Open Source

from Statistical Computing to (Generative) AI

Tobias Verbeke (Open Analytics)
September 27, 2024

Open Source

Disclaimer

Every single word I say, represents the official position of Open Analytics...

‘hacker culture’

In the beginning there were...

‘hacker culture’

… Real Programmers.

- thick glasses, white socks, read science-fiction
- played with technology and invented programming tools for fun
- **1961** MIT, Stanford, Carnegie-Mellon (AI and computer science research)
- **1969** ARPAnet: cross-net collaborations across US
- rise of UNIX (1973) and C programming language (1972)
- e.g. **Jargon File**, later published as Hacker’s Dictionary

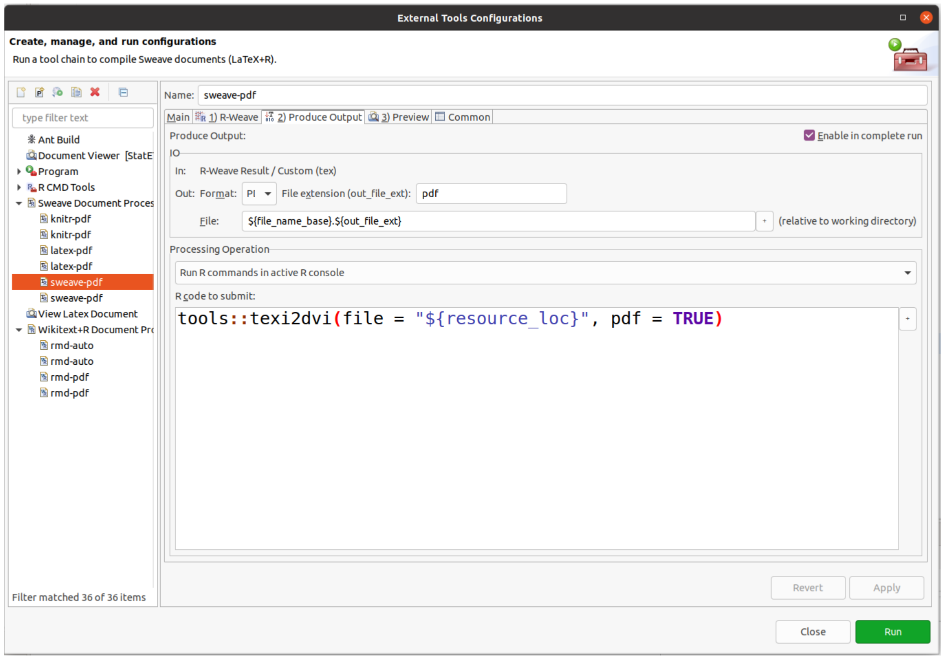
1979

- Richard Stallman joins the AI research lab at MIT
- BSc in Physics, starts work as an operating system programmer

- **time bombs** in markup language (US Copyright Act of 1976 – software becomes copyrightable)
- Stallman wrote his own: Texinfo

Texinfo

```
texi2dvi --pdf
```


External Tools Configurations

Create, manage, and run configurations

Run a tool chain to compile Sweave documents (LaTeX+R).

Name: sweave-pdf

Produce Output:

☑ Enable in complete run

IO

In: R-Weave Result / Custom (tex)

Out: Format: PI File extension (out_file_ext): pdf

File: ${file_name_base}.${out_file_ext} (relative to working directory)

Processing Operation

Run R commands in active R console

R code to submit:

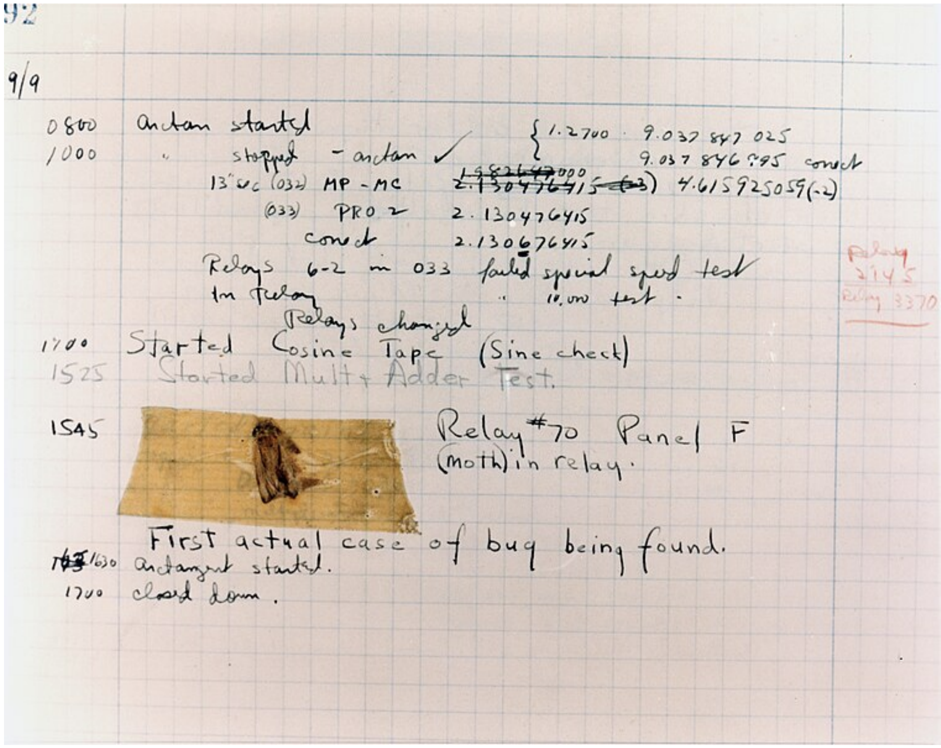
```r
tools::texi2dvi(file = "${resource_loc}", pdf = TRUE)
```

Filter matched 36 of 36 items

RMS called it a ‘crime against humanity’

1980

- new laser **printer**, access to source code refused



92

9/9

0800 arctan started

1000 " stopped - arctan ✓

13" sec (032) MP - MC ~~1.2982647000~~ ~~2.130476415 (-3)~~) 4.615925059(-2)

(033) PRO 2 2.130476415

correct 2.130676415

{1.2700 9.037 847 025

9.037 846 995 correct

Relays 6-2 in 033 failed special speed test

In relay " 10,000 test .

Relays changed

Relay 2145

Relay 3370

1100 Started Cosine Tape (Sine check)

1525 Started Mult + Adder Test.

1545 Relay #70 Panel F (moth) in relay.

First actual case of bug being found.

~~143~~ 1630 antangent started.

1700 closed down.

- previous printer: modified source: message when printer jam, message when job done
- software / **recipe** analogy
- scientific collaboration vs. rise of proprietary software, proprietary UNIX

1983

- GNU project in september 1983
- GNU: Gnu's not unix – recursive acronym (cf. hacker culture)
- UNIX philosophy:
    - small is beautiful
    - make each program do one thing well
    - make every program a filter etc.

```
head /var/log/syslog | awk '{print $4}' | sort | uniq -c | sort -rn
```

GNU implemented one tool at a time

1985

- FSF – https://www.fsf.org/
- founded in 1985 by Richard Stallman (‘RMS’, Saint IGNUcius)
- worked a.o. on (Church of) emacs
- goal: promote the ideals of free software

free as in **free beer** vs. free as in **free speech**

FREE SOFTWARE
FOUNDATION

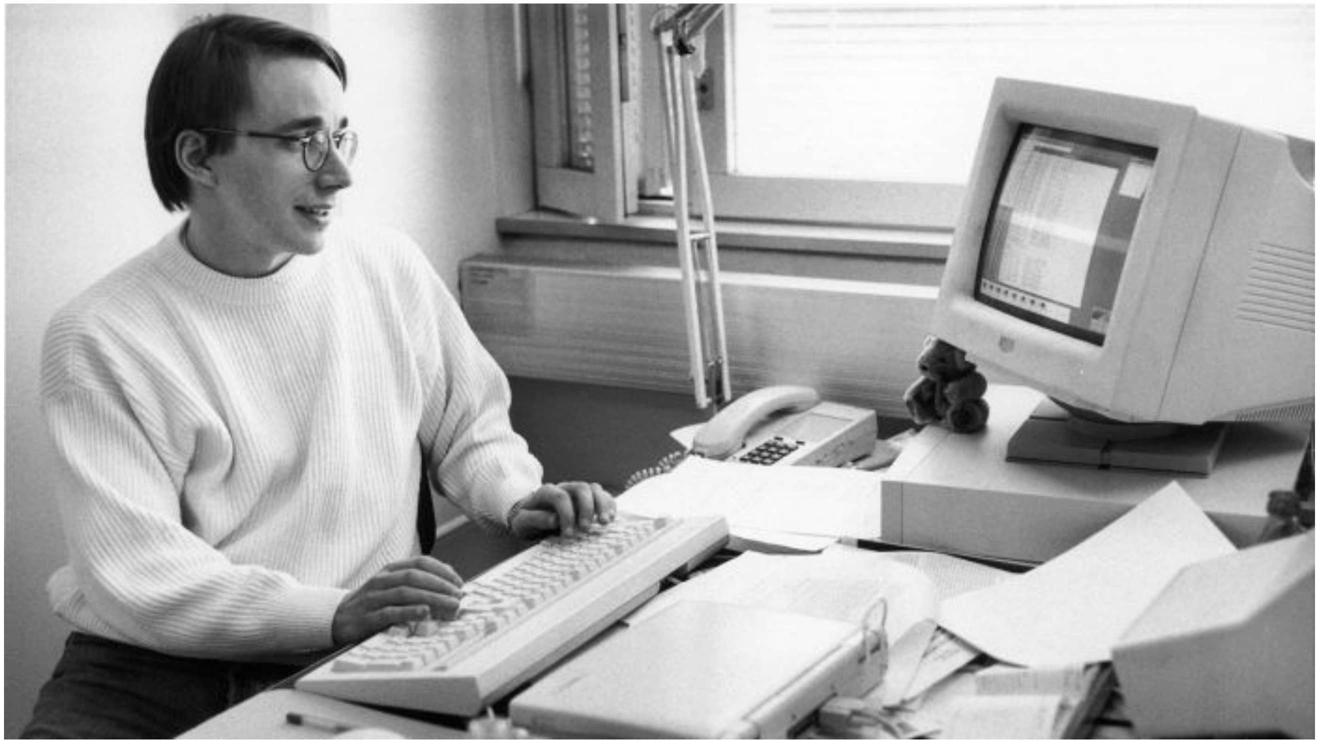
Fundamental Freedoms

formalized his experience (1986)

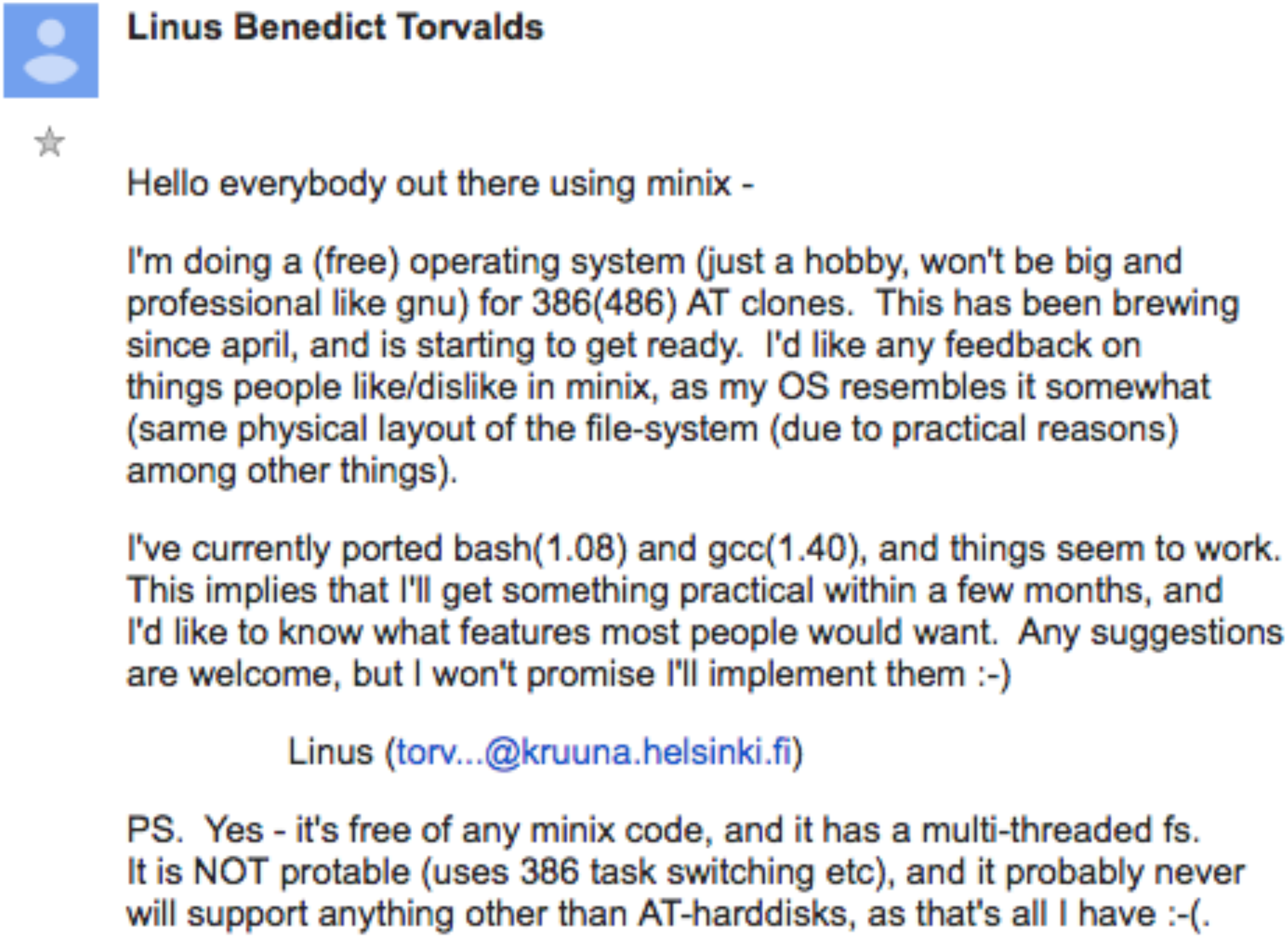
- The freedom to **run** the program as you wish, for any purpose (freedom 0).
- The freedom to **study** how the program works, and **change** it so it does your computing as you wish (freedom 1). Access to the source code is a precondition for this.
- The freedom to **redistribute** copies so you can help your neighbor (freedom 2).
- The freedom to **distribute** copies of your **modified** versions to others (freedom 3). By doing this you can give the whole community a chance to benefit from your changes. Access to the source code is a precondition for this.

Linux

- Linus Torvalds, the guy who created **Git**
- Swedish speaking minority in Finland
- classic geek in childhood
- 2nd year of computer science at university: gets a course on C programming and on UNIX
- "I want this at home too"
- did not have any C nor UNIX experience, but started work



Linux – original usenet post on `os.unix.minix`



Linus Benedict Torvalds

Hello everybody out there using minix -

I'm doing a (free) operating system (just a hobby, won't be big and professional like gnu) for 386(486) AT clones. This has been brewing since april, and is starting to get ready. I'd like any feedback on things people like/dislike in minix, as my OS resembles it somewhat (same physical layout of the file-system (due to practical reasons) among other things).

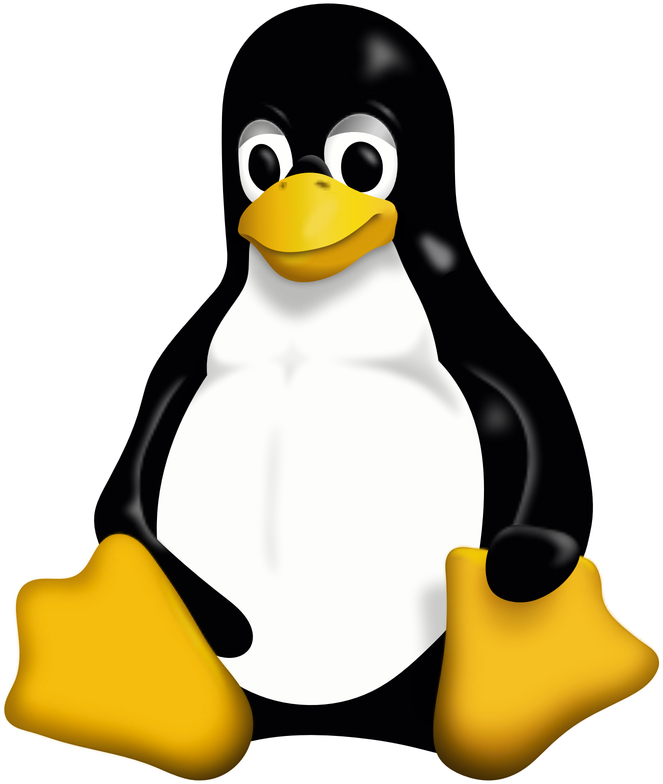
I've currently ported bash(1.08) and gcc(1.40), and things seem to work. This implies that I'll get something practical within a few months, and I'd like to know what features most people would want. Any suggestions are welcome, but I won't promise I'll implement them :-)

Linus (torv...@kruuna.helsinki.fi)

PS. Yes - it's free of any minix code, and it has a multi-threaded fs. It is NOT protable (uses 386 task switching etc), and it probably never will support anything other than AT-harddisks, as that's all I have :-(.

Linux

- 1992: LINUX is obsolete (Andy Tanenbaum, author of handbook on OS design and minix)
- ”would have failed my course”
- 1993: bitten by a ferocious penguin in Canberra Zoo while traveling to Linux.Conf.Au. Contracted ‘penguinitis’ (fixation by flightless fat waterfowl)
- 1996: Tux



From personal OS to 96.3% of the world’s servers.

The Cathedral and the Bazaar

- Eric S. Raymond (ESR), hacker worked a.o. on GNU tools

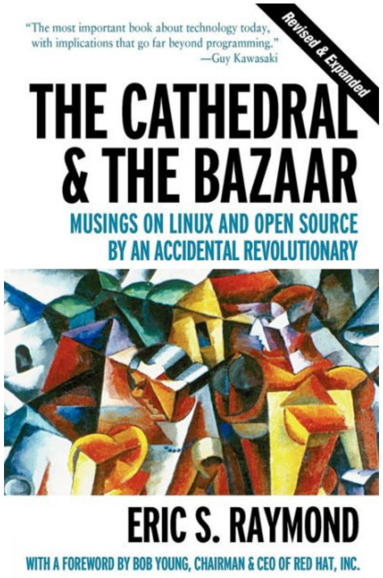
- paper first presented Linux Kongress in **1997** (**CatB**), then presented at Perl conference
- analysis of Linux' development model, application to **fetchmail** (which ESR authored)

The Cathedral and the Bazaar

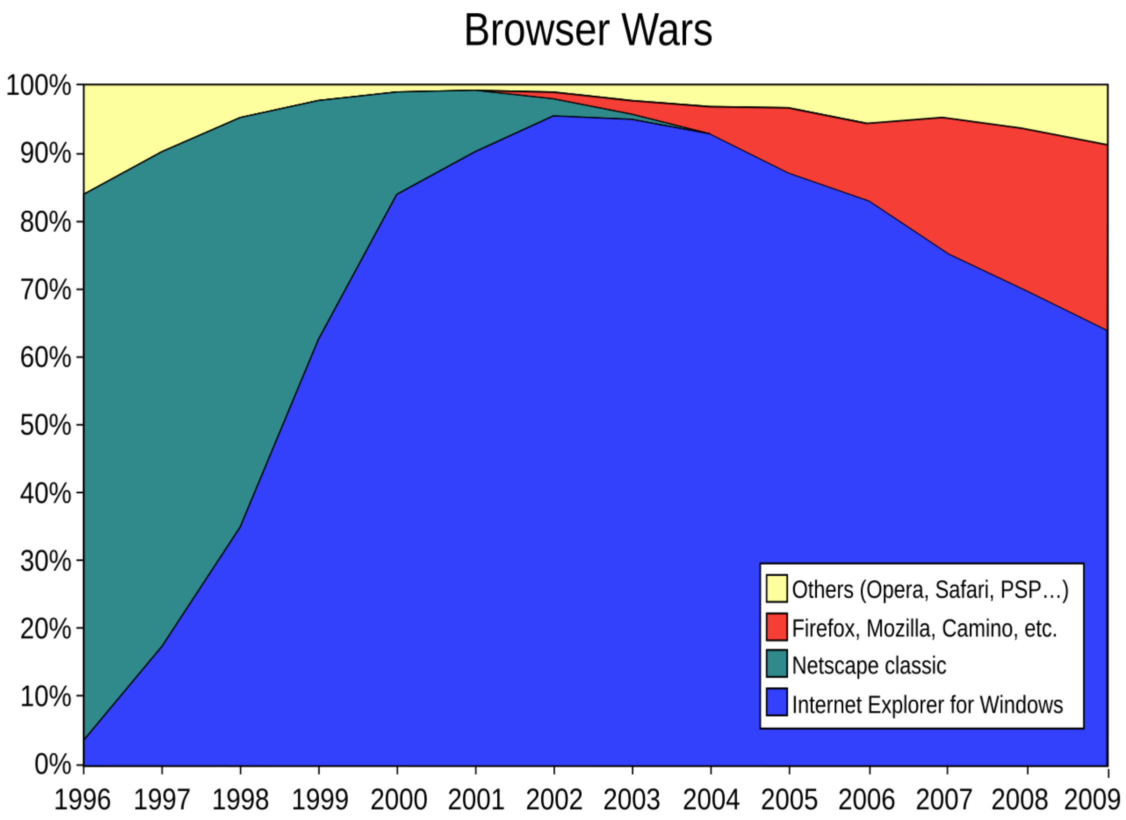
quiet, reverent, **cathedral** vs. babbling **bazaar** of different agendas and approaches

- develop software with the entire world, cf. nascent internet (WWW – 1989)
- "release early, release often"
- parallellize debugging / bug fixing
- recognize good ideas people have
- communication skills ('world domination')

Linus's law: "given enough eyeballs, all bugs are shallow"

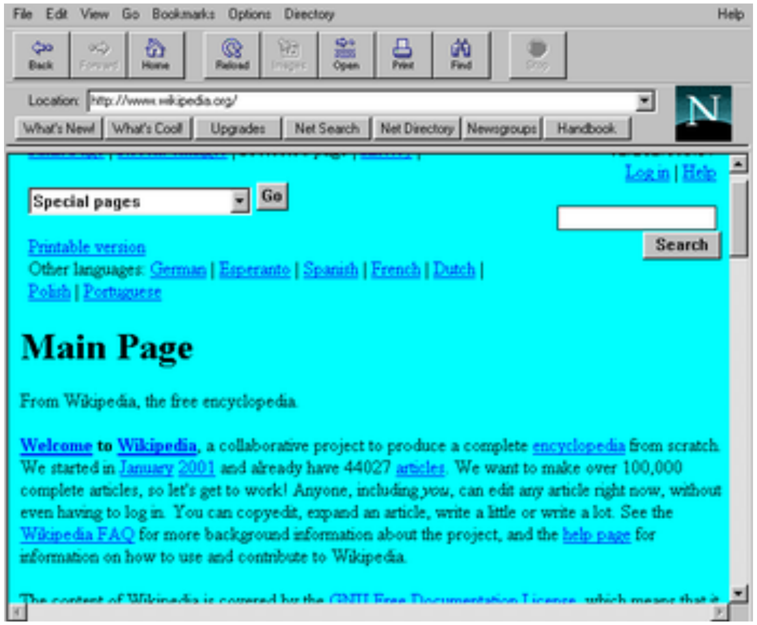

"The most important book about technology today, with implications that go far beyond programming." —Guy Kawasaki
Revised & Expanded
THE CATHEDRAL & THE BAZAAR
MUSINGS ON LINUX AND OPEN SOURCE BY AN ACCIDENTAL REVOLUTIONARY
ERIC S. RAYMOND
WITH A FOREWORD BY BOB YOUNG, CHAIRMAN & CEO OF RED HAT, INC.

Browser wars



- ‘browser wars’:
  - first browser war: IE vs. Netscape Navigator (1995–2001)
  - second browser war: IE vs. Firefox vs. Google Chrome (2004–2017)

Netscape


File Edit View Go Bookmarks Options Directory Help
Back
Forward
Home
Reload
Images
Open
Print
Find
Stop
Location: http://www.wikipedia.org/
What's New!
What's Cool!
Upgrades
Net Search
Net Directory
Newsgroups
Handbook
Log in | Help
Special pages
Go
Search
Printable version
Other languages: German | Esperanto | Spanish | French | Dutch | Polish | Portuguese

Main Page

From Wikipedia, the free encyclopedia.

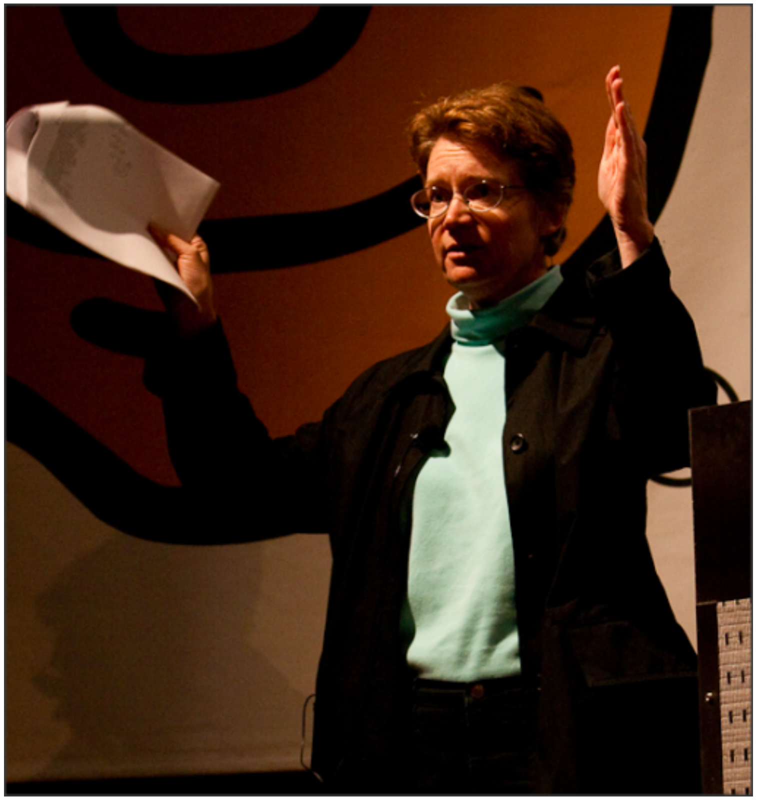
Welcome to Wikipedia, a collaborative project to produce a complete encyclopedia from scratch. We started in January 2001 and already have 44027 articles. We want to make over 100,000 complete articles, so let's get to work! Anyone, including *you*, can edit any article right now, without even having to log in. You can copyedit, expand an article, write a little or write a lot. See the Wikipedia FAQ for more background information about the project, and the help page for information on how to use and contribute to Wikipedia.

The content of Wikipedia is covered by the GNU Free Documentation License, which means that it

- CatB not only insightful essay, but also highly influential
- attended talk on CatB and reported to management of Netscape
- **January 1998**: announcement of a source code release for Navigator
- open sourcing of Netscape, became Mozilla, became **Firefox**
- Netscape Public License, later Mozilla Public License (**MPL**)

open source (the word)

- concern that conservative business people were put off
- afraid **freedom pitch** would stall development of Linux in the **business world** (vs. academic world)
- moral arguments, rather than pragmatic ones, alienate potential allies and hurt the end goal of removing code secrecy
- cf. veil of ignorance (John Rawls) vs. ignore half a century of economics
- meetings following Netscape’s announcement with ESR and a.o. **Christine Peterson** (February 3, 1998)



OSI

- The Open Source Initiative, founded by **ESR** and Bruce Perens in February **1998**

open source
initiative®

- OSI defines open source as a “development method for software that harnesses the power of distributed peer review and transparency of process. The promise of open source is better quality, higher reliability, more flexibility, lower cost, and an end to predatory vendor lock-in“
- Maintains the open source definition (Bruce Perens), approves licenses (‘**OSI Approved Licenses**’)
- RMS: ”Open source is a development methodology; free software is a social movement.“

FOSS, FLOSS and other terminology

- FOSS
- **Libre Software** alternative for Free Software , hence **FLOSS**
- **freeware**: free, typically with all features, but proprietary
- **shareware**: proprietary: share with limited features
- **freemium** software: proprietary, but free except for version with real features

Licenses

copyleft

‘hack’ to use copyright law against itself:

- usually: used to protect your work
- here: use to tell what can be done with it in order to enforce user freedom (reciprocal)
- redistribution and derivative works must be released under a free license
- has its own symbol: 🄯

On June 1, 2001, Microsoft CEO Steve Ballmer said “Linux is a cancer that attaches itself in an intellectual property sense to everything it touches.”

GPL

- GNU **General** Public license
- computer scientist with long beard that writes software license
- 1989 (v1), 1991 (v2), 2007 (v3)
- R released under GPL (thanks Martin Maechler!!)
- **Linux** kernel released under GPL
- gcc GNU compiler collection

GPLv3
Free Software

Free as in Freedom

Cygnus

- nothing wrong with making money with software; even hints as part of manuals: paid support
- very first company to provide support for GNU software
- founded by authors on GNU c++ compiler, GNU debugger etc.
- **Cygnus** solutions (originally Cygnus support) – magnum, wingnut, lugnut, etc.
- **Cygwin** (port of GNU and POSIX layer for M$ Windows)
- merged into RedHat in 1999–2000

cygnus

Internet

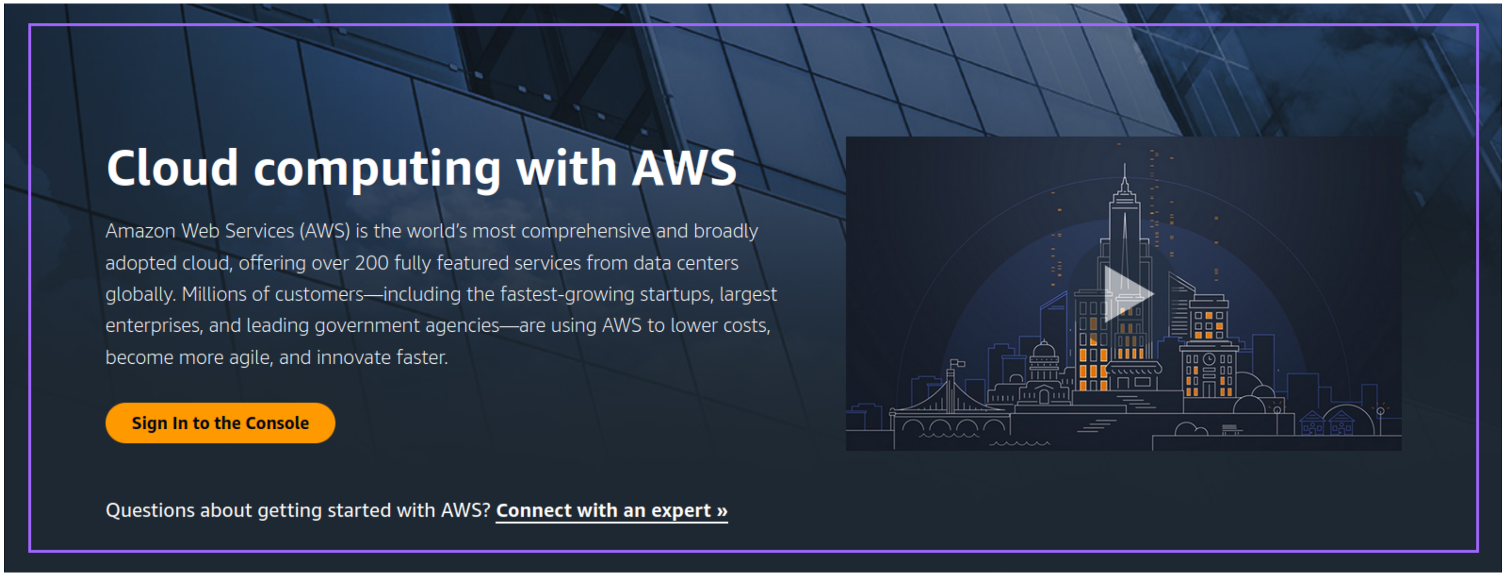
- more applications offered over the net (web application providers, ‘SaaS’)
- Affero Inc. company that wanted a GPL-like license for its **e-learning platform**: Affero General Public License
- close collaboration with FSF (a.o. Eben Moglen)

GNU Affero Public License (2000 - 2007)

- GPL: distribute software and get source code of modified versions that are distributed
- AGPL: run software over a network and still get source code of changes

Cloud era

- The cloud is just someone else's computers.
- Basic services like VMs, blob storage, file systems (2002: AWS, 2006: S3, EC2)
- Ever expanding: currently about 200 services.



Cloud computing with AWS

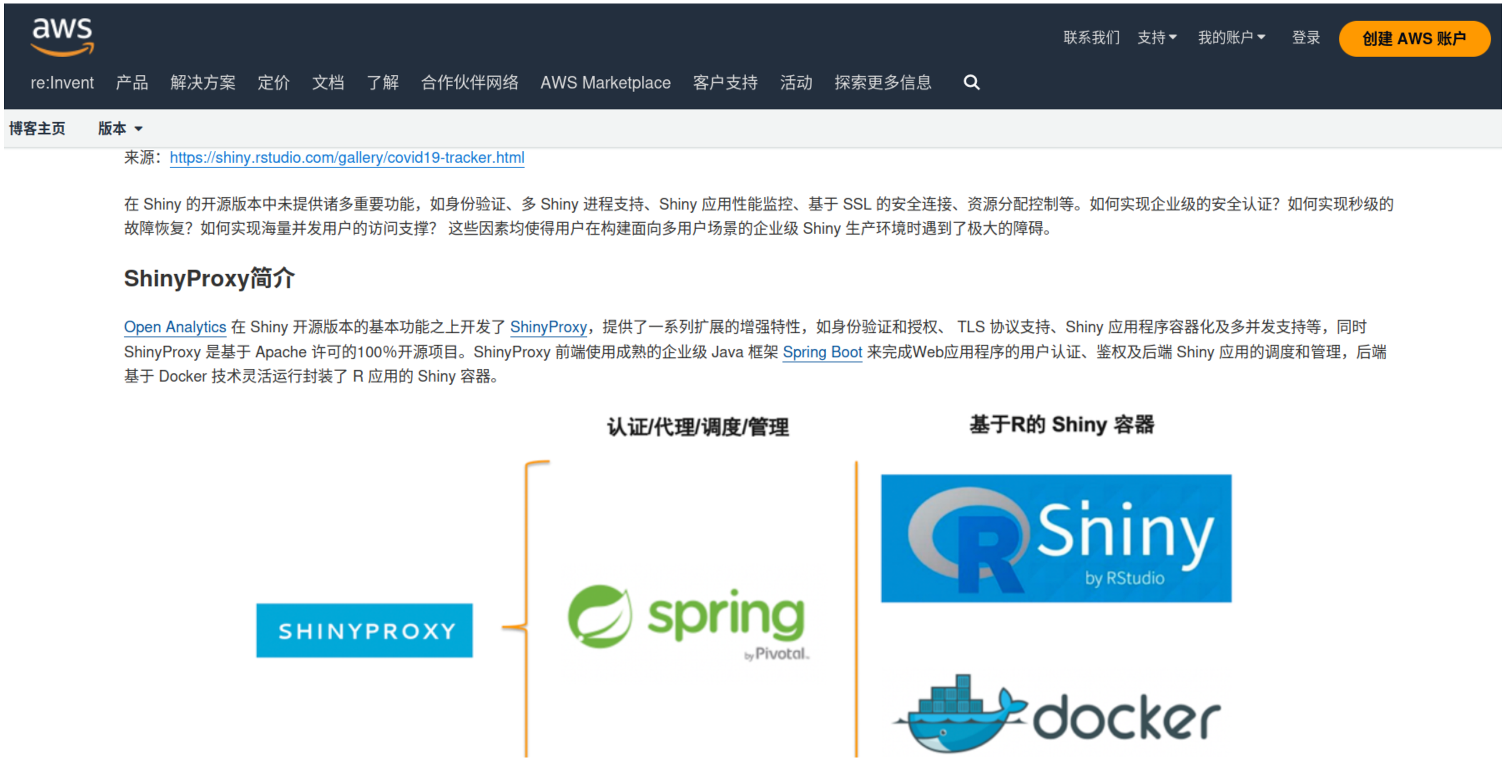
Amazon Web Services (AWS) is the world's most comprehensive and broadly adopted cloud, offering over 200 fully featured services from data centers globally. Millions of customers—including the fastest-growing startups, largest enterprises, and leading government agencies—are using AWS to lower costs, become more agile, and innovate faster.

Questions about getting started with AWS? **Connect with an expert »**

Server Side Public License

- Example: **database** services: many users turn to cloud vendors for running databases
- Cloud vendors take open source software and package it as a service
- not appreciated by venture capitalists, MongoDB introduced **SSPL**, ”fauxpen” source license
- example followed by other projects ElasticSearch, Redis
- release the entirety of their source code, including all software, APIs, and other software that would be required for a user to run an instance of the service themselves, under the SSPL.
- **forks** of the projects e.g. OpenSearch = ElasticSearch but by Amazon

So What?



来源：https://shiny.rstudio.com/gallery/covid19-tracker.html

在 Shiny 的开源版本中未提供诸多重要功能，如身份验证、多 Shiny 进程支持、Shiny 应用性能监控、基于 SSL 的安全连接、资源分配控制等。如何实现企业级的安全认证？如何实现秒级的故障恢复？如何实现海量并发用户的访问支撑？ 这些因素均使得用户在构建面向多用户场景的企业级 Shiny 生产环境时遇到了极大的障碍。

ShinyProxy简介

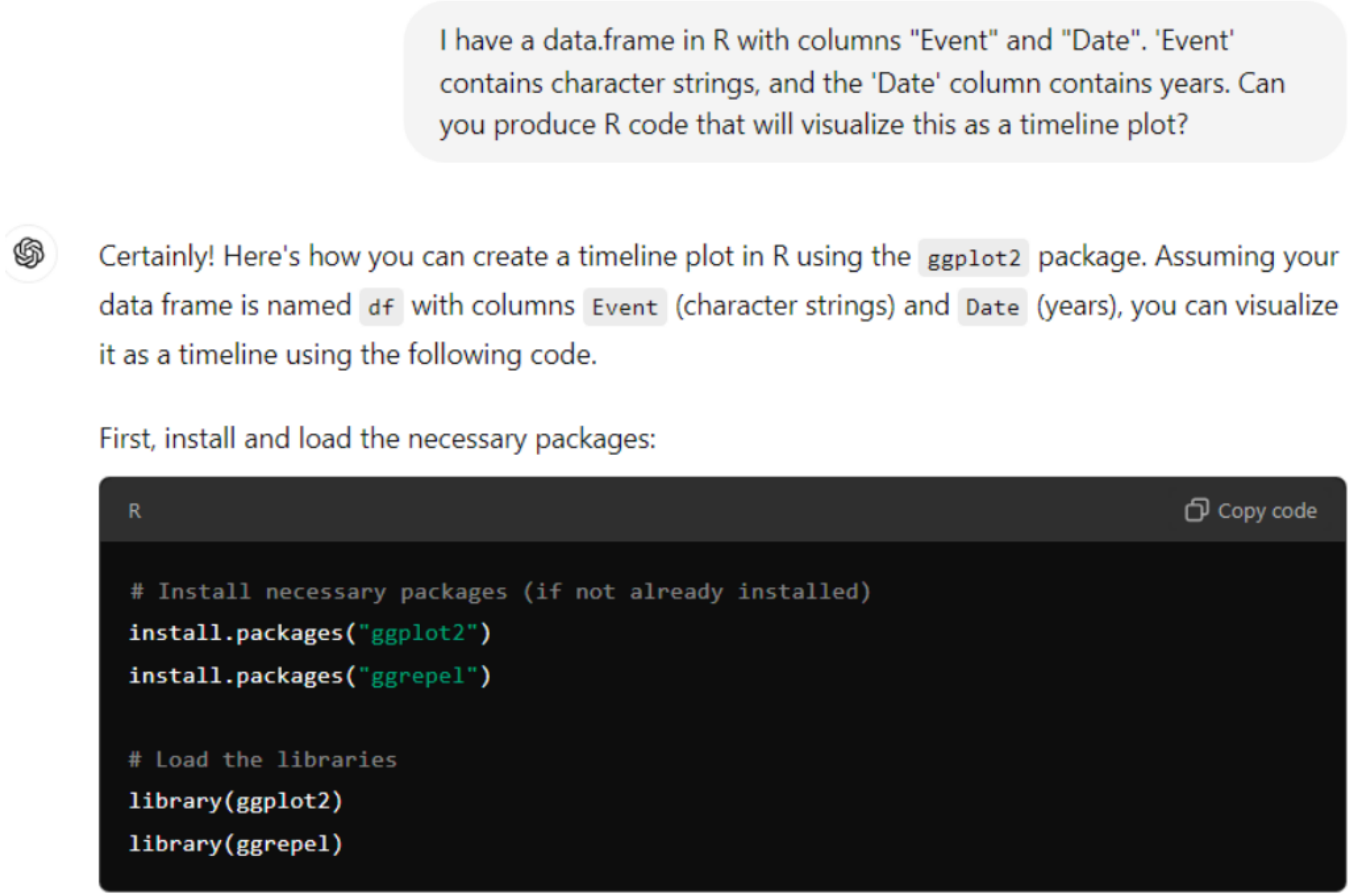
Open Analytics 在 Shiny 开源版本的基本功能之上开发了 ShinyProxy，提供了一系列扩展的增强特性，如身份验证和授权、 TLS 协议支持、Shiny 应用程序容器化及多并发支持等，同时 ShinyProxy 是基于 Apache 许可的100%开源项目。ShinyProxy 前端使用成熟的企业级 Java 框架 Spring Boot 来完成Web应用程序的用户认证、鉴权及后端 Shiny 应用的调度和管理，后端基于 Docker 技术灵活运行封装了 R 应用的 Shiny 容器。

Permissive licenses

- derivative works
- licenses
  - BSD
  - MIT
  - Apache 2.0
- opens door for **open core**
- real features require ‘enterprise edition’, ‘pro version’ etc.
- synonym for **crippleware**
- open core is for **beeeeeeeeeeeeeeeeeeeeeeeep**

Statistical Computing and AI

Timeline of Statistical Computing and AI



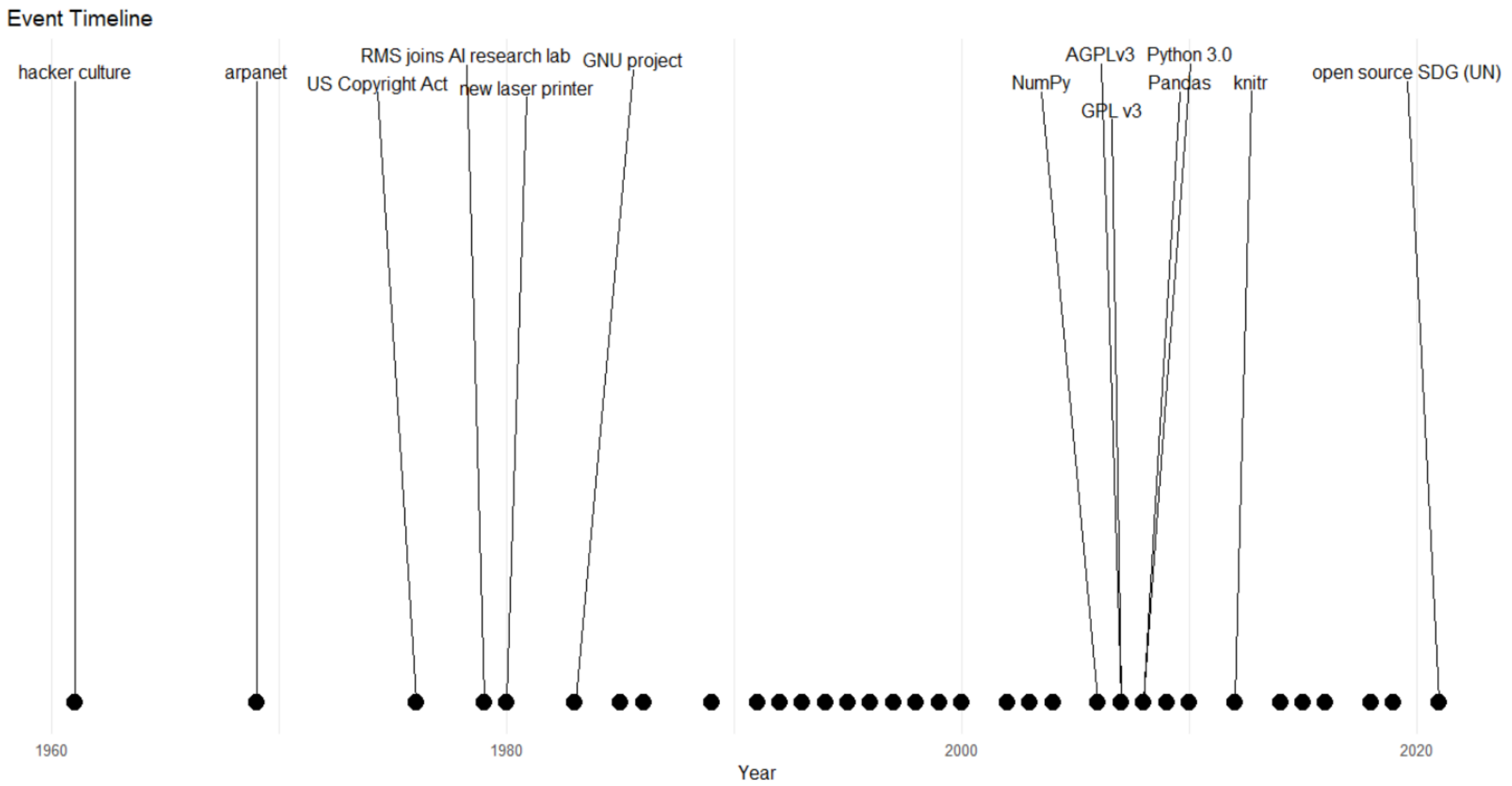
I have a data.frame in R with columns "Event" and "Date". 'Event' contains character strings, and the 'Date' column contains years. Can you produce R code that will visualize this as a timeline plot?

Certainly! Here's how you can create a timeline plot in R using the `ggplot2` package. Assuming your data frame is named `df` with columns `Event` (character strings) and `Date` (years), you can visualize it as a timeline using the following code.

First, install and load the necessary packages:

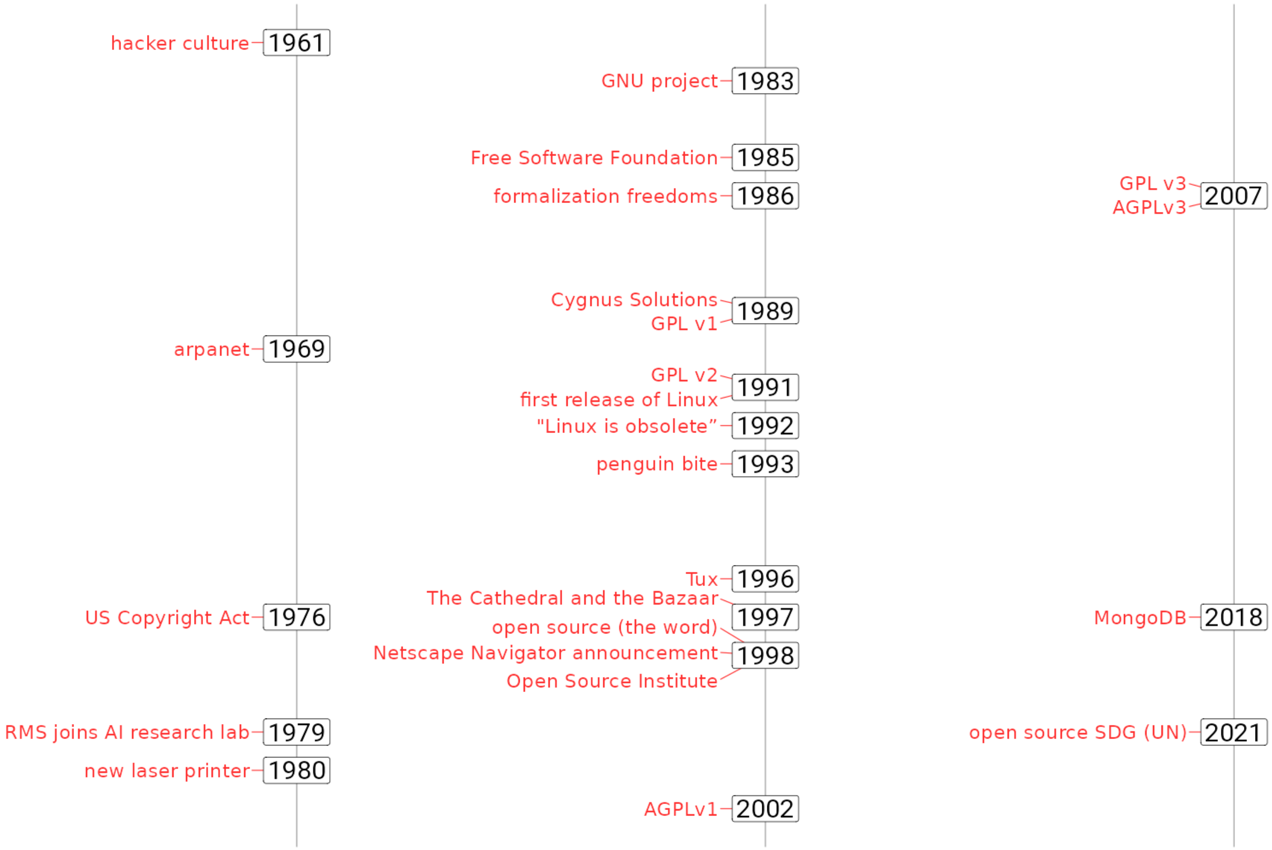
```
# Install necessary packages (if not already installed)
install.packages("ggplot2")
install.packages("ggrepel")

# Load the libraries
library(ggplot2)
library(ggrepel)
```

Timeline of Statistical Computing and AI

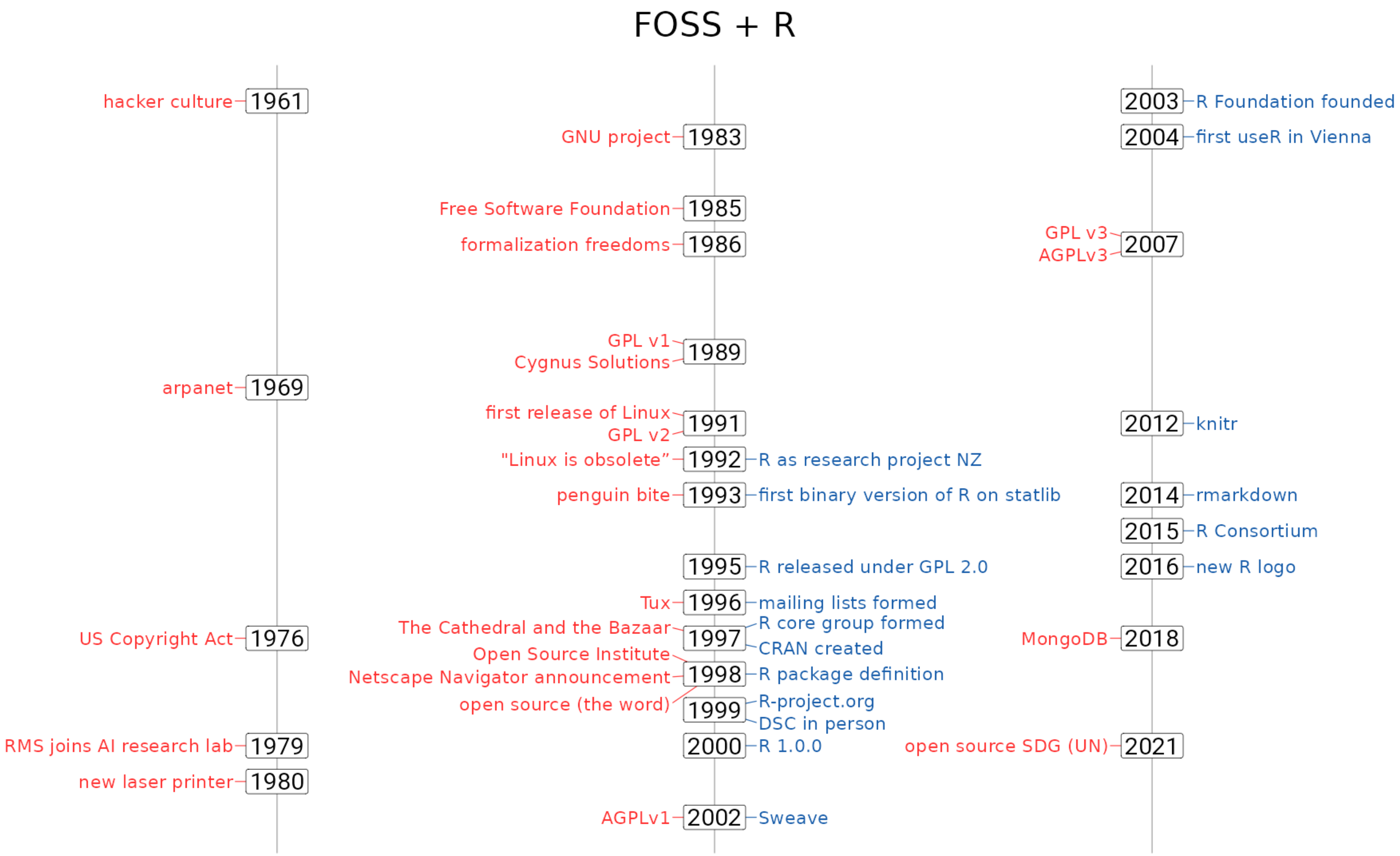


Timeline of Statistical Computing and AI

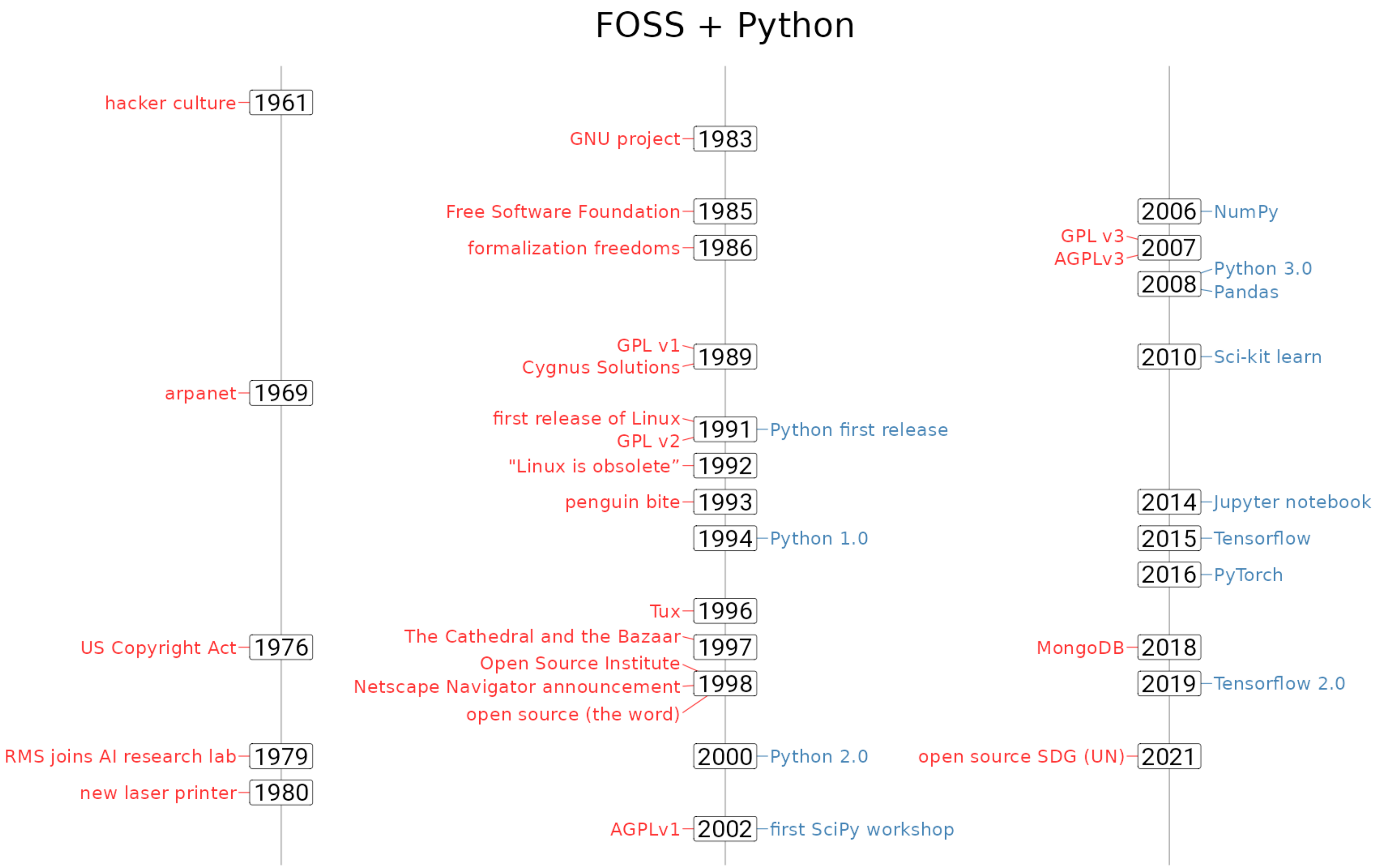
FOSS



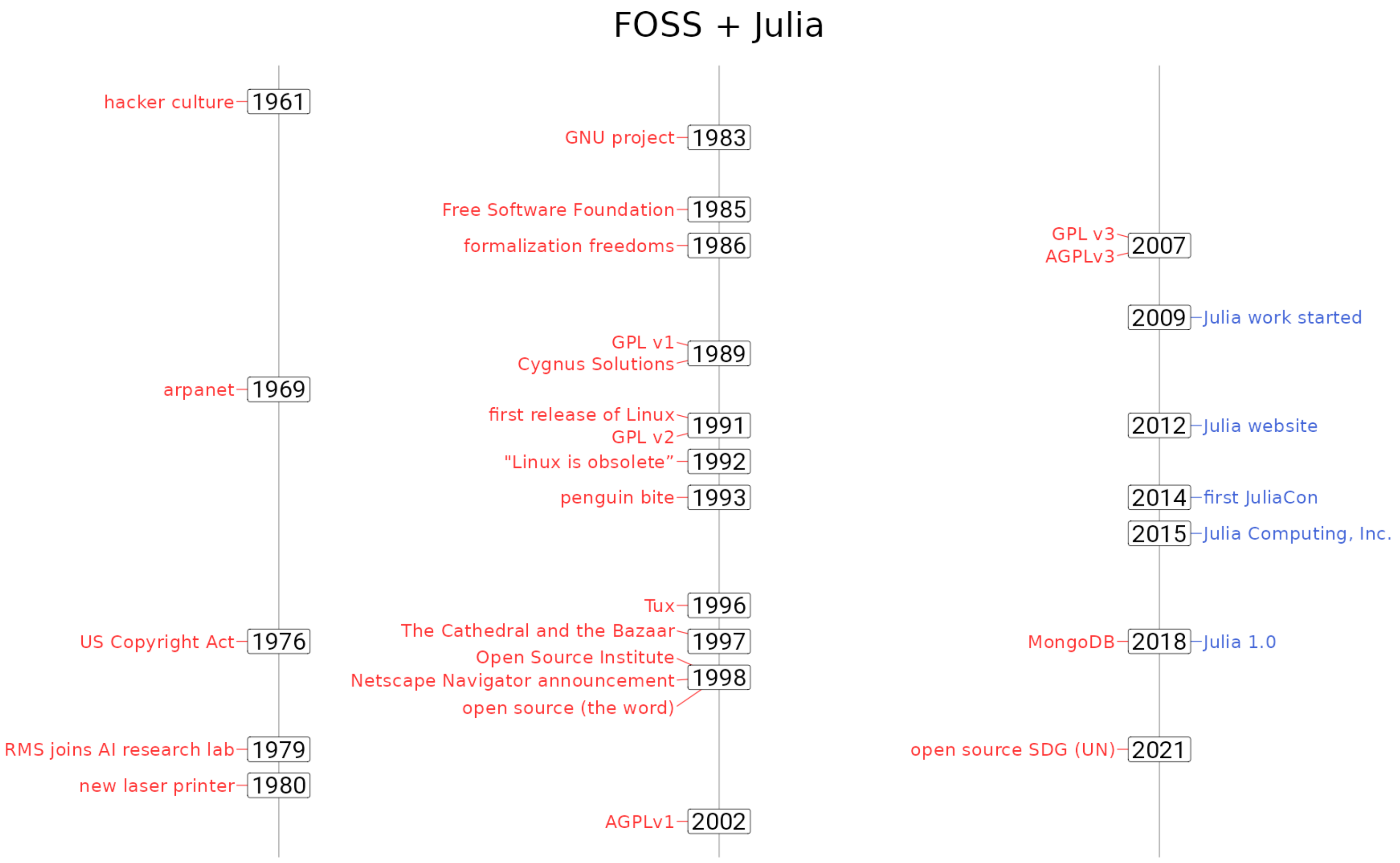
FOSS + R


FOSS + R
hacker culture 1961
arpanet 1969
US Copyright Act 1976
RMS joins AI research lab 1979
new laser printer 1980
GNU project 1983
Free Software Foundation 1985
formalization freedoms 1986
GPL v1
Cygnus Solutions 1989
first release of Linux
GPL v2 1991
"Linux is obsolete" 1992 R as research project NZ
penguin bite 1993 first binary version of R on statlib
1995 R released under GPL 2.0
Tux 1996 mailing lists formed
The Cathedral and the Bazaar 1997 R core group formed
CRAN created
Open Source Institute
Netscape Navigator announcement 1998 R package definition
open source (the word) 1999 R-project.org
DSC in person
2000 R 1.0.0
AGPLv1 2002 Sweave
2003 R Foundation founded
2004 first useR in Vienna
GPL v3
AGPLv3 2007
2012 knitr
2014 rmarkdown
2015 R Consortium
2016 new R logo
MongoDB 2018
open source SDG (UN) 2021

FOSS + Python


FOSS + Python
hacker culture 1961
arpanet 1969
US Copyright Act 1976
RMS joins AI research lab 1979
new laser printer 1980
GNU project 1983
Free Software Foundation 1985
formalization freedoms 1986
GPL v1
Cygnus Solutions 1989
first release of Linux
GPL v2 1991 Python first release
"Linux is obsolete" 1992
penguin bite 1993
1994 Python 1.0
Tux 1996
The Cathedral and the Bazaar 1997
Open Source Institute
Netscape Navigator announcement 1998
open source (the word)
2000 Python 2.0
AGPLv1 2002 first SciPy workshop
2006 NumPy
GPL v3
AGPLv3 2007
2008 Python 3.0
Pandas
2010 Sci-kit learn
2014 Jupyter notebook
2015 Tensorflow
2016 PyTorch
MongoDB 2018
2019 Tensorflow 2.0
open source SDG (UN) 2021

FOSS + Julia


FOSS + Julia
hacker culture
1961
arpanet
1969
US Copyright Act
1976
RMS joins AI research lab
1979
new laser printer
1980
GNU project
1983
Free Software Foundation
1985
formalization freedoms
1986
GPL v1
Cygnus Solutions
1989
first release of Linux
GPL v2
1991
"Linux is obsolete”
1992
penguin bite
1993
Tux
1996
The Cathedral and the Bazaar
1997
Open Source Institute
Netscape Navigator announcement
open source (the word)
1998
AGPLv1
2002
GPL v3
AGPLv3
2007
2009
Julia work started
2012
Julia website
2014
first JuliaCon
2015
Julia Computing, Inc.
MongoDB
2018
Julia 1.0
open source SDG (UN)
2021

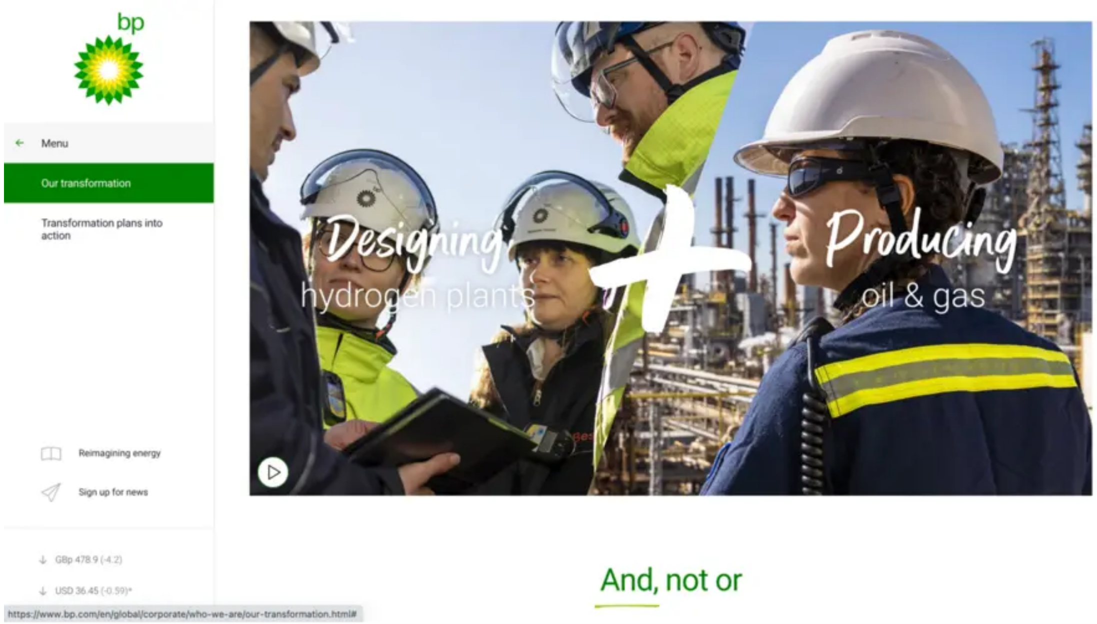
Status

- **mature ecosystems** that cover the entire data science space
- **good interoperability**, you can be multilingual to solve problems
- the problem is solved and my talk is useless

- R is GPL (free software, copyleft)
- Python PFSL (permissive, GPL compatible)
- Julia (MIT, permissive) – cf. supra

Openwashing in Statistical Computing

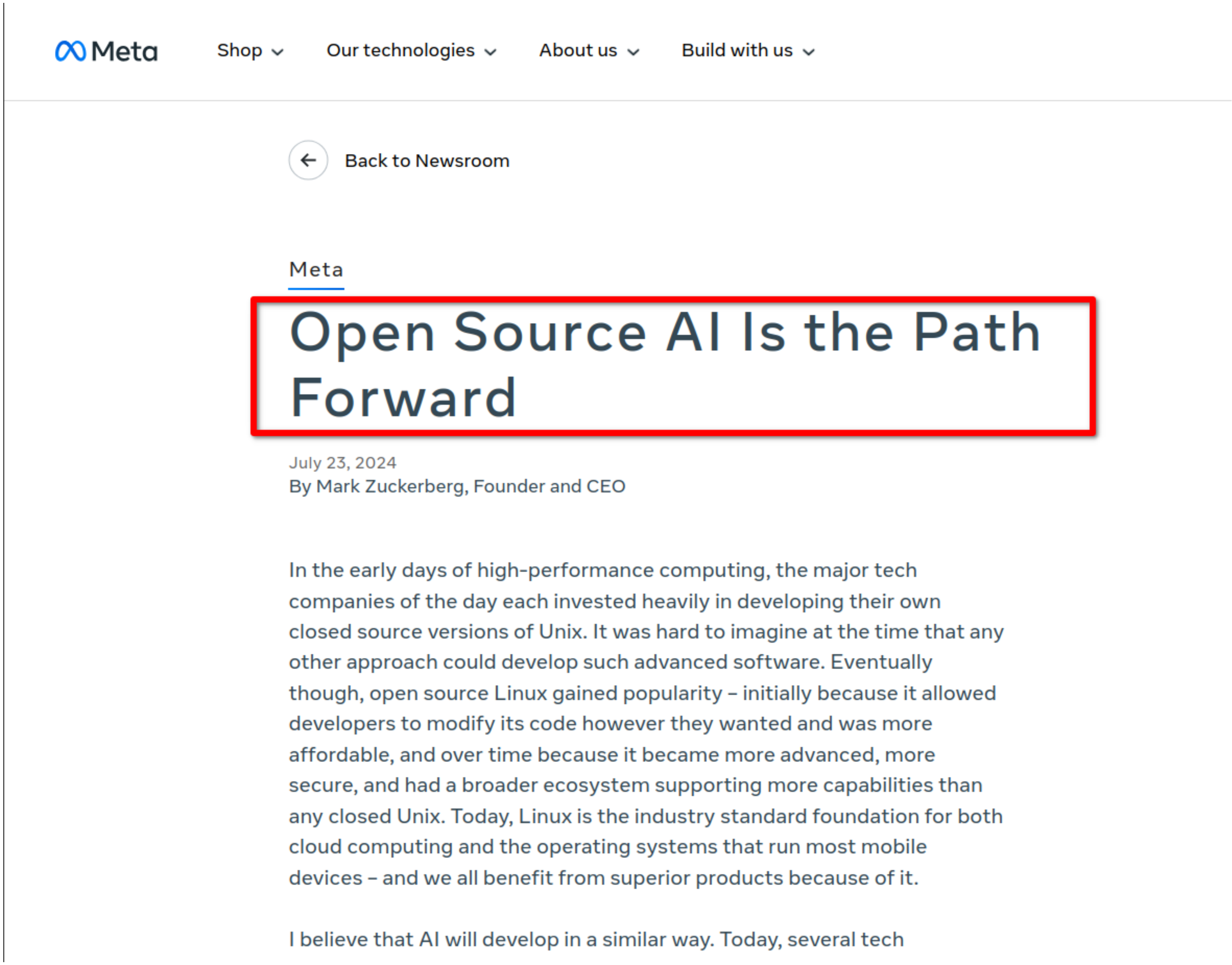
- **greenwashing** translated to open source software
- Julia: great initiative and ecosystem, but open core aftertaste (**Julia Computing Inc**, proprietary extensions by inventors of language)
- Posit (formerly **RStudio**): ‘open source data science’, all of their products are **closed source**
- started off with RStudio (open core), next generation Positron: fauxpen source license, **Elastic 2.0** (‘source available license’)



Openwashing in AI

- **OpenAI**: nothing more closed than OpenAI (“Micro$oft AI Labs“)
- **openness** being used as way to increase **market share**: post by Mark Zuckerberg: ”Open Source AI is the Path Forward“ (July 2024)
- no way to recreate the model: **not open source**
- OSI initiative set to define **open source AI**, currently running into the wall

Meta



Meta

Open Source AI Is the Path Forward

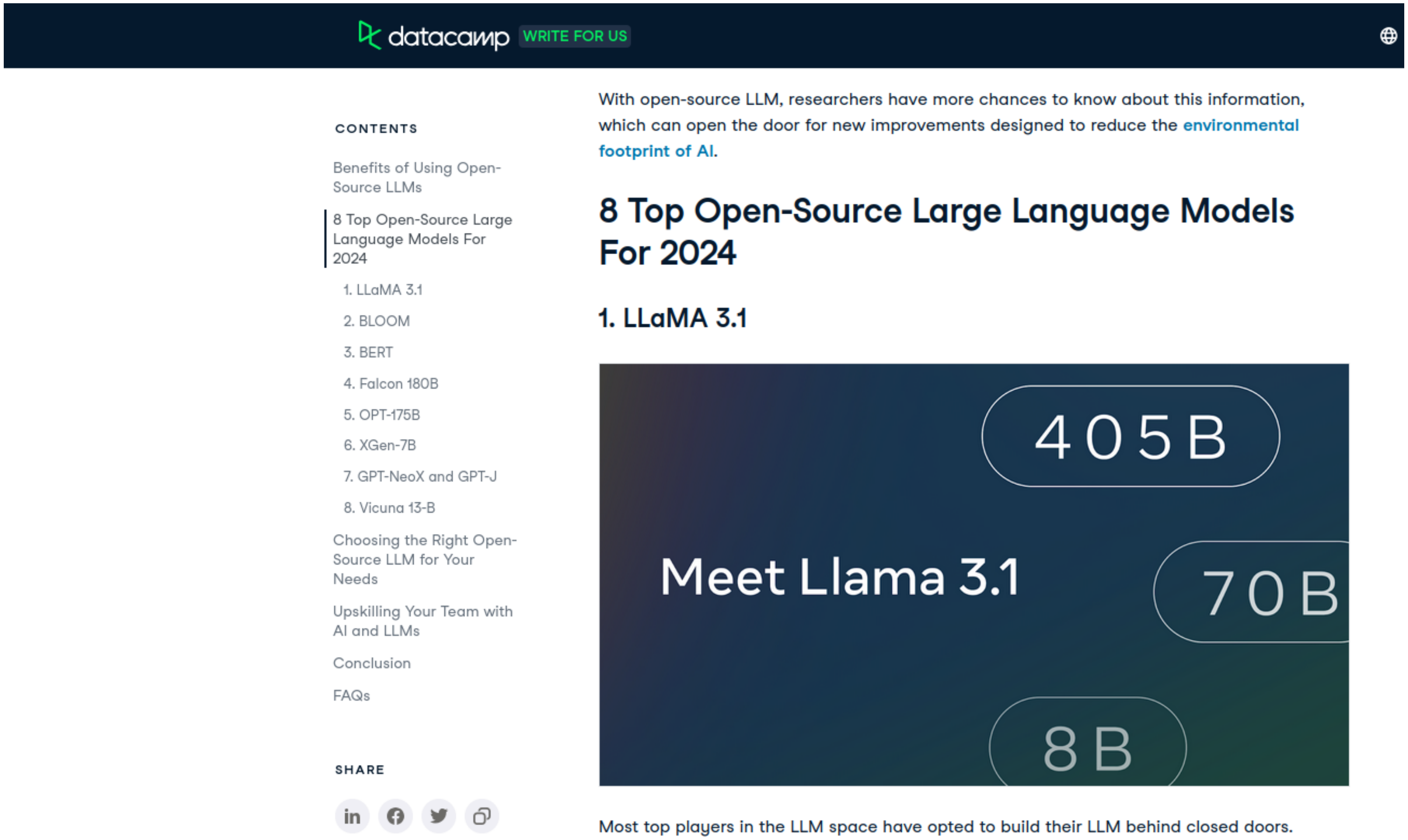
July 23, 2024

By Mark Zuckerberg, Founder and CEO

In the early days of high-performance computing, the major tech companies of the day each invested heavily in developing their own closed source versions of Unix. It was hard to imagine at the time that any other approach could develop such advanced software. Eventually though, open source Linux gained popularity – initially because it allowed developers to modify its code however they wanted and was more affordable, and over time because it became more advanced, more secure, and had a broader ecosystem supporting more capabilities than any closed Unix. Today, Linux is the industry standard foundation for both cloud computing and the operating systems that run most mobile devices – and we all benefit from superior products because of it.

I believe that AI will develop in a similar way. Today, several tech

Echos



CONTENTS

Benefits of Using Open-Source LLMs
8 Top Open-Source Large Language Models For 2024
  1. LLaMA 3.1
  2. BLOOM
  3. BERT
  4. Falcon 180B
  5. OPT-175B
  6. XGen-7B
  7. GPT-NeoX and GPT-J
  8. Vicuna 13-B
Choosing the Right Open-Source LLM for Your Needs
Upskilling Your Team with AI and LLMs
Conclusion
FAQs

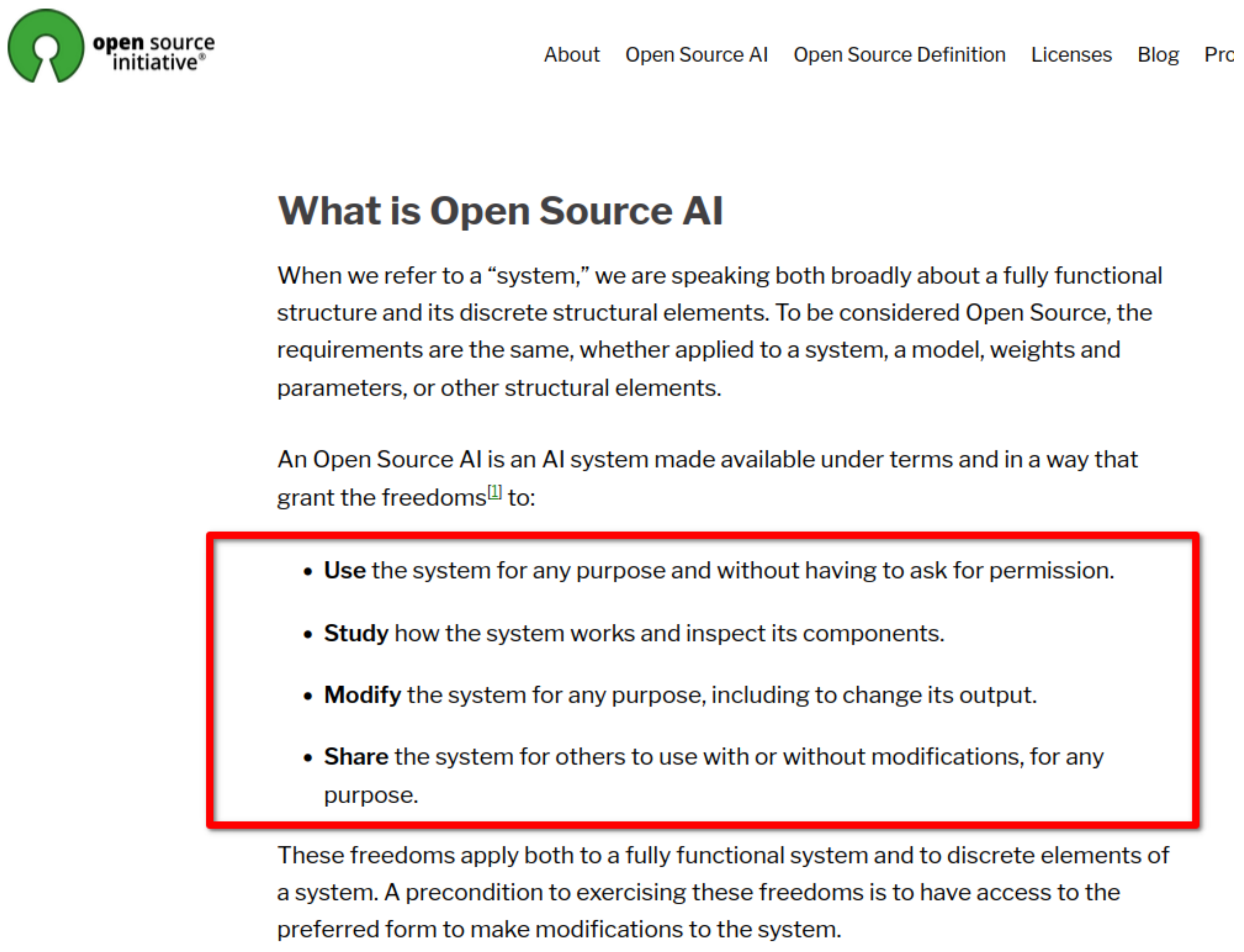
With open-source LLM, researchers have more chances to know about this information, which can open the door for new improvements designed to reduce the **environmental footprint of AI**.

8 Top Open-Source Large Language Models For 2024

1. LLaMA 3.1

Most top players in the LLM space have opted to build their LLM behind closed doors.

OSI Open Source AI



What is Open Source AI

When we refer to a “system,” we are speaking both broadly about a fully functional structure and its discrete structural elements. To be considered Open Source, the requirements are the same, whether applied to a system, a model, weights and parameters, or other structural elements.

An Open Source AI is an AI system made available under terms and in a way that grant the freedoms[1] to:

- **Use** the system for any purpose and without having to ask for permission.
- **Study** how the system works and inspect its components.
- **Modify** the system for any purpose, including to change its output.
- **Share** the system for others to use with or without modifications, for any purpose.

These freedoms apply both to a fully functional system and to discrete elements of a system. A precondition to exercising these freedoms is to have access to the preferred form to make modifications to the system.

Open Source AI

```r
data(cars) # speed in mph, distance in feet
linearModel <- lm(dist ~ speed, data = cars)
coef(linearModel)
```

```
## (Intercept)       speed
##  -17.579095    3.932409
```

```r
predict(linearModel, newdata = data.frame(speed = 150), type = "response")
```

```
##        1
## 572.2822
```

OSI Open Source AI Checklist

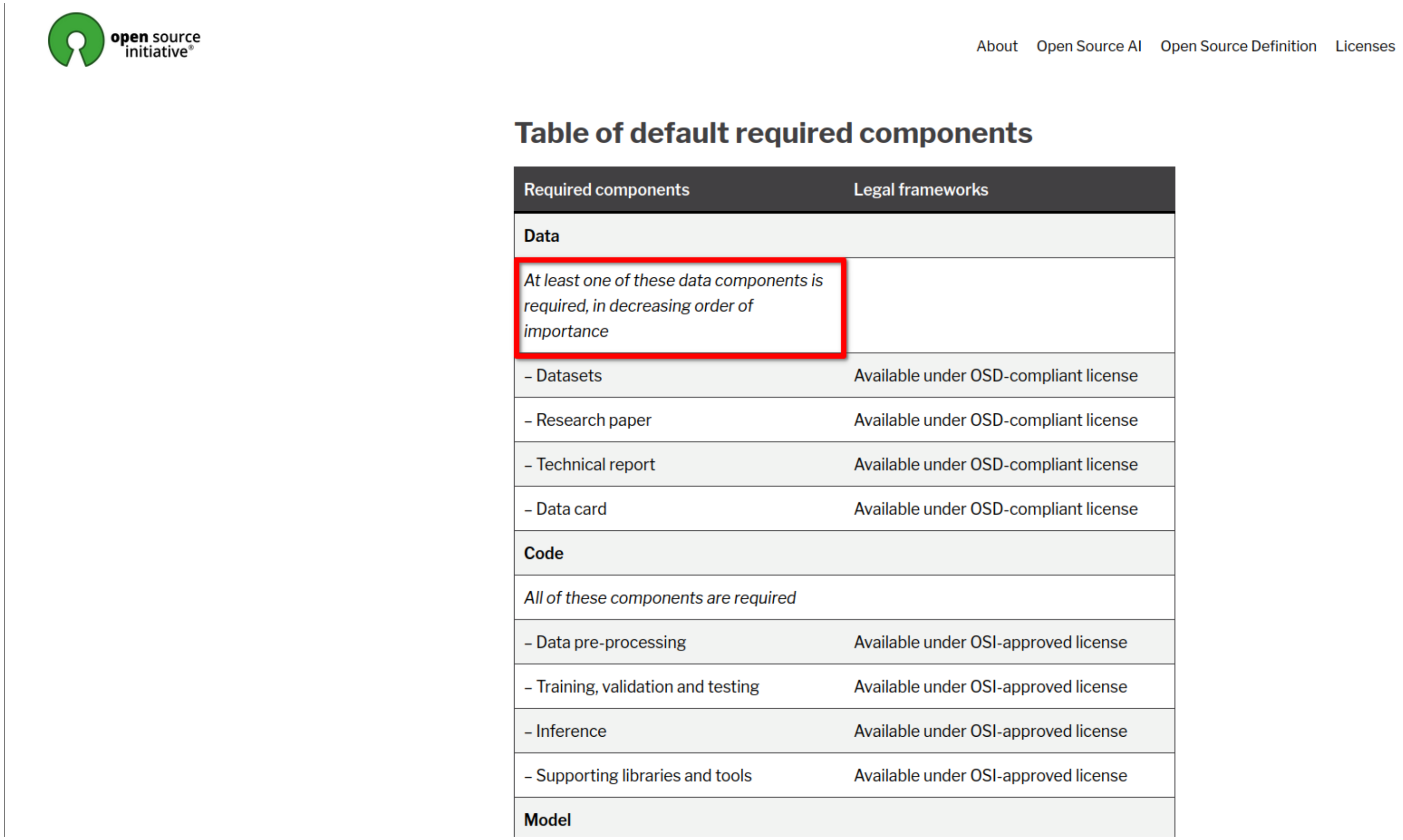


Table of default required components

Required components	Legal frameworks
Data	
At least one of these data components is required, in decreasing order of importance	
– Datasets	Available under OSD-compliant license
– Research paper	Available under OSD-compliant license
– Technical report	Available under OSD-compliant license
– Data card	Available under OSD-compliant license
Code	
All of these components are required	
– Data pre-processing	Available under OSI-approved license
– Training, validation and testing	Available under OSI-approved license
– Inference	Available under OSI-approved license
– Supporting libraries and tools	Available under OSI-approved license
Model	

Conclusion

What are you doing?

- You want full control, full transparency and **reproducibility**: you want the **source**
- You **publish** under open source license whatever can be useful to other people
- You enter **precompetitive collaborations** to build open source tooling together
- You do **nonclinical statistics and AI the GNU way**
- You help your **neighbour**…

Questions?

tobias.verbeke@openanalytics.eu

Happy hacking!