Tidy and Seamless in R

- 2020-10-14
- 유충현, Tidyverse Korea

CONTENTS

1. R History – include S and S+

Bell Lab, in AT&T (1976)

- John Chambers, Ricard Becker, Allan Wilks

StatSci (1988)

- Statistical Sciences Inc
- Start-Up, in Seattle

MathSoft (1993)

- StatSci, S 독점 라이센스 획득
- MathSoft 합병

Insightful (2001)

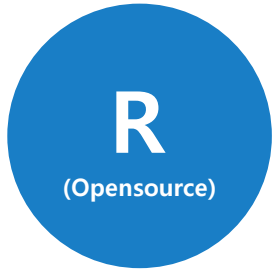
- MathSoft 합병 취소
- 사명변경
- packages on CSAN[1]

Lucent Technologies (2004)

- MathSoft로부터 라이선스 구입
- 구 AT&T

TIBCO (2008)

- Spotfire



University of Auckland (1992)

- Ross Ihaka, Robert Gentleman

R on Statlib (1993)

- http://lib.stat.cmu.edu

R Development Core Team (1997)

- The GNU Project of R Foundation
- R on CRAN[2]
- Version 0.49

Initial (1995)

- last alpha

0.16 (1997)

- White Book

1.0.0 (2000)

- S3

1.4.0 (2001)

- S4

1.5.0 (2002)

- MASS

2.0.0 (2004)

- Lazy loading

3.0.0 (2013)

- numeric index values 2^{31}

4.0.0 (2020)

- stringsAsFactors = FALSE

https://en.wikipedia.org/wiki/S_(programming_language)

https://en.wikipedia.org/wiki/S-PLUS

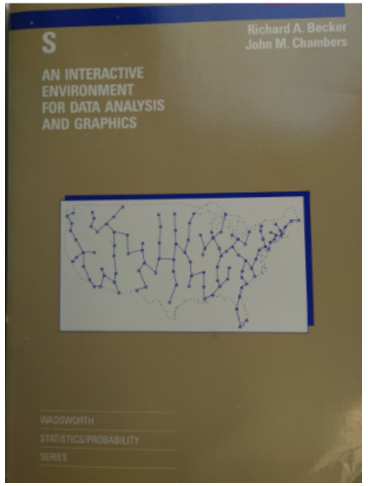
https://en.wikipedia.org/wiki/R_(programming_language)

1 : CSAN(The Comprehensive S Archive Network)

2 : CRAN (The Comprehensive R Archive Network)

S : An Interactive Environment for Data Analysis and Graphics

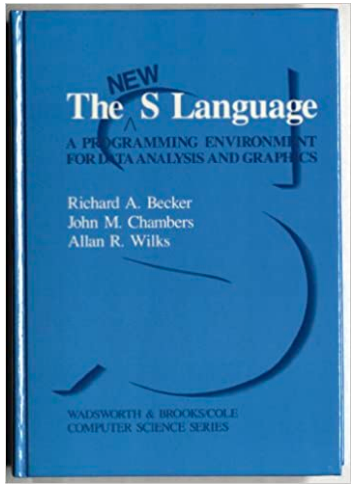
(Ricard Becker, John Chambers)



1984 Brown Book

The New S Language

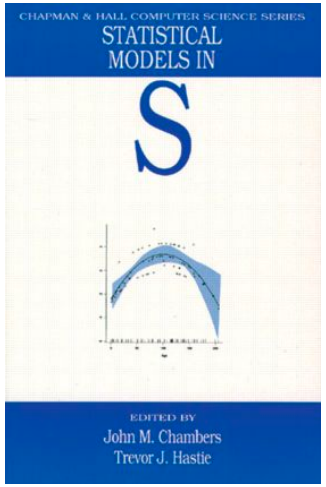
(Ricard Becker, John Chambers, Allan Wilks)



1988 Blue Book

Statistical Models in S

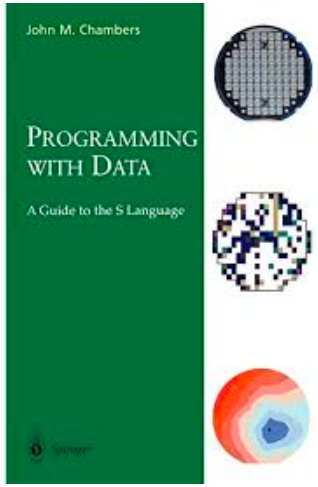
(John Chambers, Trevor Hastie)



1992 White Book

Programming with Data

(John Chambers)



1998 Green Book

R6: Encapsulated object-oriented programming for R

S2

start S3

- macros to functions
- 객체 컨셉
- Graphics device
  - X11, Postscript
- FORTRAN to C

completed S3

- macros to functions
- data frame 객체 정의
- S3 Class (OOP)
  - Method dispatch
  - Generic function

S4

- formal S4 Class
  - methods 패키지
- Connections

R5

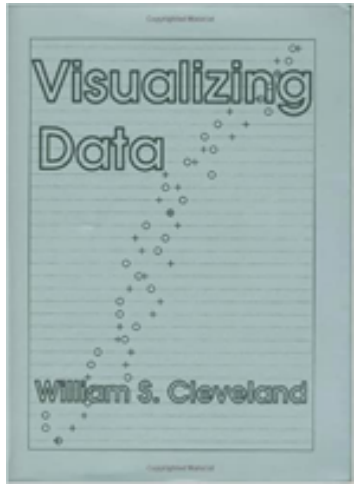
- R5 Class
  - reference class
  - using S4

R6

- R6 Class
  - public and private method
  - active binding

Visualizing Data

(William Cleveland)

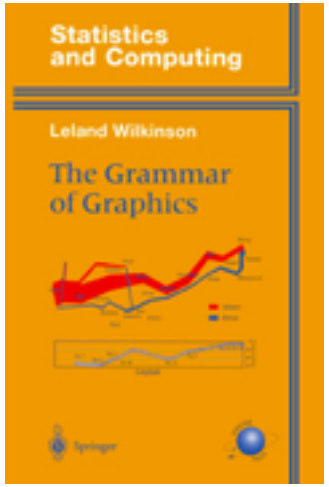


1993

trellis package in S

The Grammar of Graphics

(Leland Wilkinson)

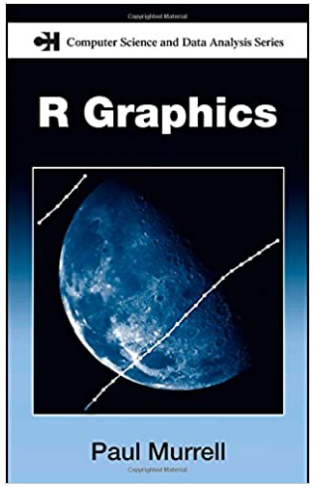


1999

Grammar Concept

R Graphics

(Paul Murrell)

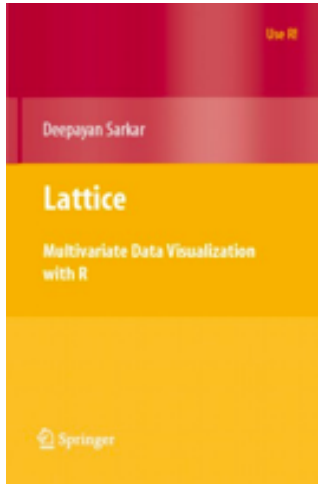


2005

grid package

Lattice

(Deepayan Sarkar)



2008

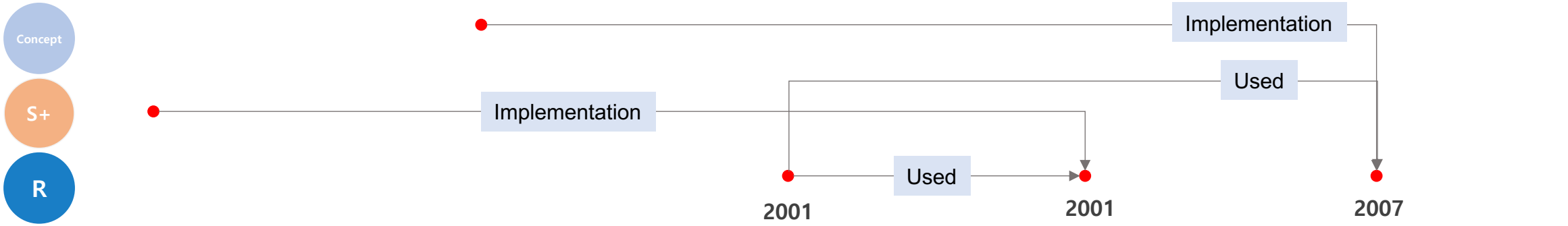
lattice package

ggplot2

(Hadley Wickham)

2009

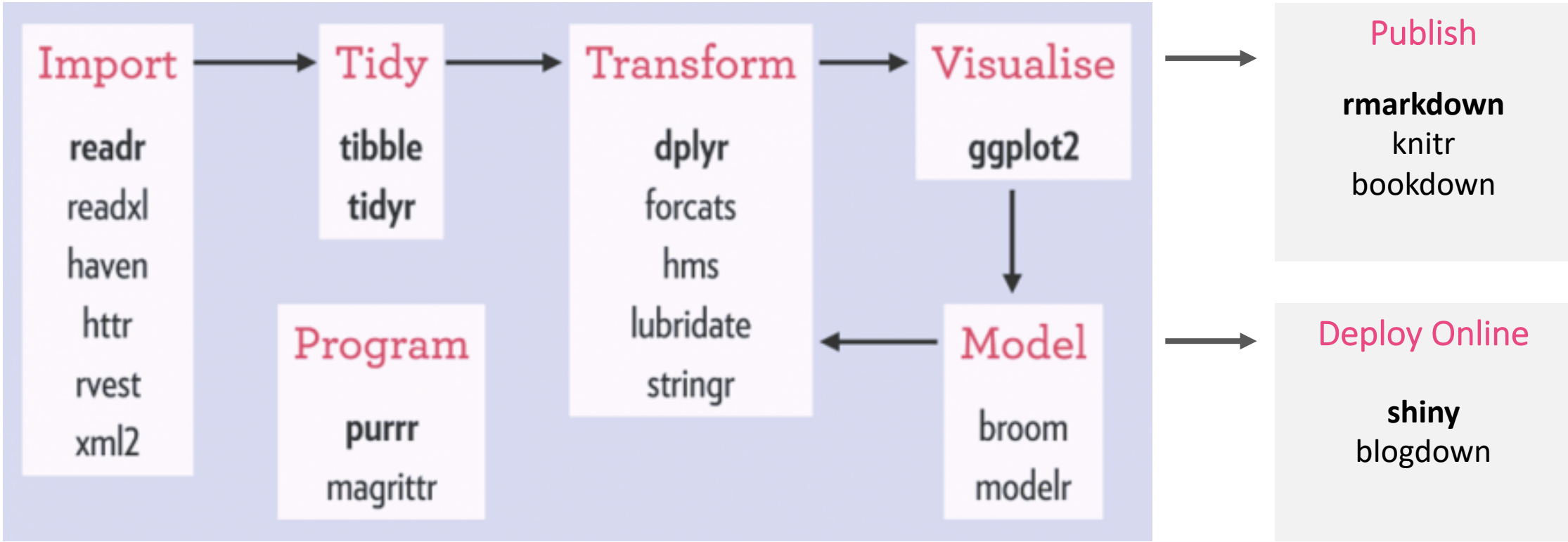
ggplot2 package



Data Scale	Package Developer	Distribution	Analytics Methods
토이 데이터 수준의 데이터 분석	통계학자 중심	상업용 통계 프로그램	연결성이 약하고 분석에만 집중된 방법론
엔터프라이즈급의 대용량 데이터의 분석	소프트웨어 개발자의 합류	오픈소스 기반의 데이터분석 에코시스템	유기적으로 연결된 Closed Loop 분석 코드 및 재사용성 제고

	Before	After
Data Scale	Toy Data data.frame read.* In-Memory	Enterprise Data tibble read_* In-Database
Package Developer	Statistician R Development Core Team Statistics	S/W Developer Hadley 사단 Machine Learning
Distribution	Commercial S-Plus (CSAN)	Open Source R (CRAN, GitHub)
Analytics Methods	R Graphics, lattice each packages	ggplot2 tidyverse, tidymodel

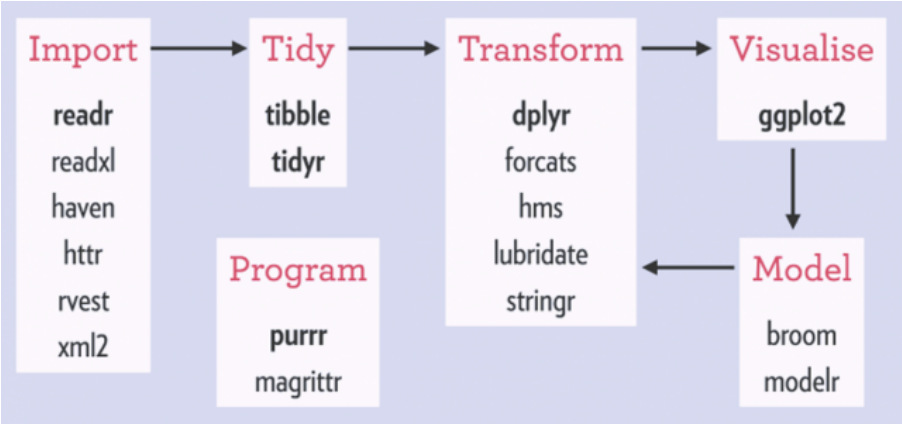
Seamless with tidyverse



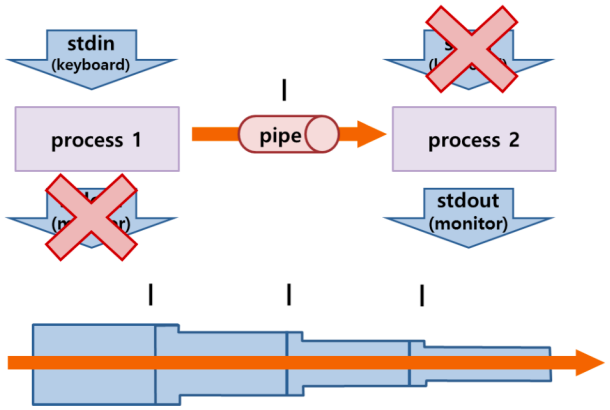
https://www.kdnuggets.com/2017/10/tidyverse-powerful-r-toolbox.html

Within Seamless

From import to model using tidyverse



Shell script using pipe



Pipe using magrittr package

Between Seamless

Rmarkdown을 이용한 Seamless 프로그래밍

- 시나리오 : 파일의 용량 증가에 따른 전통적인 read.csv()와 readr 패키지의 read_csv(), Python pandas read_csv()의 성능 비교
- 소스 : simulation_import.Rmd

1. R Chunk

- Seed data 생성

2. Bash Chunk

- 복수배 파일 생성

3. R Chunk

- 성능측정
- read.csv(), read_csv()

4. Python Chunk

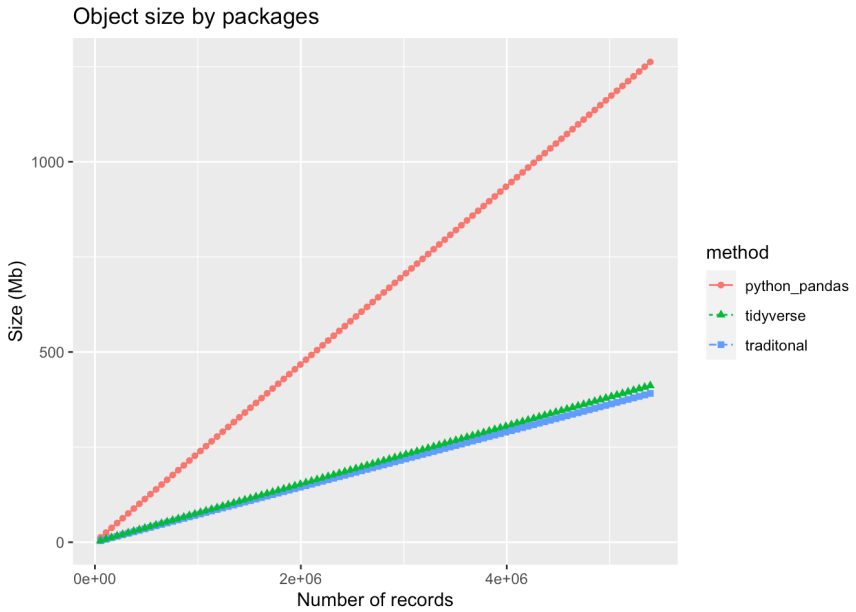
- 성능측정
- pandas's read_csv()

5. R Chunk

- 성능 비교
- 시각화

6.Bash Chunk

- 파일 삭제



reticulate package로 Python 연동하기

- 시나리오 : R에서 생성한 대통령 연설문 Term Frequency 데이터를 Python에서 word cloud 그리기
- 소스 : call_python_from_R.R, speech.rda

1. R

- 데이터 읽어 들이기
- Term Frequency 벡터 조작

2. Python 호출

- R 벡터를 python의 dictionary로 변환
- word cloud 그리기

김대중 대통령

노무현 대통령

이명박 대통령

Shell script에서 R 연동하기

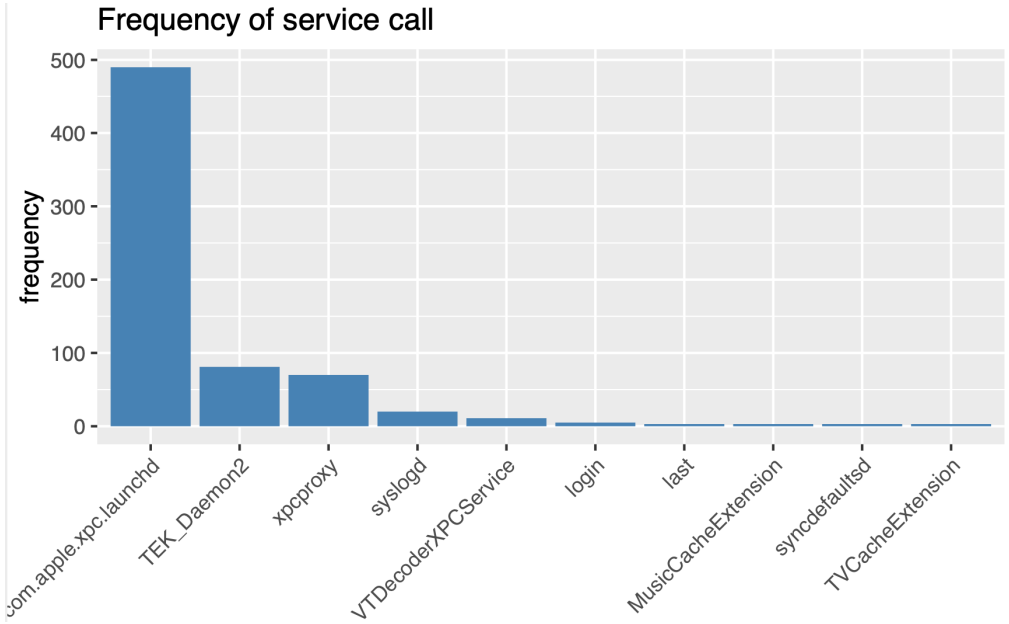
- 시나리오 : /var/log/system.log 파일에서 서비스별 호출 회수를 구하고, 이를 시각화
- 소스 : get_summary_system.sh

1. Shell Script

- pipe를 이용해서 서비스별 호출 회수 계산
- 결과를 파일에 저장

2. R 호출

- 파일을 읽어 data frame 생성
- ggplot2로 bar chart 그리기



E. O. D