LUNDS UNIVERSITET
Lunds Tekniska Högskola

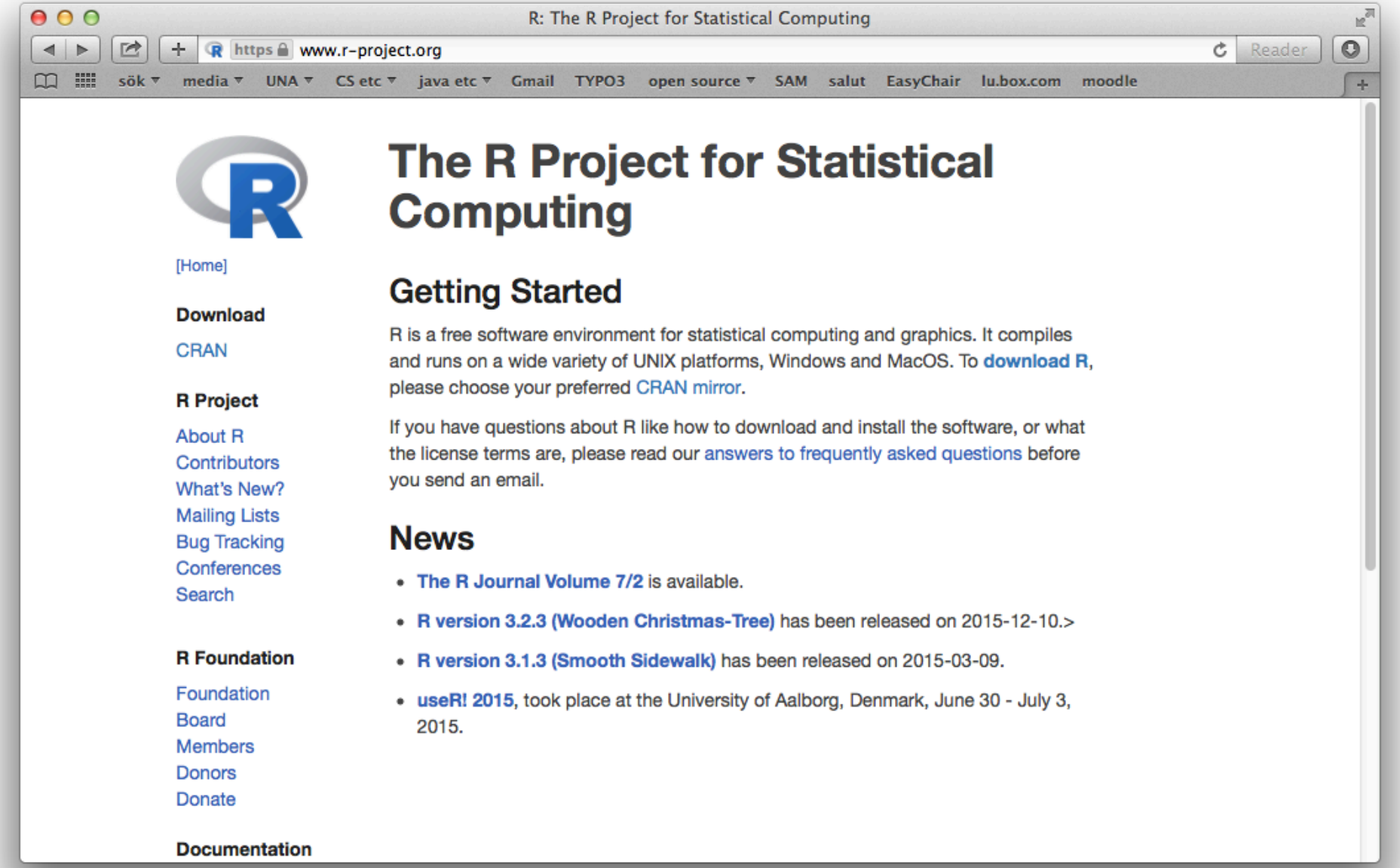
EDAA35: R-programmering

MARTIN HÖST

Idag

- Om R
- De vanligaste datatyperna
- Urval av data
- Analys av data
- Presentation av resultat: plot, utskrift, etc

www.r-project.org



Vad är R?

- En variant av S
  - Högnivåspråk för dataanalys och grafik utvecklat av AT&T Bell Laboratories. Fick ”ACM Software System Award” 1998
- Open source version (GNU GPL), mycket väl använd
  - ”free”
  - Mycket hjälp finns t ex på nätet
- ”Statistikverktyg” och programspråk

Typisk användning

- Inläsning av data
- Sortering, urval av data, etc
- Statistisk analys
  - Deskriptiv statistik
  - Mer avancerade funktioner
- Presentation av resultat
  - Skärm, fil, grafik,…
  - Spara data
- Många paket för olika funktioner

Library Highlights (Pandas)

- A fast and efficient **DataFrame** object for data manipulation with integrated indexing;
- Tools for **reading and writing data** between in-memory data structures and different formats: CSV and text files, Microsoft Excel, SQL databases, and the fast HDF5 format;
- Intelligent **data alignment** and integrated handling of **missing data**: gain automatic label-based alignment in computations and easily manipulate messy data into an orderly form;
- Flexible **reshaping** and pivoting of data sets;
- Intelligent label-based **slicing**, **fancy indexing**, and **subsetting** of large data sets;
- Columns can be inserted and deleted from data structures for **size mutability**;
- Aggregating or transforming data with a powerful **group by** engine allowing split-apply-combine operations on data sets;
- High performance **merging and joining** of data sets;
- **Hierarchical axis indexing** provides an intuitive way of working with high-dimensional data in a lower-dimensional data structure;
- **Time series**-functionality: date range generation and frequency conversion, moving window statistics, moving window linear regressions, date shifting and lagging. Even create domain-specific time offsets and join time series without losing data;
- Highly **optimized for performance**, with critical code paths written in Cython or C.
- Python with *pandas* is in use in a wide variety of **academic and commercial** domains, including Finance, Neuroscience, Economics, Statistics, Advertising, Web Analytics, and more.

https://pandas.pydata.org

Programspråket

- Dynamiskt typat
- Script-språk, körs från terminalen utan kompilering

Lab 1

- Görs som quiz i moodle
- Man måste klara alla frågorna
- Jobba i par på en inloggning i moodle. Godkännande inte i moodle (i ”SAM”)
- Ibland möjligt att ”gissa” svar… Gör inte det, ni har nytta av att förstå på nästa lab…

- ”enrolment key”: EDAA3520
  - OBS! Välj rätt kursinstans (Utvärdering av programvarusystem 2020)

Några alternativ till R

- Matlab – används av många inom elektro, signalbehandling, etc. Mycket bra på t ex matriser
- Python – generellt språk. Med ”numpy” och ”pandas” får man ungefär samma funktioner som i R (t ex ”dataramar”)

- …men har man förstått hur man jobbar i ett av språken kan man enkelt lära sig ett annat

- Visa i terminalen

Operationer

`:: :::`	åtkomst av variabler i 'namespace' *
`$ @`	komponent, @*
`[ [[`	indexering
`^`	exponent
`- +`	unär plus och minus
`:`	sekvens
`%any%`	special-operator *
`* /`	multiplikation, division
`+ -`	binär addition och subtraktion
`< > <= >= == !=`	jämförelser
`!`	negation
`& &&`	och
`| ||`	eller
`~`	för att beskriva modeller *
`-> ->>`	tilldelning, tilldelning i 'enclosing environment'*
`<- <<-`	tilldelning, tilldelning i 'enclosing environment'*
`=`	tilldelning (motsvarar '<-')
`?`	hjälp

Några funktioner för beräkningar

`sum(..., na.rm = FALSE)`	summa
`prod(..., na.rm = FALSE)`	produkt
`cumsum(x)`, `cumprod(x)`	kumulativ summa, produkt
`mean(x, trim = 0, na.rm = FALSE, ...)`	medelvärde
`median(x, na.rm = FALSE)`	median
`var(x, na.rm = FALSE)`	varians
`sd(x, na.rm = FALSE)`	standardavvikelse
`exp(x)`	e^x
`log(x, base = exp(1))`, `log10(x)`, `log2(x)`	logaritmer
`max(..., na.rm = FALSE)`	max
`min(..., na.rm = FALSE)`	minimum
`range(..., na.rm = FALSE)`	'range', dvs min och max
`which.max(x)`, `which.min(x)`	vilket värde som är max, min
`sqrt(x)`, `abs(x)`	$\sqrt{x}$, $\|x\|$
`sin(x)`, `cos(x)`, `tan(x)`	$sin()$, $cos()$, $tan()$ (radianer)
`round(x, digits = 0)`	avrundning (`digits` decimaler)
`floor(x)`, `ceiling(x)`	avrundning nedåt, uppåt
`factorial(x)`, `choose(n, k)`	$x!$, $n!/((n-k)!k!)$

Arbeta med dataobjekt

`nrow(a)`	antal rader i `a`
`ncol(a)`	antal kolumner i `a`
`rbind(a, b)`	sammanslagning där raderna i `b` kommer efter raderna i `a`, dvs en radvis sammanslagning.
`cbind(a, b)`	sammanslagning där kolumnerna i `b` kommer efter kolumnerna i `a`, dvs en kolumnvis sammanslagning.
`summary(a)`	sammanfattar innehållet i `a`, t ex medelvärde och median
`str(a)`	sammanfattar `a`:s struktur
`head(a)`	de första raderna i `a`
`tail(a)`	de sista raderna i `a`
`sort(a)`	sortera `a`
`order(a)`	index som sorterar vektor `a`, dvs `a[order(a)]` motsvarar `sort(a)`
`table(a)`	räknar hur många element i `a` det finns av varje värde
`unlist(x)`	omvandlar listan `x` till en vektor
`append(x, values, after = length(x))`	lägger in vektorn `values` i vektorn `x` efter position `after` (om `after==0` så hamnar `values` först)

Läsa data från fil

- read.csv(file, header = TRUE, sep = ",", quote = "\"", dec = ".", fill = TRUE, comment.char = "", ...)
  - För komma-separerade data-filer
- read.table()
  - Mer generell

+ många fler paket och funktioner för att importera data

Instruktioner i script

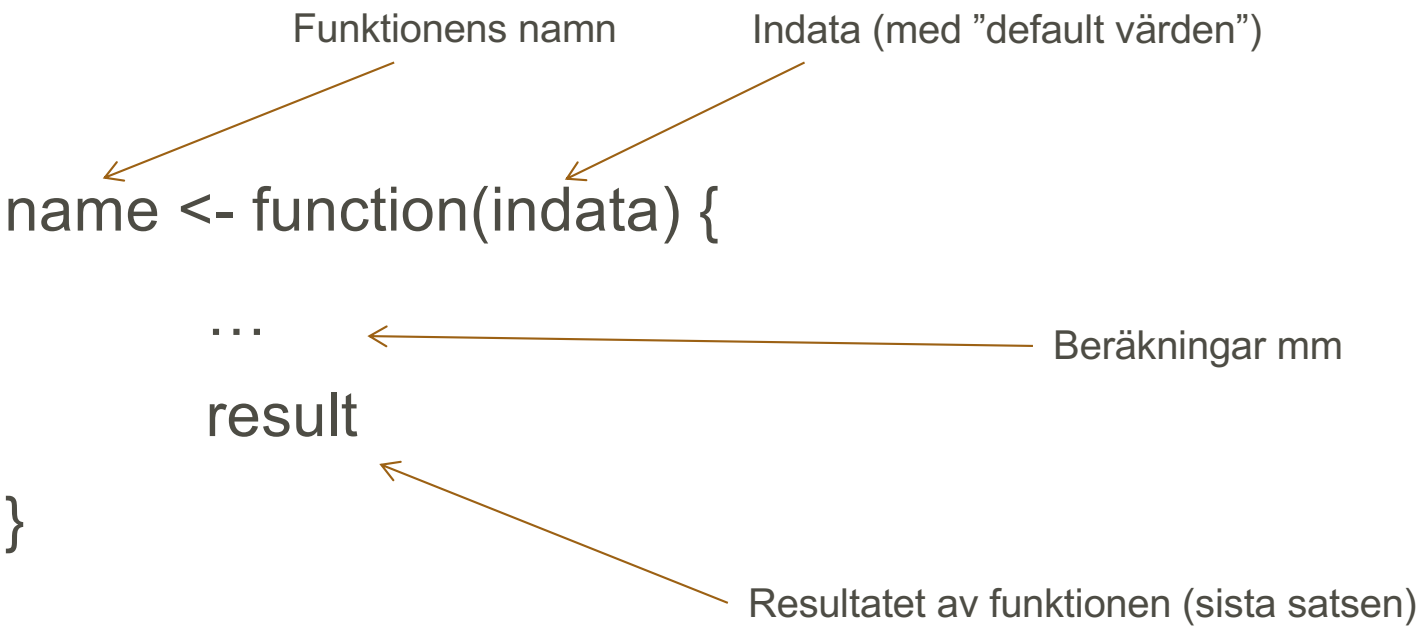
- Körs med
  > source(”filnamn”)

- T ex:

```
> source("script.R")
[1] "hej"
```

```
a <- "hej"
print(a)
```

Funktioner, definition



(Kan vara anonyma)

- Visa i terminalen