace	Color Space Manipulation	1.3-2
□	curl	A Modern and Flexible Web Client for R	2.8.1
□	data.table	Extension of `data.frame`	1.10.4
□	dichromat	Color Schemes for Dichromats	2.0-0



```
library(beepr)
```

```
> library(beepr)
>
```

Install Packages

Install from: Repository (CRAN, CRANextra)

? Configuring Repositories

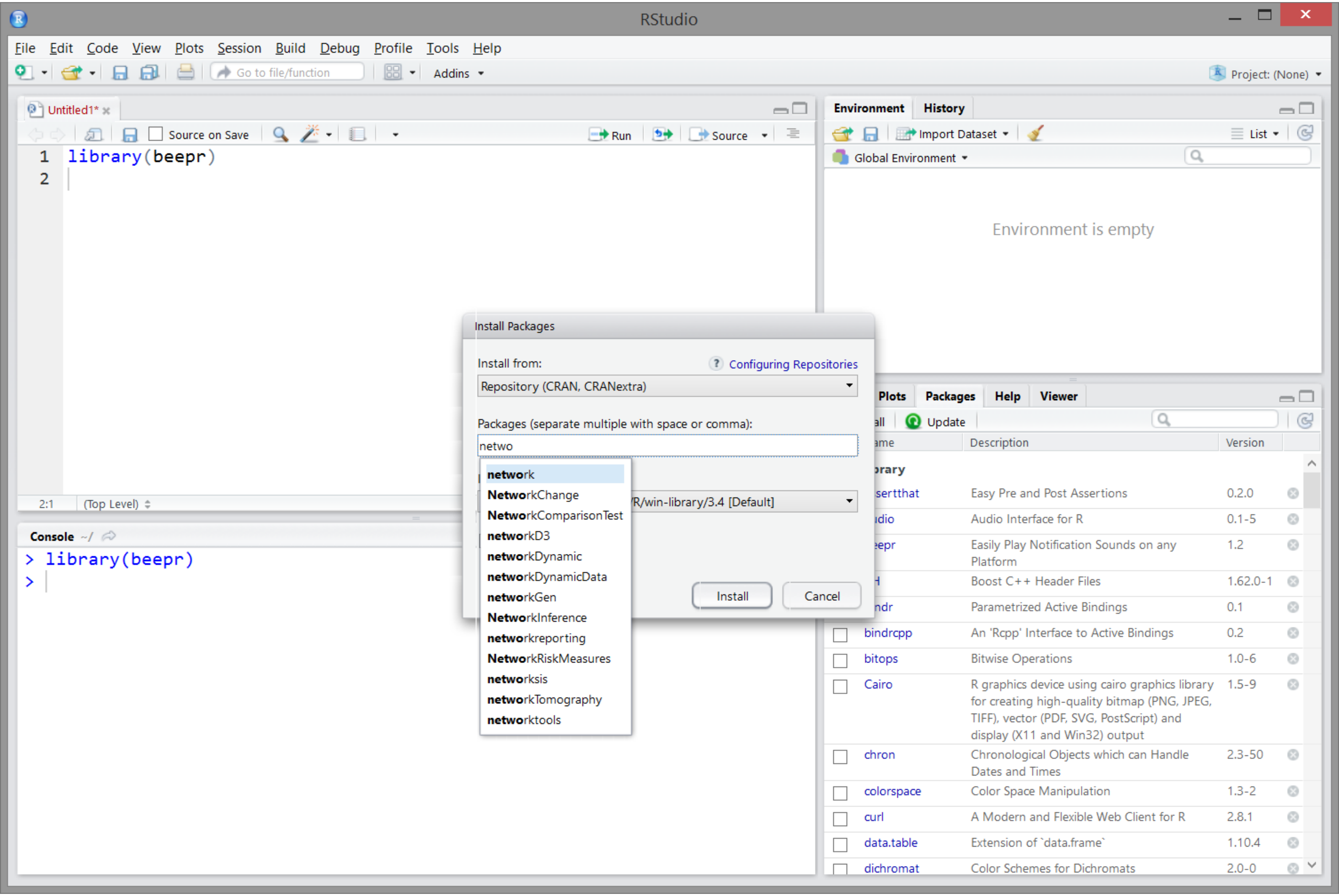
Packages (separate multiple with space or comma):

Install to Library: C:/Users/Lukas/Documents/R/win-library/3.4 [Default]

☑ Install dependencies

Install Cancel

	ame	Description	Version
	orary		
	sertthat	Easy Pre and Post Assertions	0.2.0
	udio	Audio Interface for R	0.1-5
	eepr	Easily Play Notification Sounds on any Platform	1.2
	H	Boost C++ Header Files	1.62.0-1
	ndr	Parametrized Active Bindings	0.1
□	bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2
□	bitops	Bitwise Operations	1.0-6
□	Cairo	R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output	1.5-9
□	chron	Chronological Objects which can Handle Dates and Times	2.3-50
□	colorspace	Color Space Manipulation	1.3-2
□	curl	A Modern and Flexible Web Client for R	2.8.1
□	data.table	Extension of `data.frame`	1.10.4
□	dichromat	Color Schemes for Dichromats	2.0-0



```
library(beepr)
```

Console:

```
> library(beepr)
>
```

Environment is empty

Install Packages

Install from: Configuring Repositories

Repository (CRAN, CRANextra)

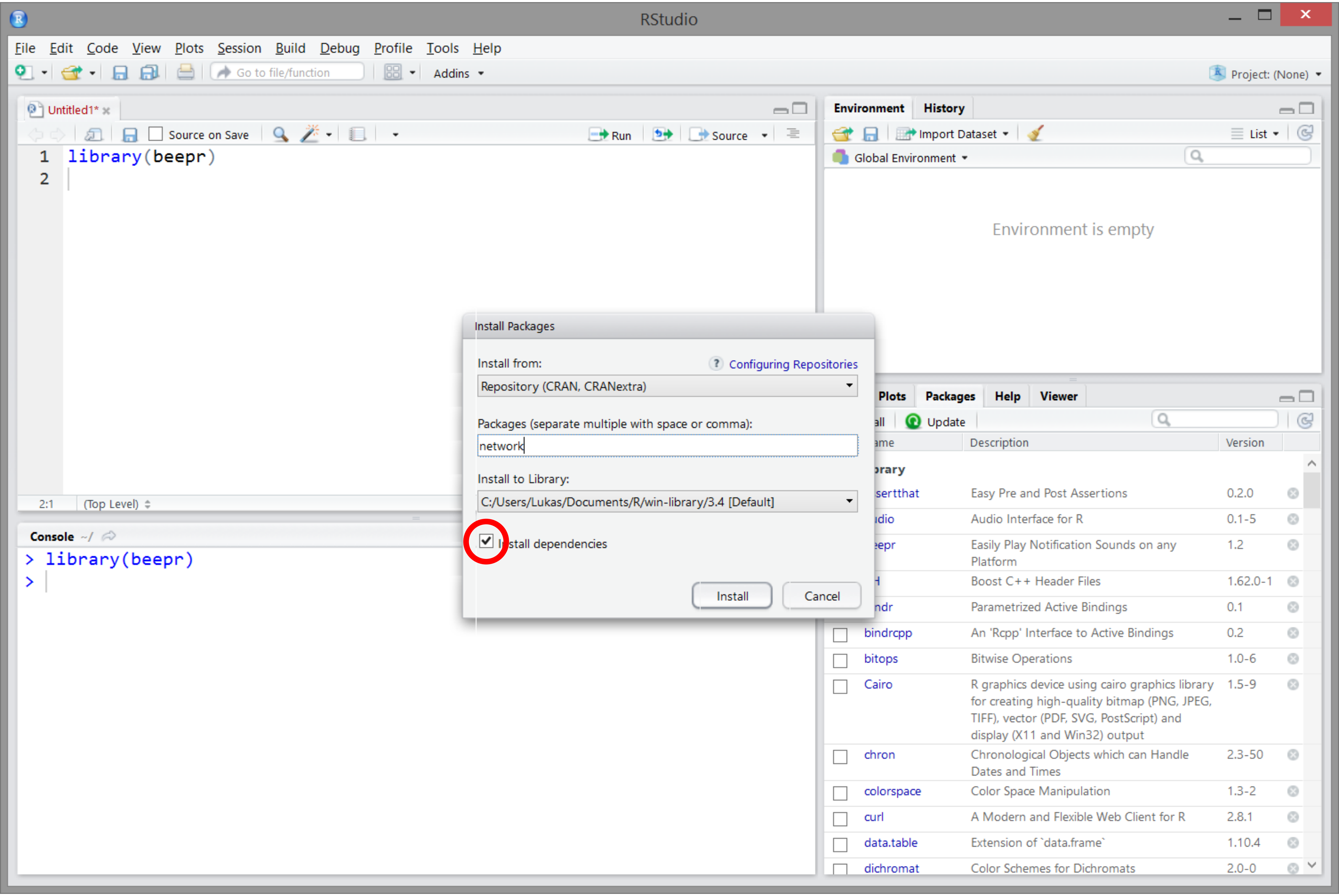
Packages (separate multiple with space or comma):

netwo

/R/win-library/3.4 [Default]

network
NetworkChange
NetworkComparisonTest
networkD3
networkDynamic
networkDynamicData
networkGen
NetworkInference
networkreporting
NetworkRiskMeasures
networksis
networkTomography
networktools

ame	Description	Version
orary		
sertthat	Easy Pre and Post Assertions	0.2.0
dio	Audio Interface for R	0.1-5
epr	Easily Play Notification Sounds on any Platform	1.2
H	Boost C++ Header Files	1.62.0-1
ndr	Parametrized Active Bindings	0.1
□ bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2
□ bitops	Bitwise Operations	1.0-6
□ Cairo	R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output	1.5-9
□ chron	Chronological Objects which can Handle Dates and Times	2.3-50
□ colorspace	Color Space Manipulation	1.3-2
□ curl	A Modern and Flexible Web Client for R	2.8.1
□ data.table	Extension of `data.frame`	1.10.4
□ dichromat	Color Schemes for Dichromats	2.0-0



```
library(beepr)
```

Console:

```
> library(beepr)
>
```

Install Packages

Install from: ? Configuring Repositories

Repository (CRAN, CRANextra)

Packages (separate multiple with space or comma):

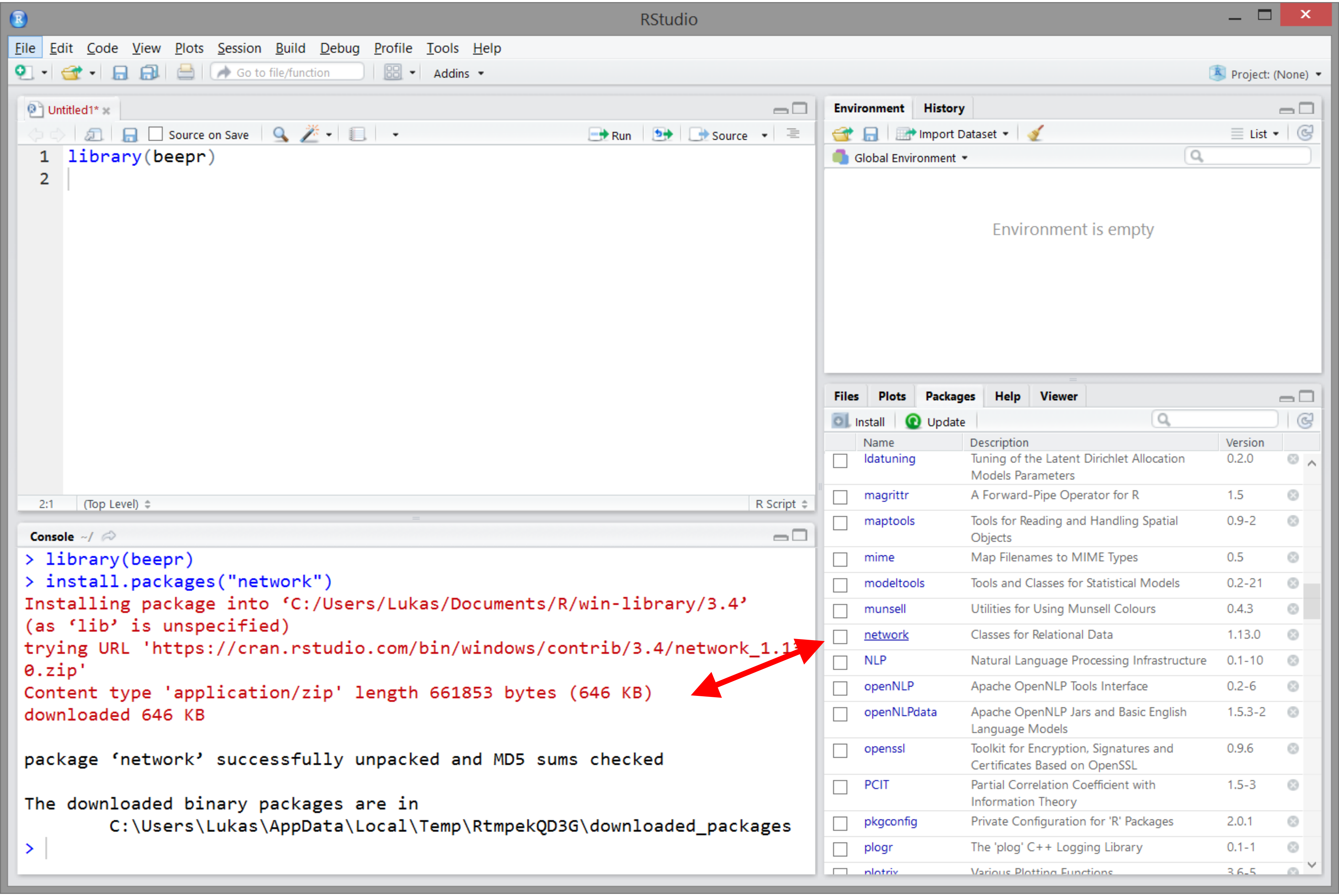
network

Install to Library:

C:/Users/Lukas/Documents/R/win-library/3.4 [Default]

☑ Install dependencies

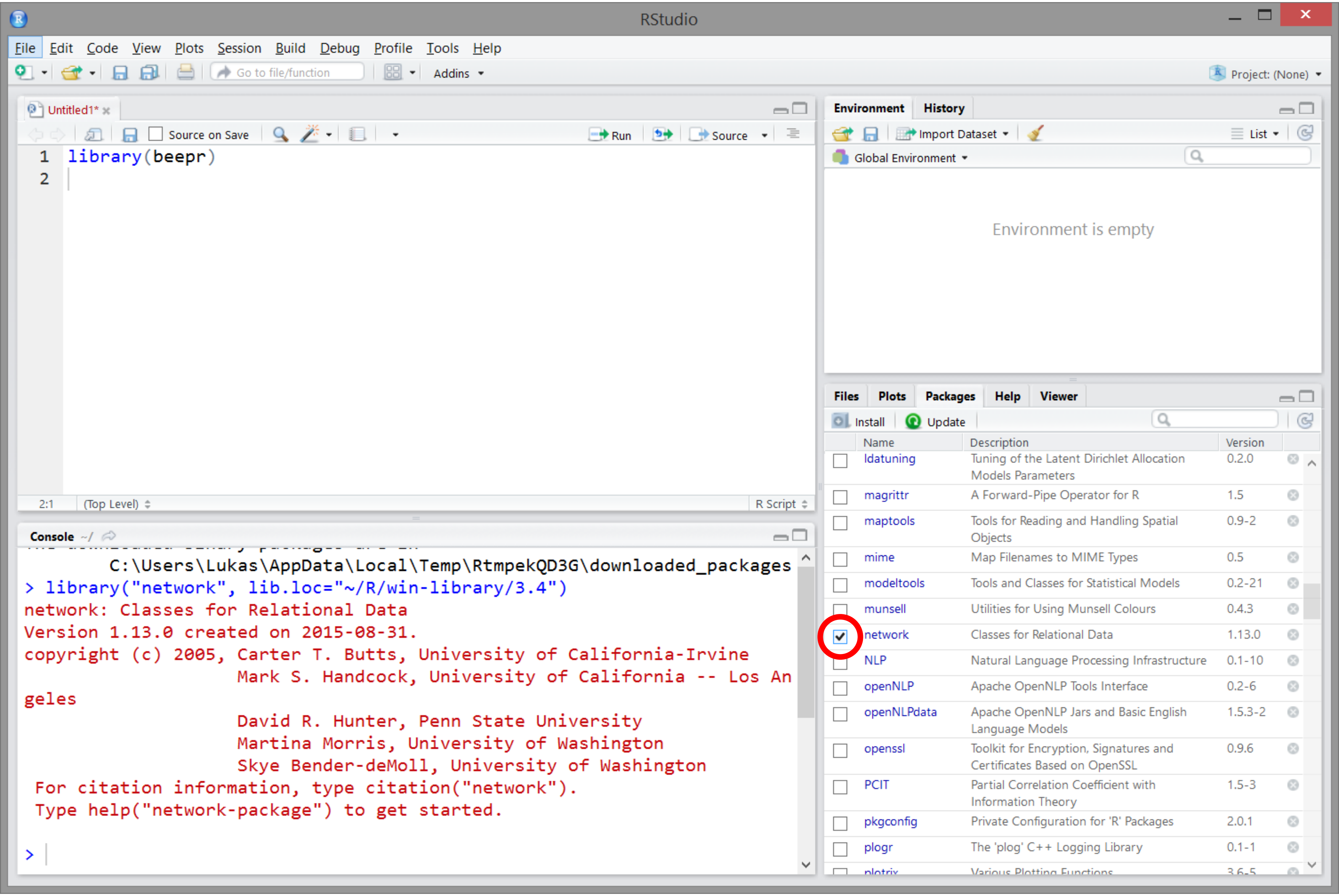
	ame	Description	Version
	rary		
	sertthat	Easy Pre and Post Assertions	0.2.0
	udio	Audio Interface for R	0.1-5
	eepr	Easily Play Notification Sounds on any Platform	1.2
	H	Boost C++ Header Files	1.62.0-1
	ndr	Parametrized Active Bindings	0.1
□	bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2
□	bitops	Bitwise Operations	1.0-6
□	Cairo	R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output	1.5-9
□	chron	Chronological Objects which can Handle Dates and Times	2.3-50
□	colorspace	Color Space Manipulation	1.3-2
□	curl	A Modern and Flexible Web Client for R	2.8.1
□	data.table	Extension of `data.frame`	1.10.4
□	dichromat	Color Schemes for Dichromats	2.0-0



```
library(beepr)
```

```
> library(beepr)
> install.packages("network")
Installing package into ‘C:/Users/Lukas/Documents/R/win-library/3.4’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.4/network_1.1
0.zip'
Content type 'application/zip' length 661853 bytes (646 KB)
downloaded 646 KB

package ‘network’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Lukas\AppData\Local\Temp\RtmpekQD3G\downloaded_packages
>
```

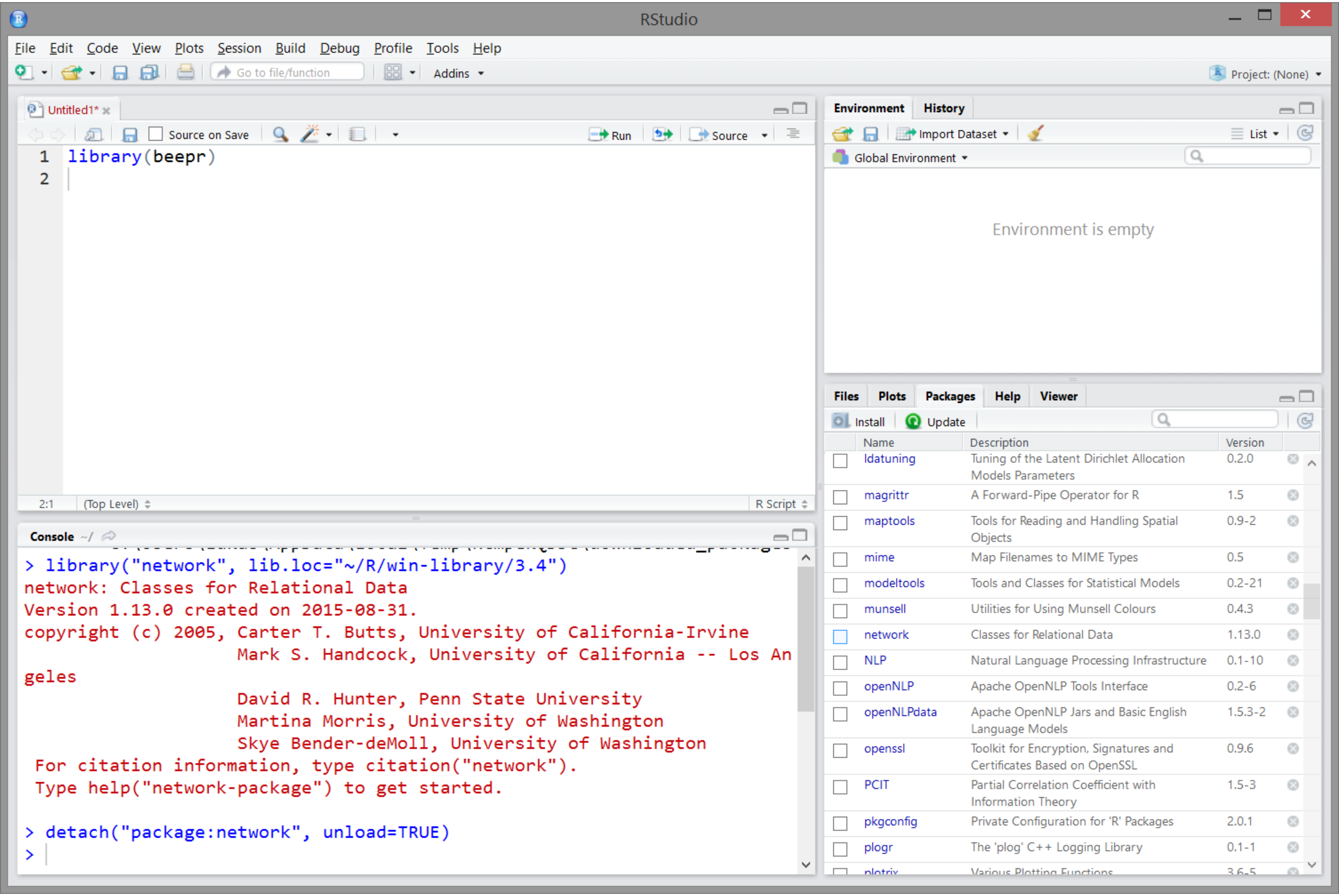
	Name	Description	Version
□	ldatuning	Tuning of the Latent Dirichlet Allocation Models Parameters	0.2.0
□	magrittr	A Forward-Pipe Operator for R	1.5
□	maptools	Tools for Reading and Handling Spatial Objects	0.9-2
□	mime	Map Filenames to MIME Types	0.5
□	modeltools	Tools and Classes for Statistical Models	0.2-21
□	munsell	Utilities for Using Munsell Colours	0.4.3
□	network	Classes for Relational Data	1.13.0
□	NLP	Natural Language Processing Infrastructure	0.1-10
□	openNLP	Apache OpenNLP Tools Interface	0.2-6
□	openNLPdata	Apache OpenNLP Jars and Basic English Language Models	1.5.3-2
□	openssl	Toolkit for Encryption, Signatures and Certificates Based on OpenSSL	0.9.6
□	PCIT	Partial Correlation Coefficient with Information Theory	1.5-3
□	pkgconfig	Private Configuration for 'R' Packages	2.0.1
□	plogr	The 'plog' C++ Logging Library	0.1-1
□	plotrix	Various Plotting Functions	3.6-5



```
library(beepr)
```

```
The downloaded binary packages are in
        C:\Users\Lukas\AppData\Local\Temp\RtmpekQD3G\downloaded_packages
> library("network", lib.loc="~/R/win-library/3.4")
network: Classes for Relational Data
Version 1.13.0 created on 2015-08-31.
copyright (c) 2005, Carter T. Butts, University of California-Irvine
                    Mark S. Handcock, University of California -- Los An
geles
                    David R. Hunter, Penn State University
                    Martina Morris, University of Washington
                    Skye Bender-deMoll, University of Washington
 For citation information, type citation("network").
 Type help("network-package") to get started.

> 
```

	Name	Description	Version
□	ldatuning	Tuning of the Latent Dirichlet Allocation Models Parameters	0.2.0
□	magrittr	A Forward-Pipe Operator for R	1.5
□	maptools	Tools for Reading and Handling Spatial Objects	0.9-2
□	mime	Map Filenames to MIME Types	0.5
□	modeltools	Tools and Classes for Statistical Models	0.2-21
□	munsell	Utilities for Using Munsell Colours	0.4.3
☑	network	Classes for Relational Data	1.13.0
□	NLP	Natural Language Processing Infrastructure	0.1-10
□	openNLP	Apache OpenNLP Tools Interface	0.2-6
□	openNLPdata	Apache OpenNLP Jars and Basic English Language Models	1.5.3-2
□	openssl	Toolkit for Encryption, Signatures and Certificates Based on OpenSSL	0.9.6
□	PCIT	Partial Correlation Coefficient with Information Theory	1.5-3
□	pkgconfig	Private Configuration for 'R' Packages	2.0.1
□	plogr	The 'plog' C++ Logging Library	0.1-1
□	plotrix	Various Plotting Functions	3.6-5



```
library(beepr)
```

Environment is empty

```
> library("network", lib.loc="~/R/win-library/3.4")
network: Classes for Relational Data
Version 1.13.0 created on 2015-08-31.
copyright (c) 2005, Carter T. Butts, University of California-Irvine
                    Mark S. Handcock, University of California -- Los An
geles
                    David R. Hunter, Penn State University
                    Martina Morris, University of Washington
                    Skye Bender-deMoll, University of Washington
 For citation information, type citation("network").
 Type help("network-package") to get started.

> detach("package:network", unload=TRUE)
> 
```

	Name	Description	Version
□	ldatuning	Tuning of the Latent Dirichlet Allocation Models Parameters	0.2.0
□	magrittr	A Forward-Pipe Operator for R	1.5
□	maptools	Tools for Reading and Handling Spatial Objects	0.9-2
□	mime	Map Filenames to MIME Types	0.5
□	modeltools	Tools and Classes for Statistical Models	0.2-21
□	munsell	Utilities for Using Munsell Colours	0.4.3
□	network	Classes for Relational Data	1.13.0
□	NLP	Natural Language Processing Infrastructure	0.1-10
□	openNLP	Apache OpenNLP Tools Interface	0.2-6
□	openNLPdata	Apache OpenNLP Jars and Basic English Language Models	1.5.3-2
□	openssl	Toolkit for Encryption, Signatures and Certificates Based on OpenSSL	0.9.6
□	PCIT	Partial Correlation Coefficient with Information Theory	1.5-3
□	pkgconfig	Private Configuration for 'R' Packages	2.0.1
□	plogr	The 'plog' C++ Logging Library	0.1-1
□	plotrix	Various Plotting Functions	3.6-5

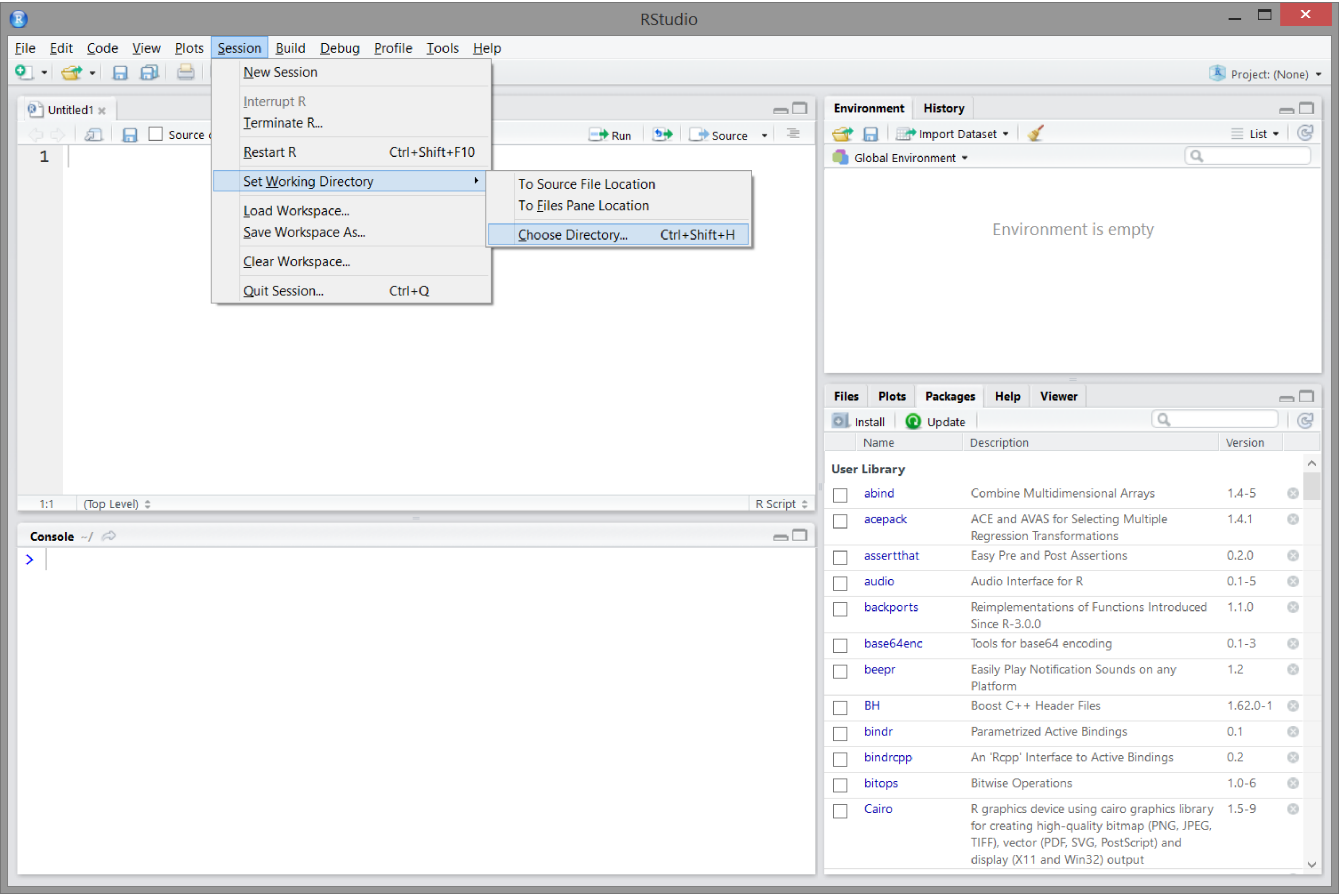
Basic R functions

```
c() # combine two or more elements into an object

class() # explore elements’ data class
length() # explore number of first dim. of object
dim() # explore dimensions of two-dimensional obj.
nrow() # number of rows
ncol() # number of columns

head() # first few rows of data
tail() # last few rows of data
str() # explore structure of object

names() # names in the named vector - one dimension
rownames() # names of rows - two dimensions
colnames() # names of columns - two dimensions
```

Working directory

- Folder, where all imports and exports are taking place – enough to set once
- Makes data import and export easier
- Functions `setwd()` and `getwd()`
- Does not accept single backslash in Win path
  - Replace backslash \ with forwardslash / or double backslash \\

```
setwd("C:\\Users\\Lukas\\Documents\\R intro")

setwd("C:/Users/Lukas/Documents/R intro")
```



Name	Description	Version
User Library		
□ abind	Combine Multidimensional Arrays	1.4-5
□ acepack	ACE and AVAS for Selecting Multiple Regression Transformations	1.4.1
□ assertthat	Easy Pre and Post Assertions	0.2.0
□ audio	Audio Interface for R	0.1-5
□ backports	Reimplementations of Functions Introduced Since R-3.0.0	1.1.0
□ base64enc	Tools for base64 encoding	0.1-3
□ beepr	Easily Play Notification Sounds on any Platform	1.2
□ BH	Boost C++ Header Files	1.62.0-1
□ bindr	Parametrized Active Bindings	0.1
□ bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2
□ bitops	Bitwise Operations	1.0-6
□ Cairo	R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output	1.5-9

Data output

- Save entire workspace
  - Save all R objects you’ve created so far
  - Allows to return to work/backup current work
- Save particular object
  - Export data to tabular objects
  - CSV as most common format

CSV - most common data format

- **C**omma-**S**eparated **V**alues
- Tabular data separated by commas (separator/delimiter) or other signs (tabulator, space, semicolon)
- CSV file (.csv), TSV file (.tsv) – always a **text file** (.txt)
- Must have **same number of columns** (separators)

```
cars,type,price,consumption,emissions,expensive
BMW,3,1200000,6.2,0,0
Audi,A4,1164000,5.9,0,0
VW,Passat,950500,6.2,NA,NA
```

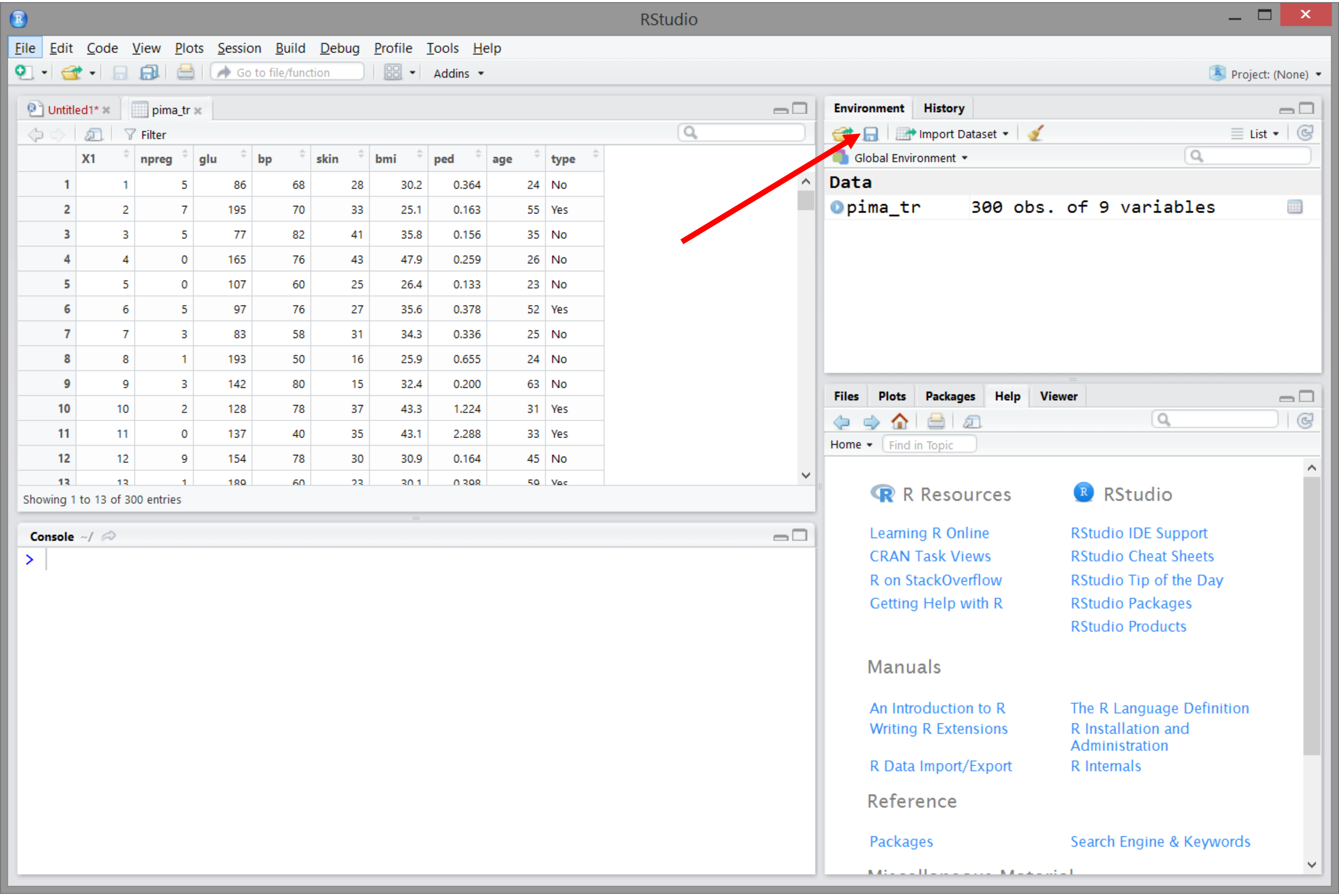
CSV – other examples

```
cars;type;price;consumption;emissions
BMW;3;1200000;6.2;0
Audi;A4;1164000;5.9;0
VW;Passat;950500;6.2;0
```

```
"cars" "type" "price" "consumption"  "emissions"
"BMW" "3" "1,200,000" "6.2" "0"
"Audi" "A4" "1,164,000" "5.9" "0"
"VW" "Passat" "950,500" "6.2" "0"
```

```
cars,type,price,consumption,emissions
BMW,3,1,200,000,6.2
Audi,A4,1,164,000,5.9
VW,Passat,950,500,6.2
```

Bad data – improper use of comma delimiter results in uneven # of rows



	X1	npreg	glu	bp	skin	bmi	ped	age	type
1	1	5	86	68	28	30.2	0.364	24	No
2	2	7	195	70	33	25.1	0.163	55	Yes
3	3	5	77	82	41	35.8	0.156	35	No
4	4	0	165	76	43	47.9	0.259	26	No
5	5	0	107	60	25	26.4	0.133	23	No
6	6	5	97	76	27	35.6	0.378	52	Yes
7	7	3	83	58	31	34.3	0.336	25	No
8	8	1	193	50	16	25.9	0.655	24	No
9	9	3	142	80	15	32.4	0.200	63	No
10	10	2	128	78	37	43.3	1.224	31	Yes
11	11	0	137	40	35	43.1	2.288	33	Yes
12	12	9	154	78	30	30.9	0.164	45	No
13	13	1	189	60	23	30.1	0.398	59	Yes

Showing 1 to 13 of 300 entries

```
>
```

Data

pima_tr 300 obs. of 9 variables

R Resources

Learning R Online
CRAN Task Views
R on StackOverflow
Getting Help with R

RStudio

RStudio IDE Support
RStudio Cheat Sheets
RStudio Tip of the Day
RStudio Packages
RStudio Products

Manuals

An Introduction to R
Writing R Extensions
R Data Import/Export
The R Language Definition
R Installation and Administration
R Internals

Reference

Packages
Search Engine & Keywords

Miscellaneous Material

Exporting object – tabular

- Function `write.table()`
- Name of file must be specified
- Easy to import to Excel or other software

```r
frequencies <- c(92,89,94,72,59)

write.table(frequencies,
            "frequencies.csv",
            sep = ",",
            row.names = FALSE,
            col.names = TRUE,
            fileEncoding = "UTF-8")
```

Exporting object – unstructured

- Function `writeLines()`
- Has basically no arguments
- Saves the whole object as one text

```
frequencies <- c(92,89,94,72,59)

writeLines(frequencies,
           "frequencies.txt")
```

Text analysis in R

Text analysis in R

- Most prominent package for text analysis is “tm” (stands for text mining)
  - Provides tools corpus creation, text manipulation, term-document matrix creation
  - Easily allows to read text documents as corpus
- Competing packages – “quanteda”
  - Developed by Ken Benoit (WordScores)
  - Provides some TA methods
  - Overlaps with “tm” package – if both packages loaded, it will generate conflicts (feature, not bug)

Corpus

- `getSources()` provides list of available sources
  - Files inside a directory – `DirSource()`
  - Text inside a vector – `VectorSource()`
  - Dataframe, XML, links to web-sites, …
- `Corpus()` creates a corpus object out of text sources

Corpus

```
my.texts <-  "C:\\Users\\Lukas\\Desktop\\data\\"

directory.source <- DirSource(directory = my.texts)

text.corpus <- Corpus(directory.source)
```

Corpus operations – functions

- Useful functions:
  - `removePunctuation()` – remove all punctuation
  - `removeWords()` – remove stopwords
  - `stripWhitespace()` – remove duplicate white space
  - `removeNumbers()` – remove all numbers
  - `stemDocument()` – stem document
  - `plainTextDocument()` – turn document into tm package's plain text format

Corpus operations

- `tm_map()` function allows to apply manipulations over the corpus data

```
edited.corpus <- text.corpus

edited.corpus <- tm_map(edited.corpus, removeNumbers)

edited.corpus <- tm_map(edited.corpus, removePunctuation)

edited.corpus <- tm_map(edited.corpus, stripWhitespace)

edited.corpus <- tm_map(edited.corpus,
                        removeWords,
                        stopwords("english"))
```

Term-document matrix

- Function `TermDocumentMatrix()`
  - Terms in rows
  - Documents in columns
- `DocumentTermMatrix()` creates inverse TDM
- Output is non-standard matrix object
  - If matrix operations are needed, it must be converted to basic matrix format with `as.matrix()` function

Term-document matrix

```
tdm <- TermDocumentMatrix(edited.corpus)

dtm <- DocumentTermMatrix(edited.corpus)

tdm.matrixed <- as.matrix(tdm)
```

Useful functions in “tm”

- `removeSparseTerms()`
  - Removes terms to a defined sparsity of the TDM matrix – removes terms which are used sparsely across documents
- `findFreqTerms()`
  - Lists most frequent terms across the TDM matrix
  - Does **not provide frequencies**, though
- `findAssocs()`
  - Correlation of appearance of a term with other terms across TDM – returns **Pearson’s r**

Frequencies

- `findFreqTerms()` shows frequent terms
  - Has two attributes defining bounds – `lowfreq`, `highfreq`
- Easier to calculate frequencies separately
  - Convert TDM to matrix with `as.matrix()`
  - Calculate sums of rows with `rowSums()`
  - Sort the vector with `sort()` with `decreasing` attribute

```
tdm.matrixed <- as.matrix(tdm)

frequencies <- rowSums(tdm.matrixed)

frequencies <- sort(frequencies, decreasing = T)
```

Wordclouds

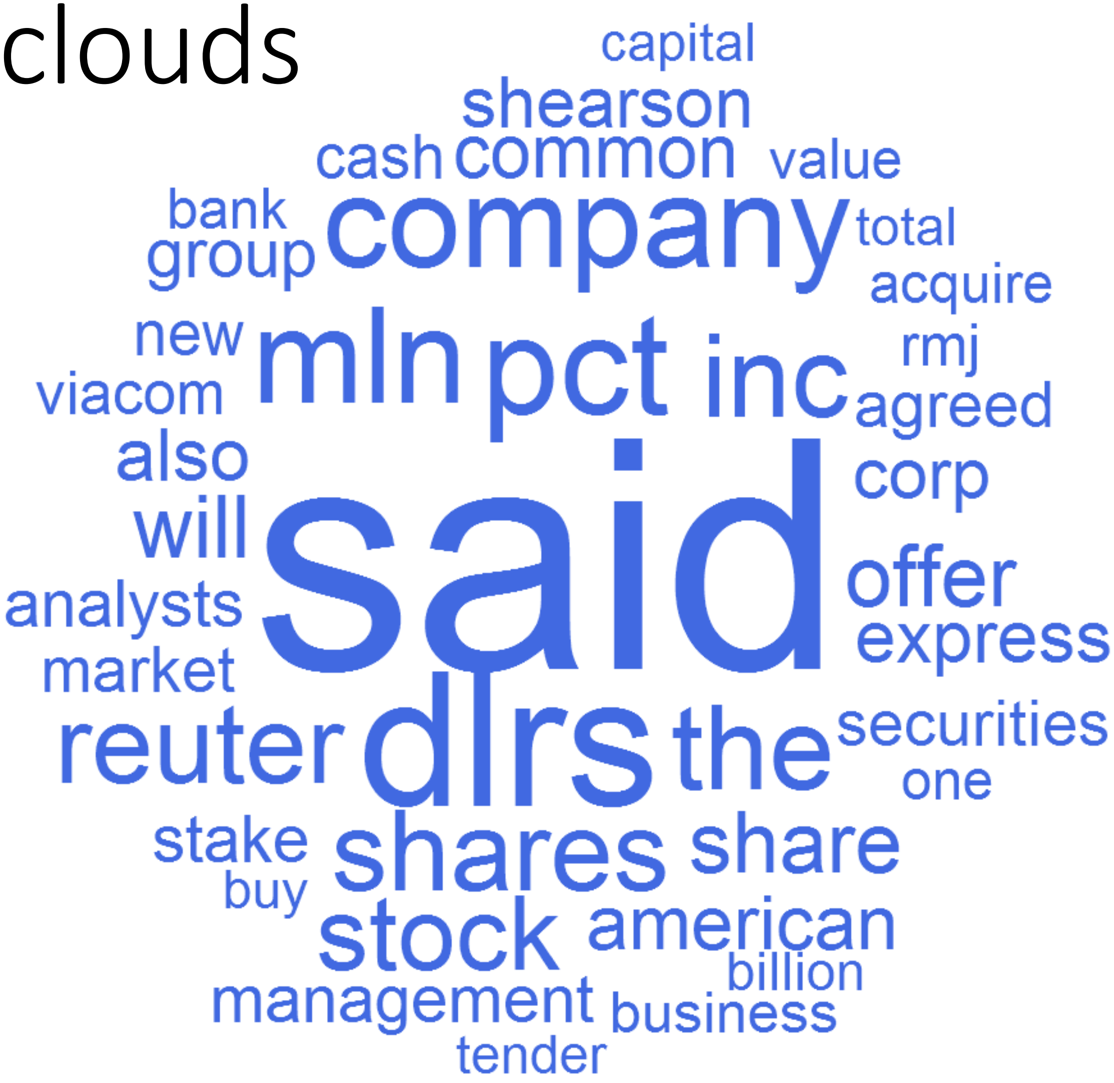
- Package “wordcloud”
- Function `wordcloud()`

Attribute	Description
`words`	Terms
`freq`	Frequencies of terms
`scale`	Two values in `c()` function to bound upper and lower scale
`max.words`	Maximum number of words rendered
`random.order`	Binary - should terms be placed in random order?
`rot.per`	Percentage of terms placed vertically
`colors`	Color or color palette
`random.color`	Binary – should colors be assigned randomly or based on the word frequency?

Wordclouds

```
tdm.matrixed <- as.matrix(tdm)

frequencies <- rowSums(tdm.matrixed)

frequencies <- sort(frequencies,decreasing = T)

terms <- names(frequencies)

library(wordcloud)

wordcloud(words = terms,
        freq = frequencies,
        scale = c(5,0.5),
        max.words = 150,
        random.order = F,
        rot.per = 0,
        colors = "red")
```

Wordclouds


capital
shearson
cash
common
value
bank
company
total
group
acquire
new
mln
pct
inc
rmj
viacom
agreed
also
corp
will
said
offer
analysts
express
market
reuter
dlrs
the
securities
one
stake
shares
share
buy
stock
american
billion
management
business
tender