



Carnegie Mellon University

Introduction to Software Curation and Preservation

Eric Kaltman

CLIR Fellow for Data Curation in the Sciences

Overview of Workshop

1. Why Software Preservation?
2. What is Software Curation?
 - a. Scope and Description
 - b. Migration
 - c. Reproduction and Access
 - d. Outreach and Culture
3. Trends
4. Efforts at CMU

Overview of Workshop

1. Why Software Preservation?
2. What is Software Curation?
 - a. Scope and Description
 - b. Migration
 - c. Reproduction and Access
 - d. Outreach and Culture
3. Trends
4. Efforts at CMU

What is Software?

Source Code

- Instructions to a computational device to do something

Executable binaries

- Compiled source code that executes on in a target environment

Documentation

- Descriptions of implementation and use

What is Software?

Data encoded to be executed by a specific computational context ('execution context')

Execution Context - the collection of hardware and software dependencies required to allow for execution

Hardware Dependencies - CPUs, system architecture, peripherals, displays, etc.

Software Dependencies - Compiled libraries, support programs, APIs, etc.

Why Software Preservation?

History

- Modern society is built on and from interactions with software

Progress

- Modern research, regardless of the field, is tied to software infrastructures

Pedagogy

- Access to legacy software and systems for teaching and critical engagement.

Practicality

- Software at the root of commercial progress, ties to competition and legal issues

History and Reproducibility

Research reproducibility

- Maintaining citable access to previous results, analyzes and data sets
- Shifting research practices toward an archival and curatorial mindset
- Models for sustainable software development

History, legacy and maintenance

- Preserving the historical record, the intellectual history of humanity
- Aligning with a maintenance narrative instead of an innovation one
- Access to the past will be dependent on software for the rest of time

Access to software is also access to the files produced and interpreted by software

Digital Dark Ages

“People think that bits are somehow immortal because somehow they’re this ethereal thing in cyberspace...It could be that the format of those bits, the way in which they are interpreted requires a piece of software to figure out what the bits mean. How they should be presented as an image or a video or how you should interact with it in a spreadsheet, but the software doesn’t exist. What if the operating system that the old software used to run on doesn’t exist anymore? What if the latest software doesn’t know how to read the formats of those complex digital objects? Guess what? That information is gone.

There isn’t a systematic way to ensure the information that we create today will still be usable 100 years from now. That’s why I’m worried about a ***digital dark ages***.

Vint Cerf, co-inventor of TCP/IP



Ensuring the Longevity of Digital Information

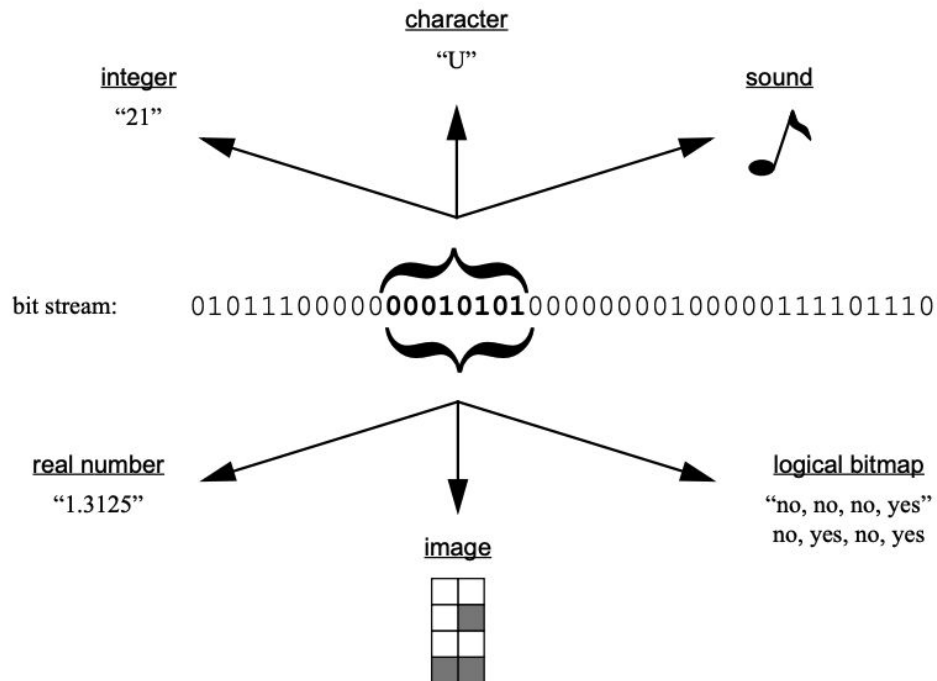


Figure 4: A bit stream can represent anything at all

Jeff Rothenberg
Scientific American
1995

Overview of Workshop

1. Why Software Preservation?
2. What is Software Curation?
 - a. **Scope and Description**
 - b. Migration
 - c. Reproduction and Access
 - d. Outreach and Culture
3. Trends
4. Efforts at CMU

Curation

Scope

- What is historically important?
- What will be important for future research efforts?

Description and Standards

- How do we describe software for future access, study and use?

Migration

- How do we transition and preserve data across time?

Storage and Access

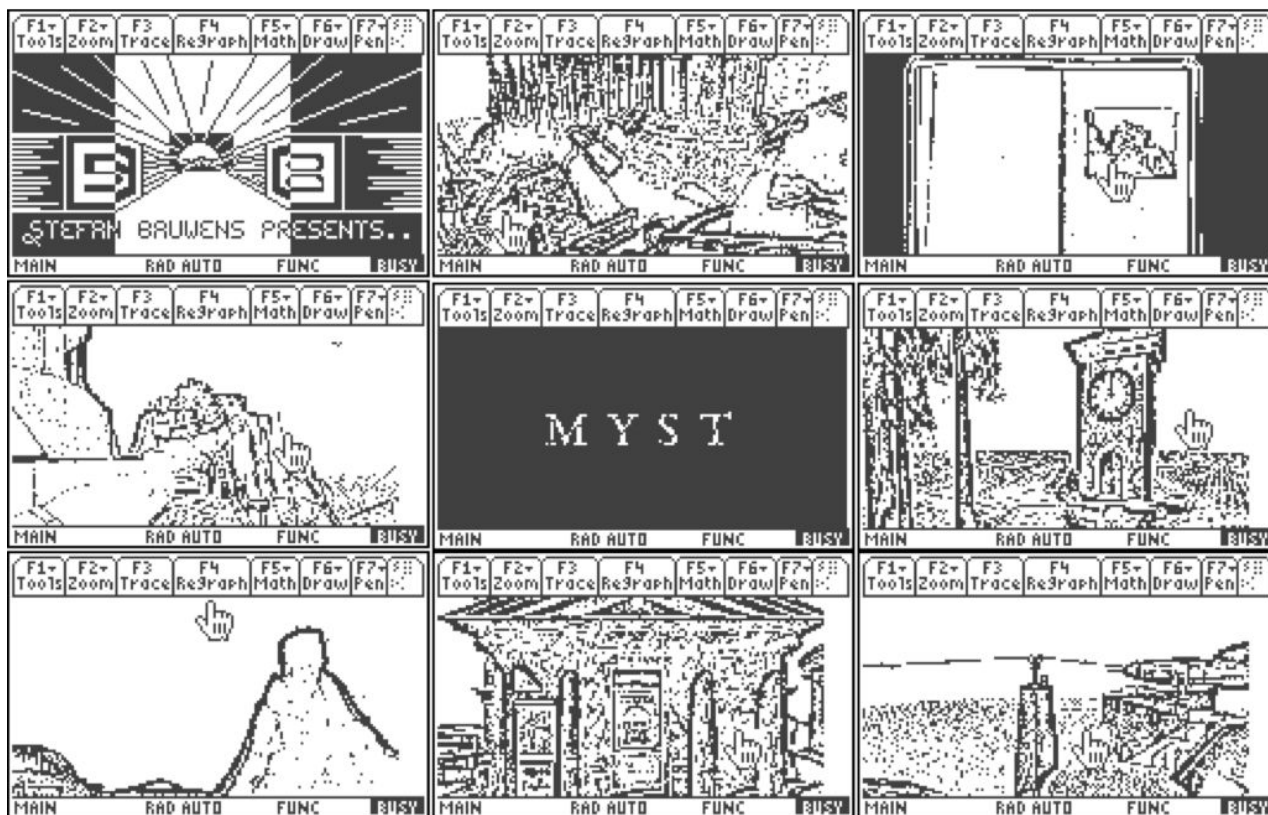
- Where do we put these things?
- How do we find them again?

What is Important?

A medical supply company in Miami had received a delivery of botulin, which was to be processed into Botox and distributed. However, it was misprocessed, and a dangerous concentrate was distributed. The FDA had all of the information needed to identify the recipients, but the information was in a file created with a 2003 version of a popular business software application. The 2004 version available to the FDA could not open the data file. The manufacturer of the software was also unable to supply the relevant version.

It so happened that one of the agents involved in the case was familiar with the NSRL, and had in fact provided software to us earlier in the year. He called, explained the situation, and asked if we had the 2003 version of the software. We did! The agent then arranged for an FDA contact to come to NIST, get the software, and put it on a jet to Miami. The people working the case in Miami were able to install the old version, open the data file, and trace the paths of the botulin.

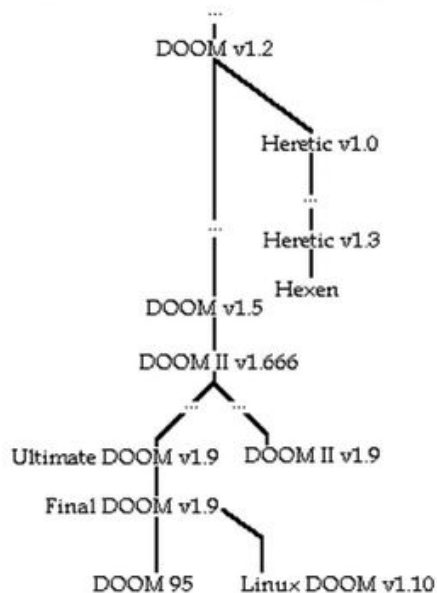
Preserving.Exe: Toward a National Strategy for Software Preservation - 2013



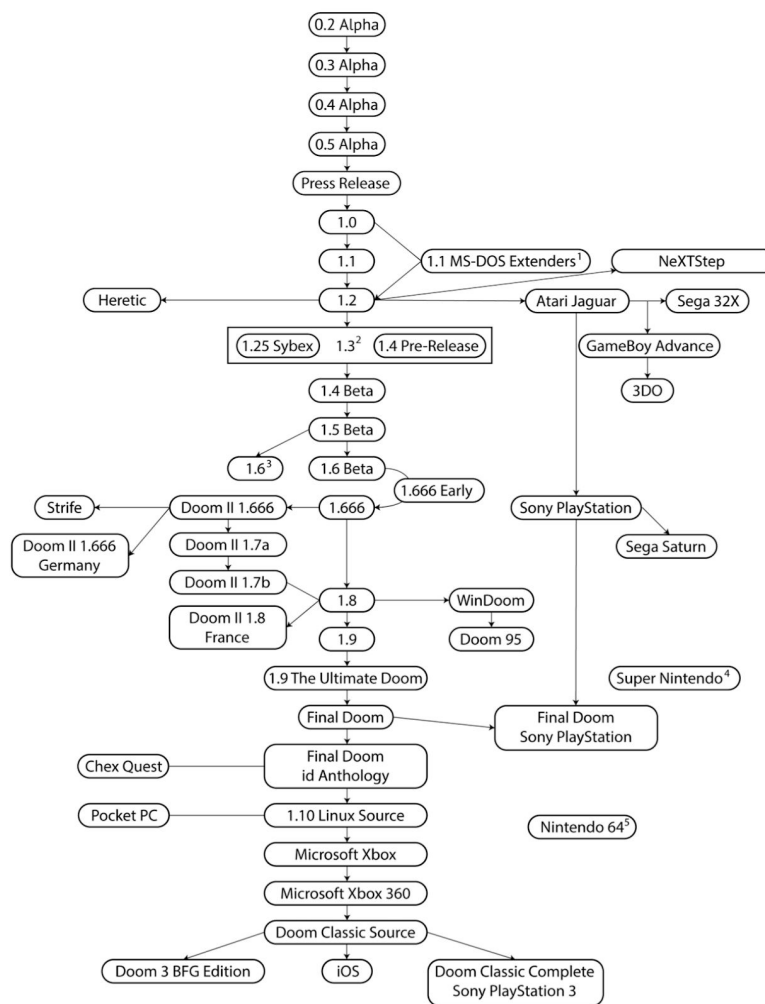








Filename	File Size	Release Date	notes
doom0_2.zip	254k	Feb. 4, 1993	alpha version - doom0_2.zip downloaded by HEL from www.doomworld.com/pageofdoom , 18 July 2008
doom0_4.zip	950k	Apr. 2, 1993	alpha version
doom0_5.zip	1264k	May 22, 1993	alpha version
doom_pre_beta.zip	2555k	Oct. 4, 1993	press-release beta version
doom1_0.zip	2113k	Dec. 10, 1993	v0.99 shareware version
doom1_1.zip	2160k	Dec 16, 1993	v1.1 shareware version
doom1_2.zip	2203k	Feb 17, 1994	shareware version
doom14bt.zip	2246k	Jun. 28, 1994	v1.4beta shareware version
doom15bt.zip	2262k	Jul. 8, 1994	v1.5beta shareware version
doom16bt.zip	2234k	Aug. 3, 1994	v1.6beta shareware version
dm1666sw.zip	2293k	Sep. 1, 1994	v1.666 shareware
doom_v18.zip	2423k	Jan 23, 1995	v1.8 shareware
doom19s.zip	2393k	N/A	v1.9 shareware version



1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017

GNU/Linux Distributions Timeline

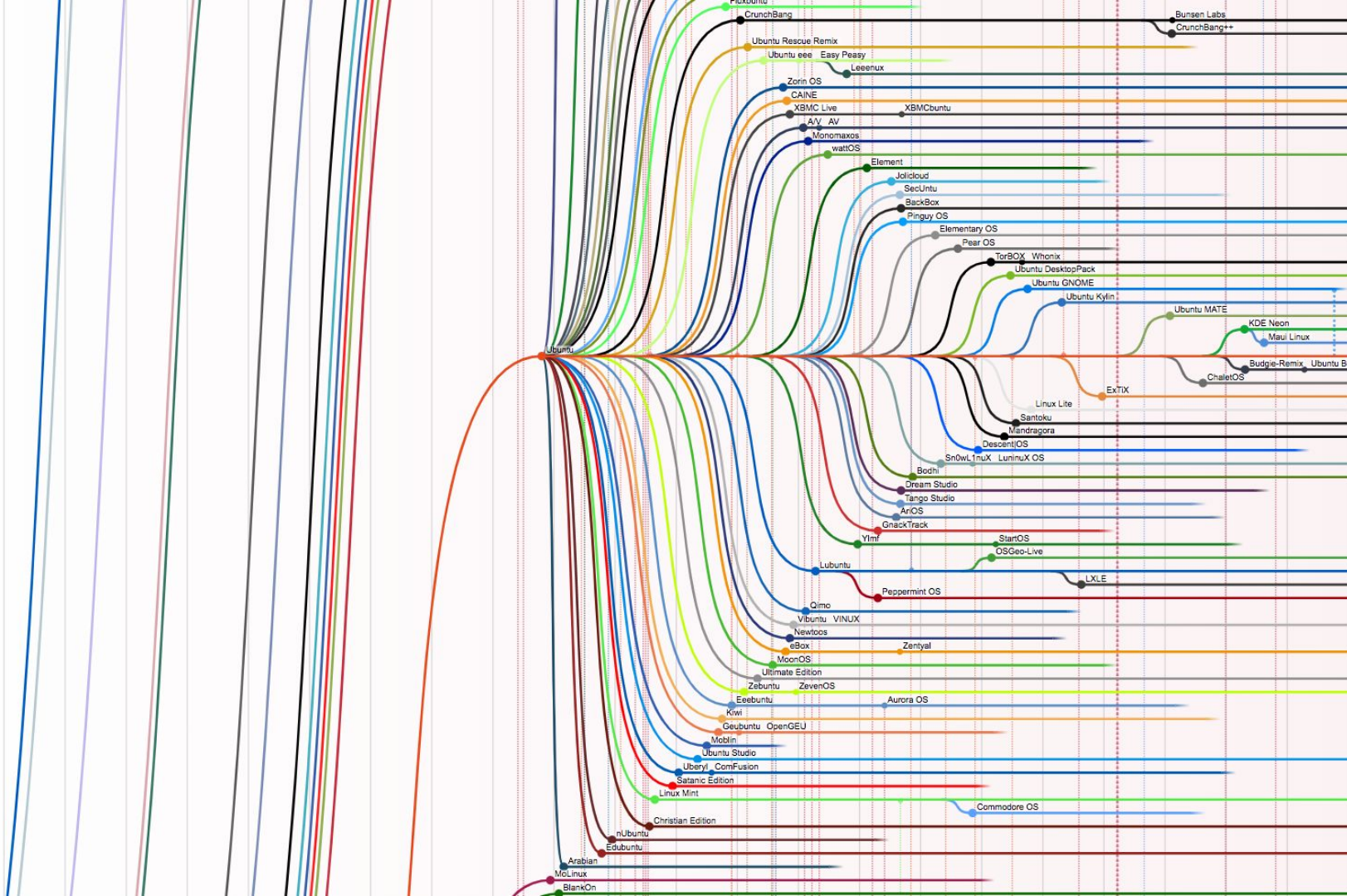
Version 17.10

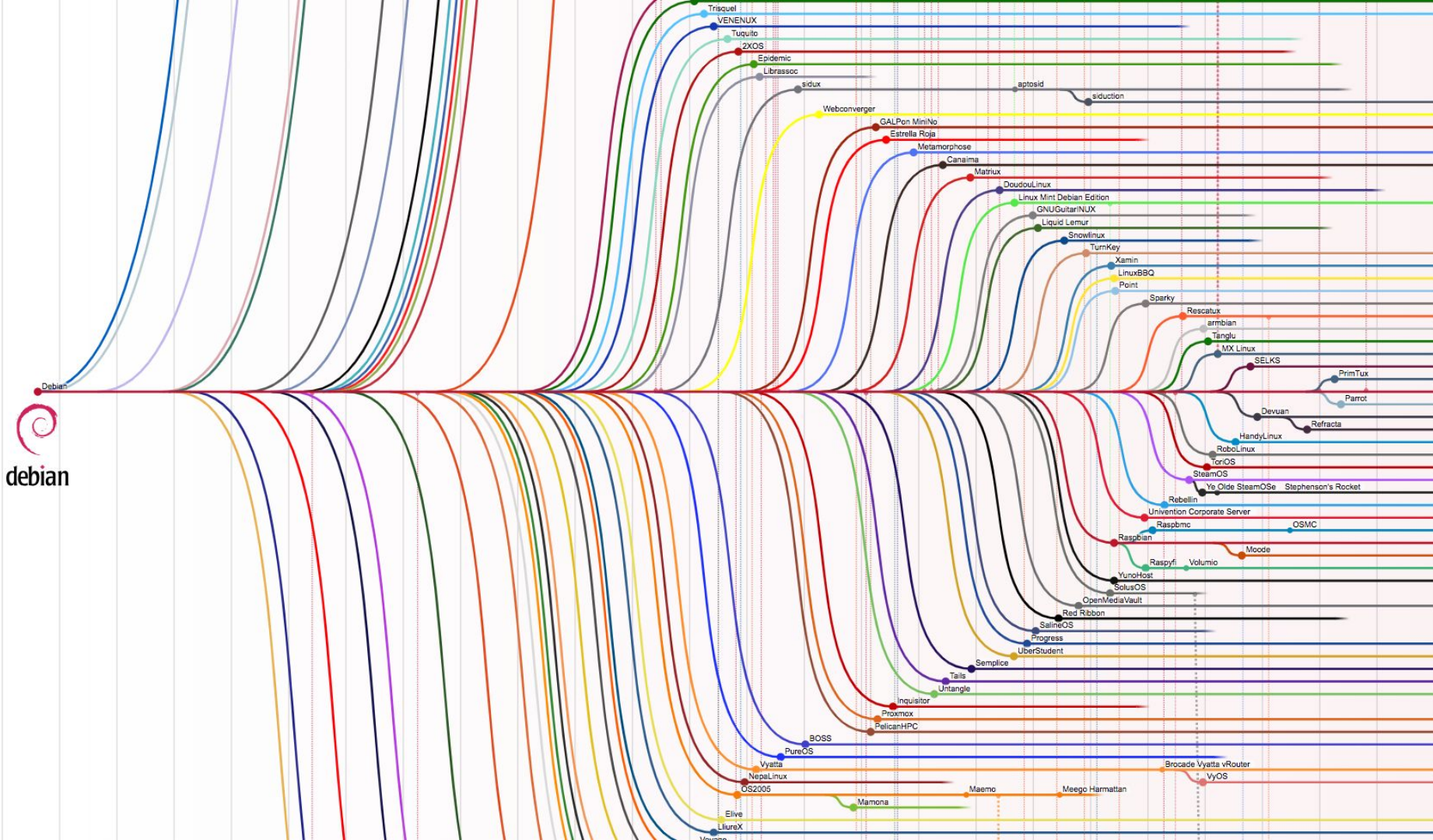
© Andreas Lundqvist, Donjan Rodic, Mohammed A. Mustafa
© Konimex, Fabio Loli and contributors
<https://github.com/FabioLoli/linuxtimeline>
Original source: futurist.se/gidt
Published under the GNU Free Documentation License

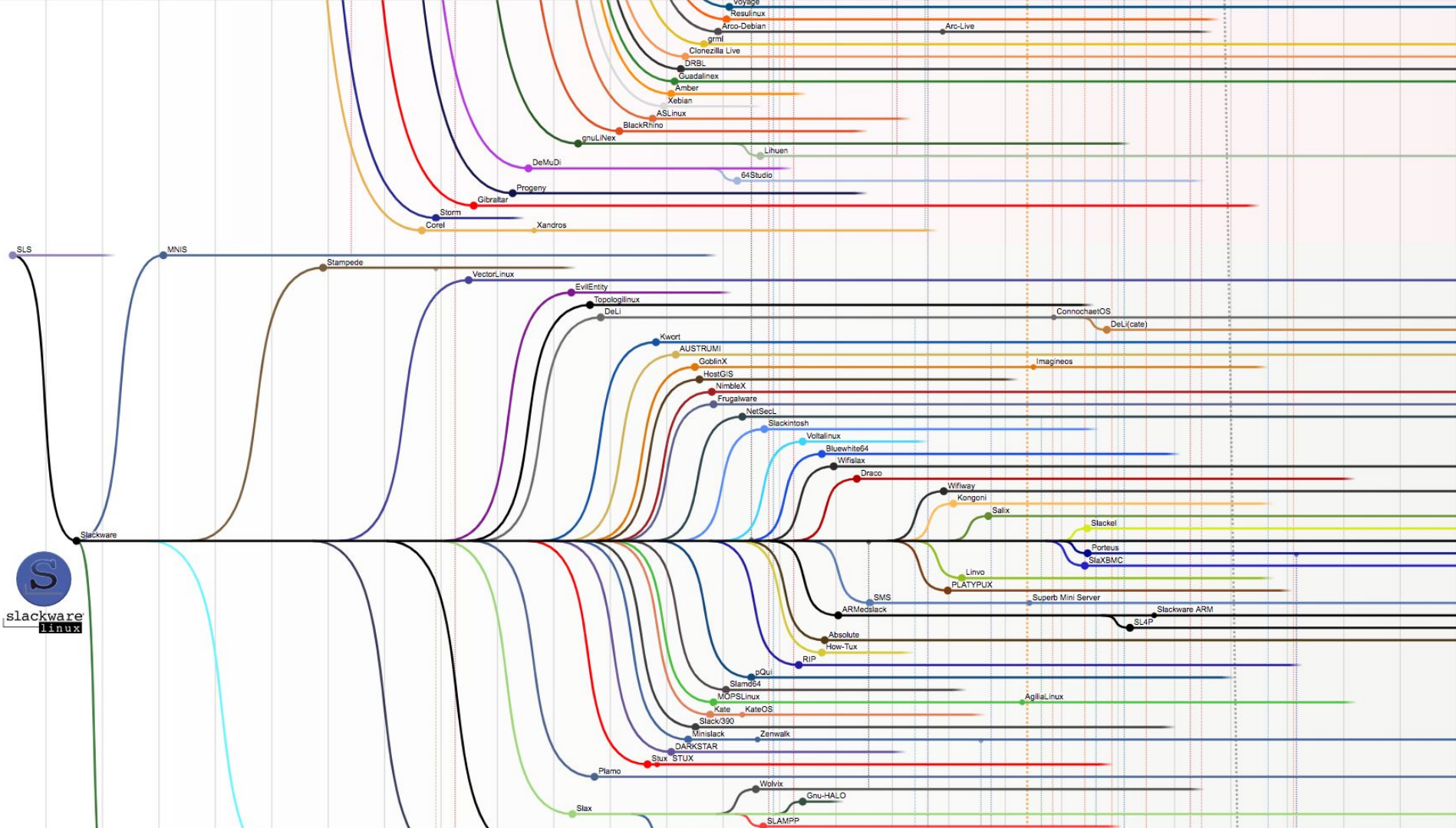


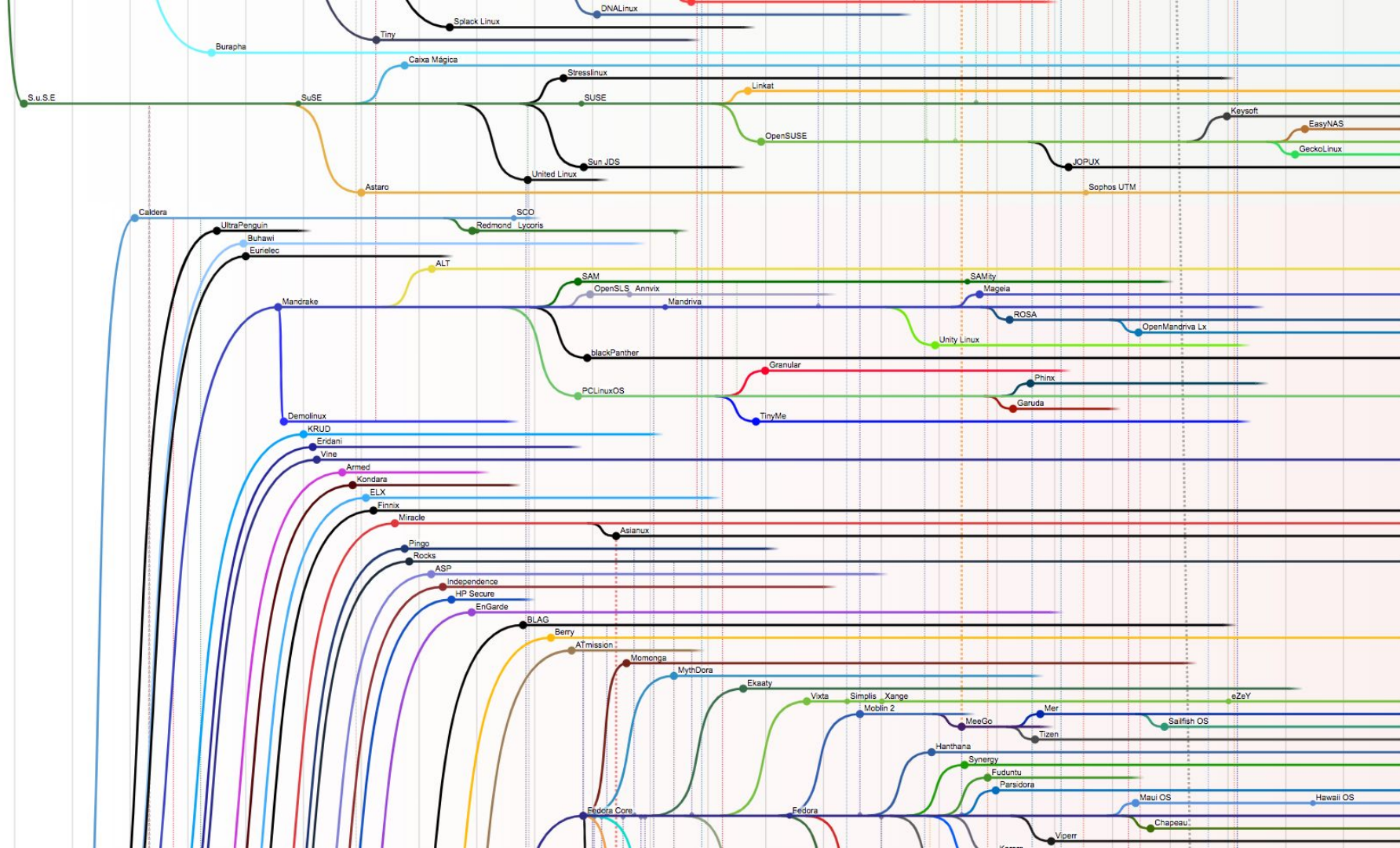
- Influence, developer switching
- Releasing, substantial code flow, project overtaking
- Developer & code sharing, project merging

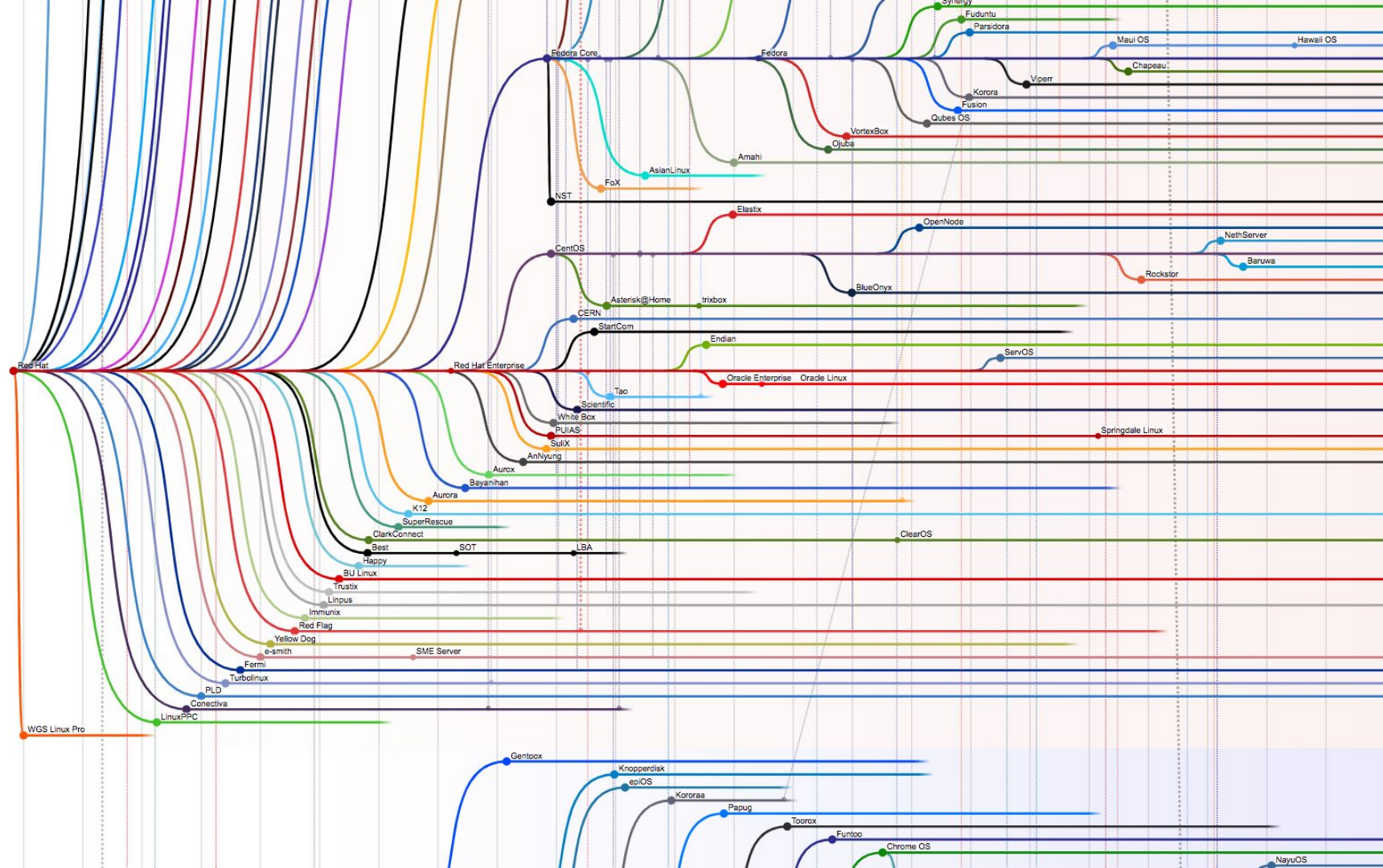


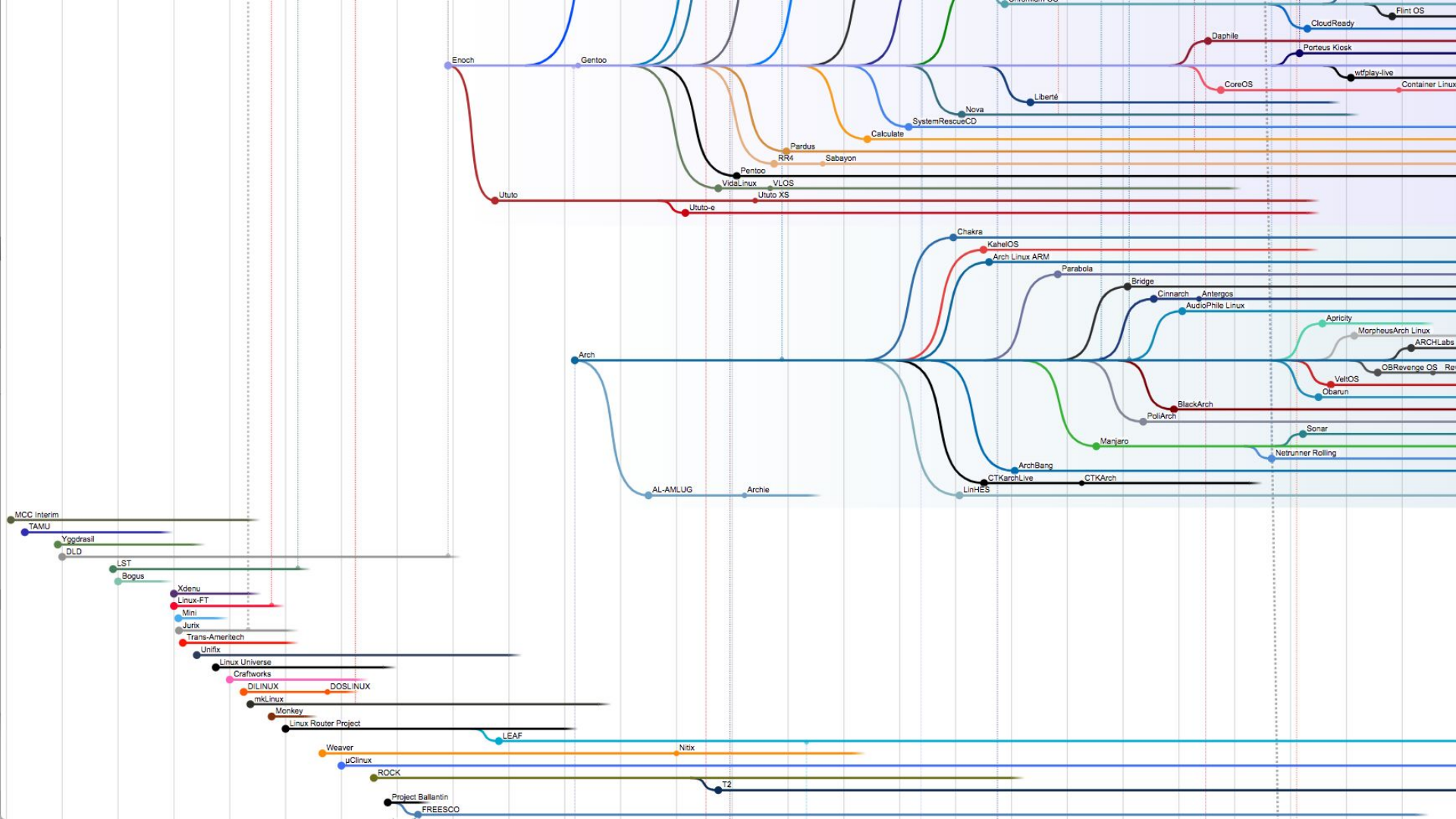


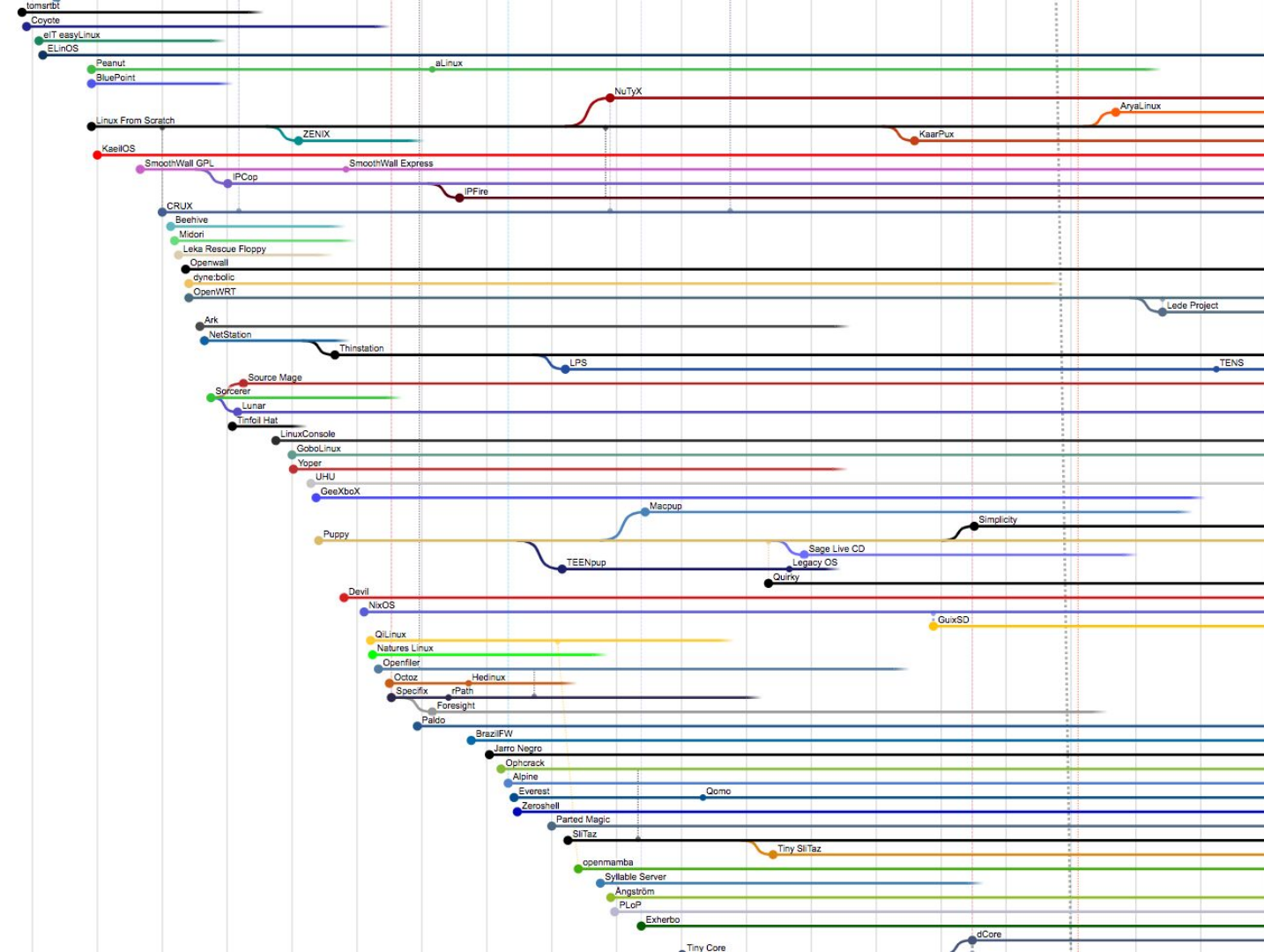


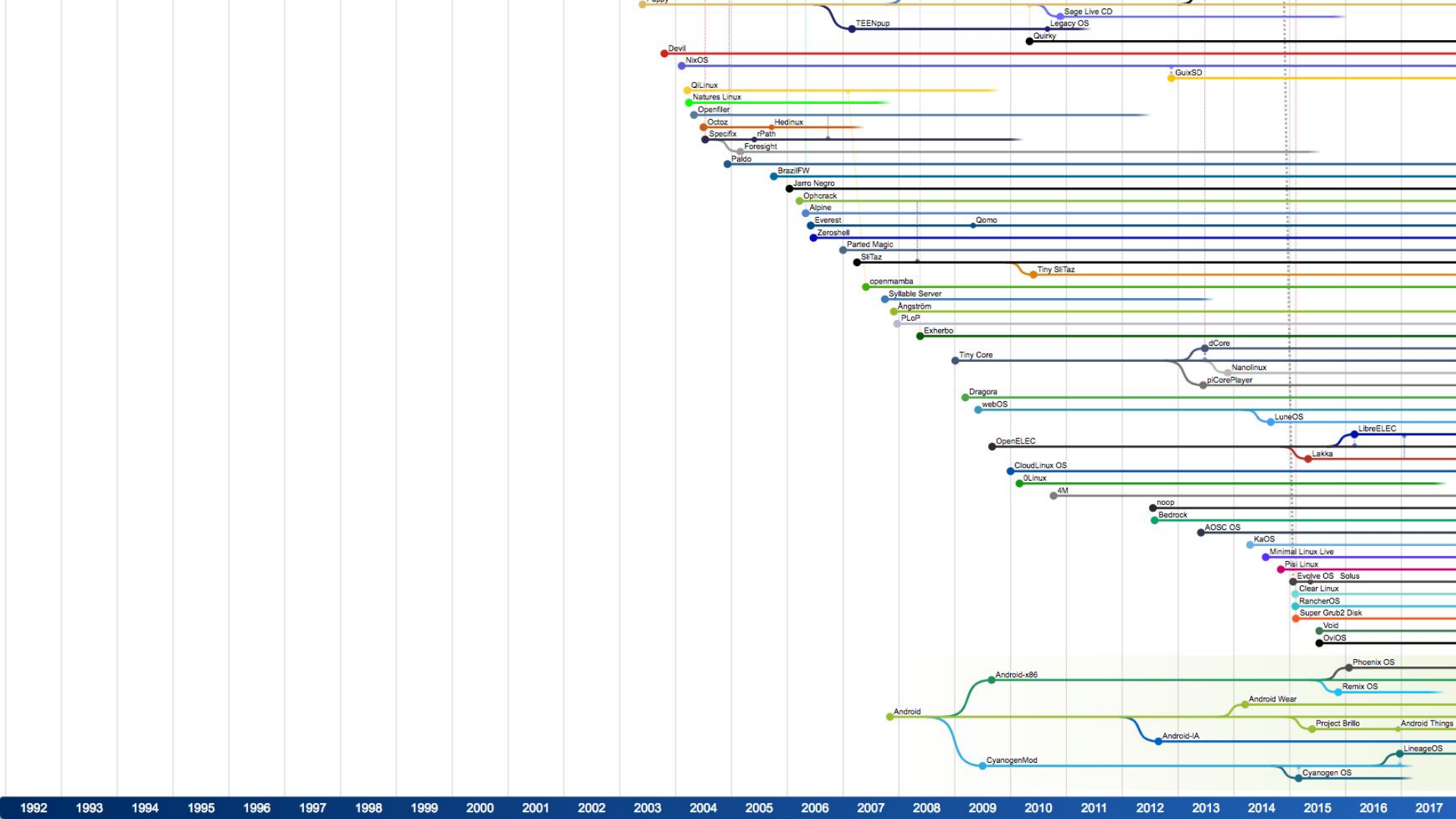












The Oregon Trail



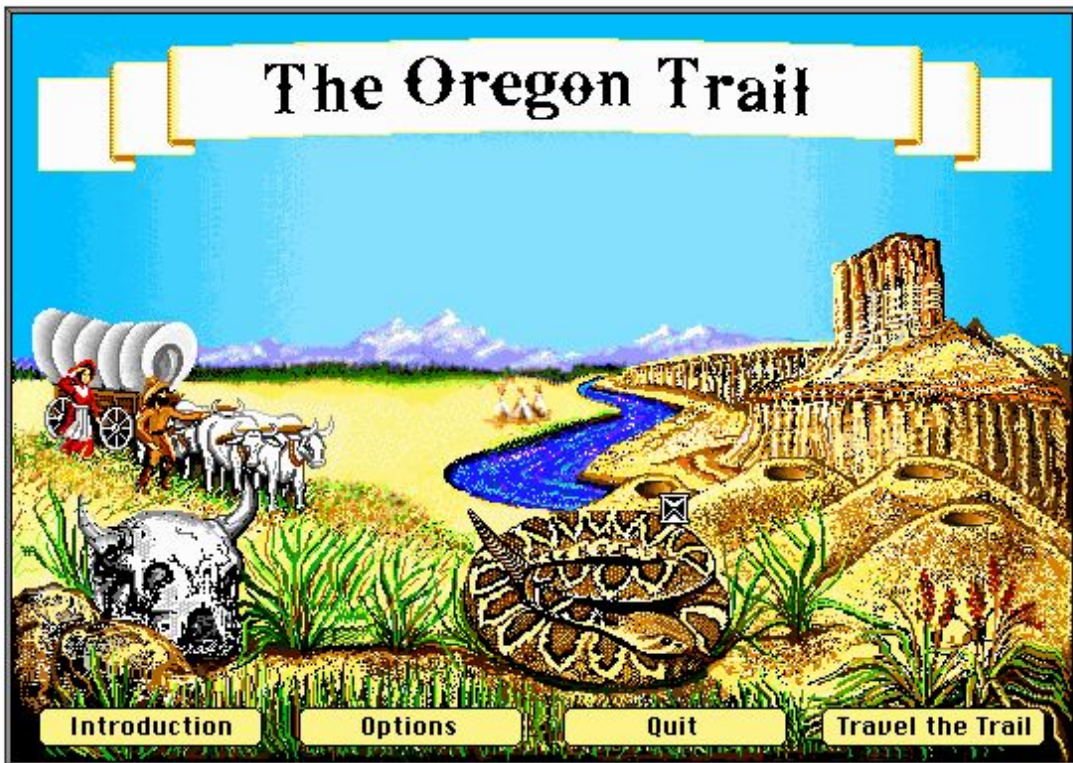
You may:

1. Travel the trail
2. Learn about the trail
3. See the Oregon Top Ten
4. Turn sound off

What is your choice?



The Oregon Trail





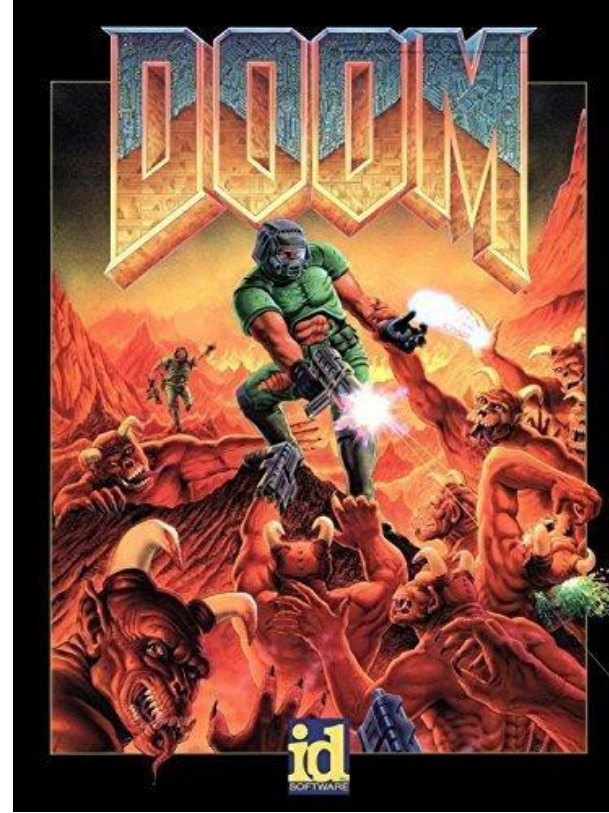
Instant Play

New Game

Endless telegraph

THE **OREGON TRAIL**

**Carnegie
Mellon
University**



Carnegie
Mellon
University

Overview of Workshop

1. Why Software Preservation?
2. What is Software Curation?
 - a. Scope and Description
 - b. Migration**
 - c. Reproduction and Access
 - d. Outreach and Culture
3. Trends
4. Efforts at CMU

Migration

Imaging and Storage

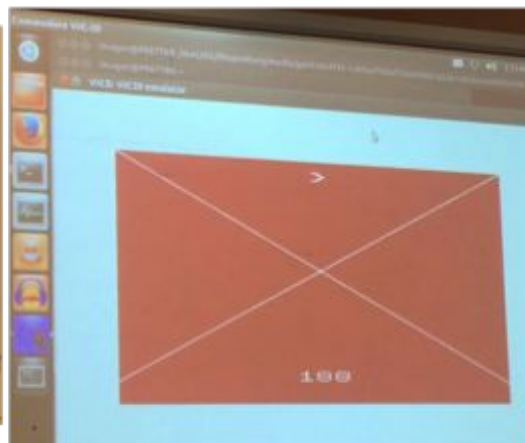
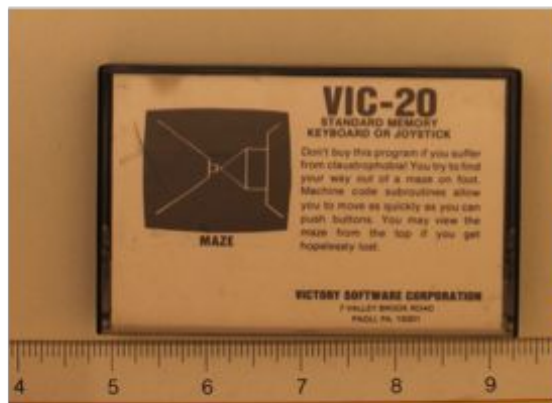
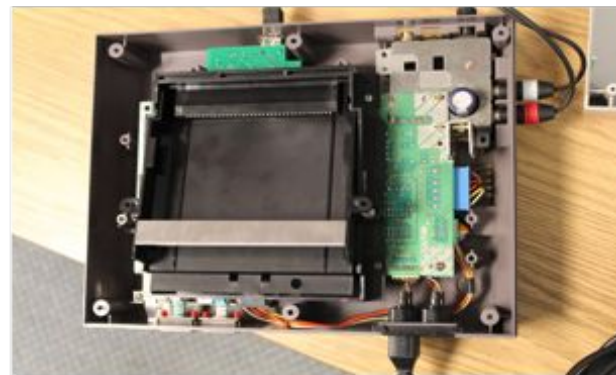
Digital Forensic Workflows

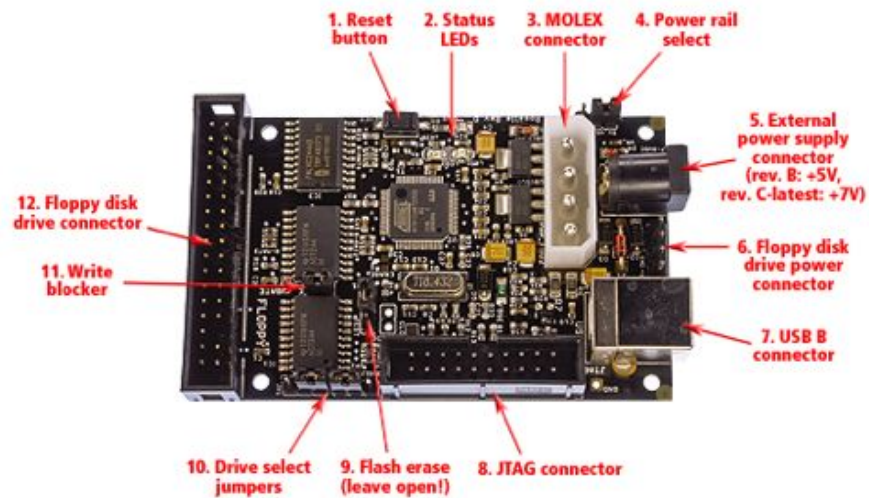
Digital Rights Management

Digital Migration Strategies

Migration



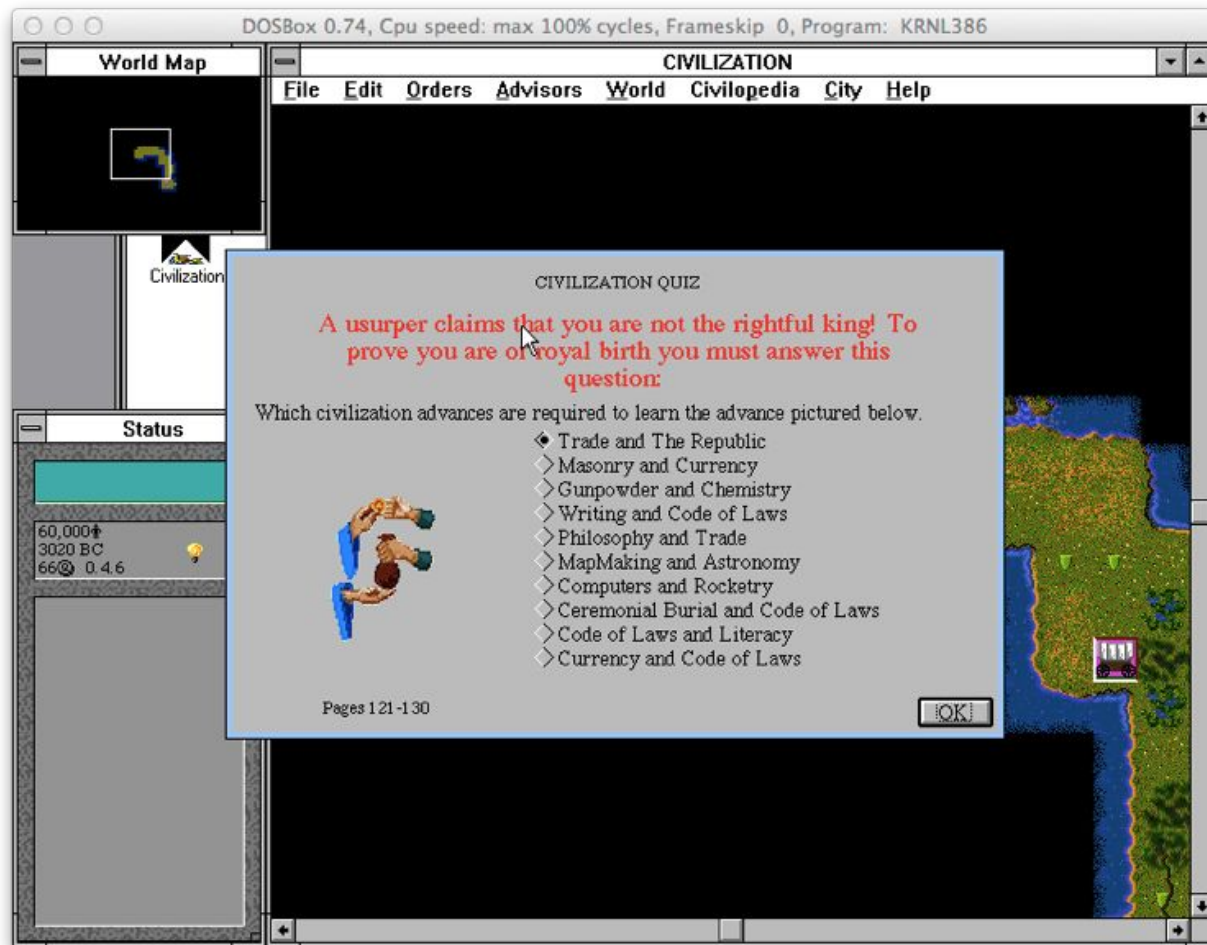




BitCurator

CONSORTIUM





Digital Migration Strategies

Once imaged or ingested, need a *forever* storage strategy

- Repositories change and evolve, data needs to migrate with those changes
- Provenance information becomes more important

Current repositories are still playing significant catch up

No current long-term solutions available for software data

Overview of Workshop

1. Why Software Preservation?
2. What is Software Curation?
 - a. Scope and Description
 - b. Migration
 - c. Reproduction and Access**
 - d. Outreach and Culture
3. Trends
4. Efforts at CMU

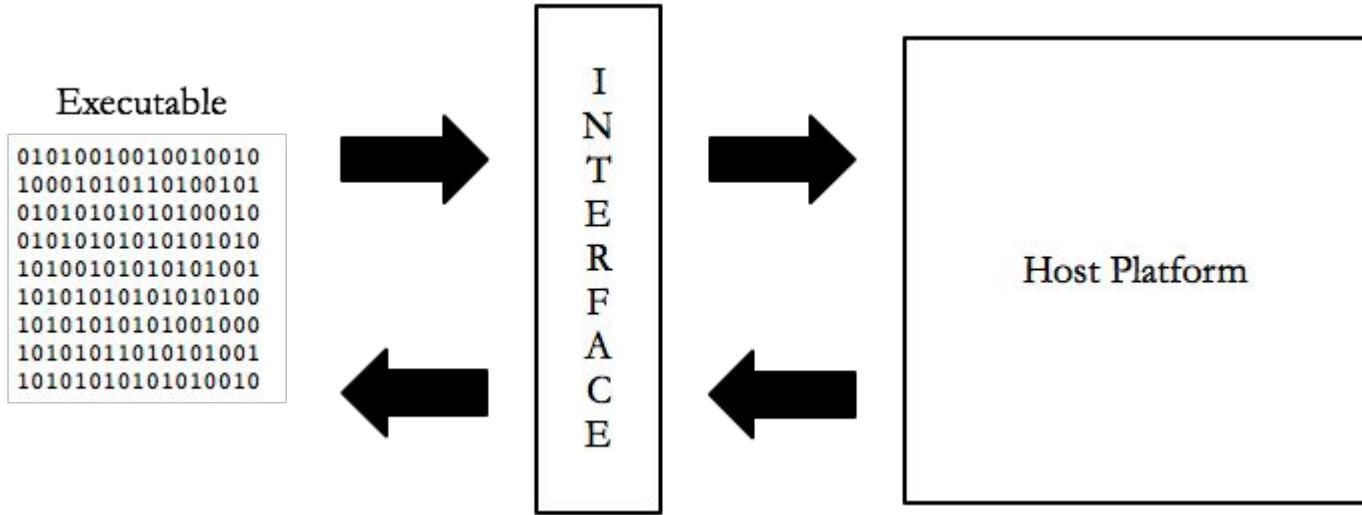
Reproduction Strategies

Virtualization

Containerization

Emulation

Hardware Preservation





Client Side Emulation

Local executable emulation

JavaScript in browser emulation

Pros:

- Lower latency
- Locally inspectable
- If browser based, shareable and single requirement

Cons:

- Dependent on specific system configuration
- Not easily shareable if local executable
- Legally dubious

Server Side Emulation

Emulation in cloud

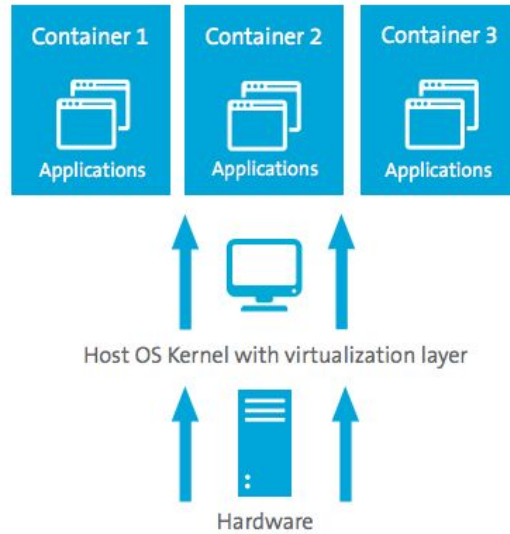
Pros:

- Legally more appealing
- Management and maintenance are centralized
- More easy to roll into services

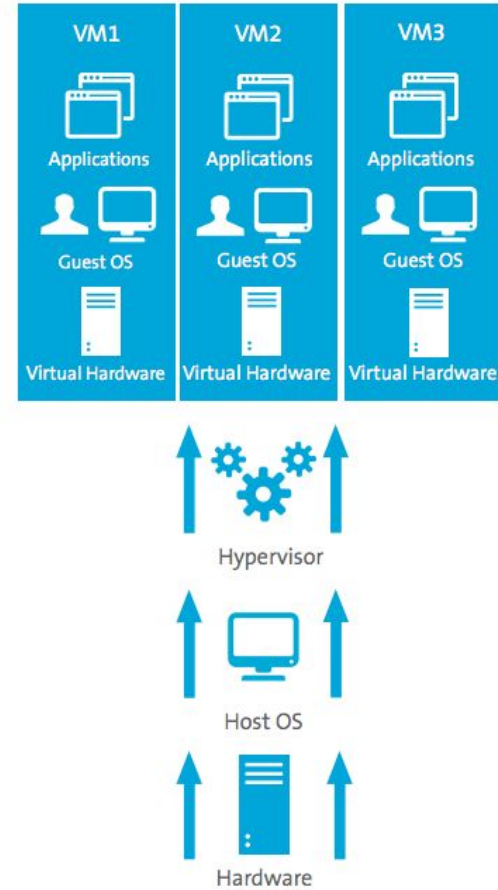
Cons:

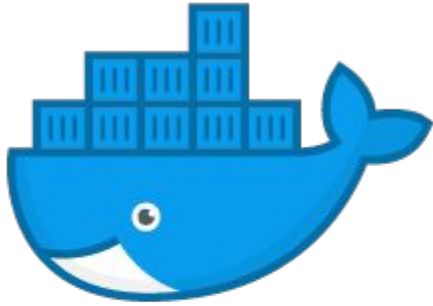
- Latency issues
- Much less introspection
- Cloud is a preservation issue in of itself

Virtualization using containers



Virtualization using hypervisors





docker



kubernetes



MESOS

Carnegie
Mellon
University



Hardware Preservation

Cultural software objects are designed for specific hardware and interactions

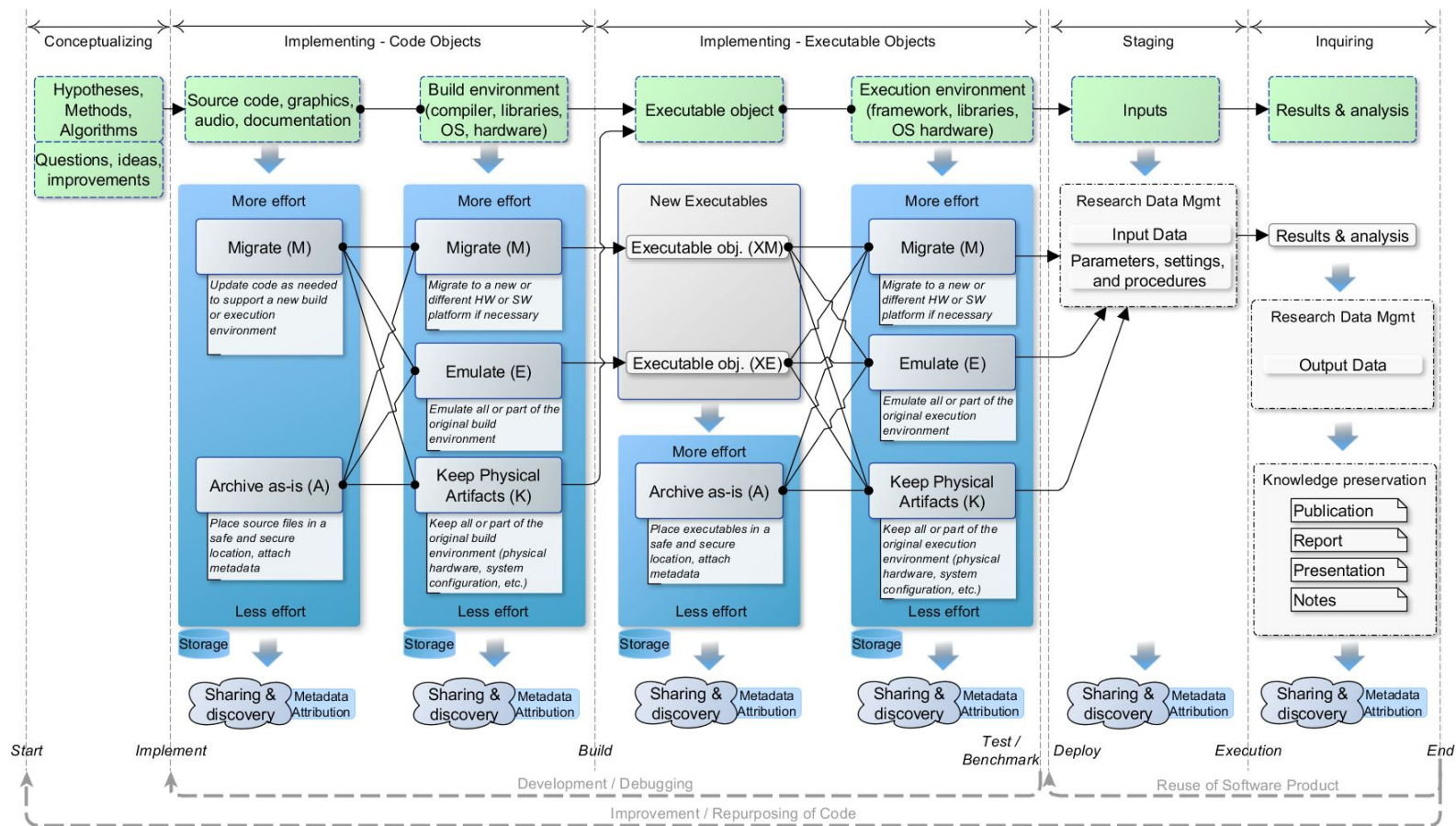
- Computer Games
- Interactive Art Installation
- Digital Art
- Other digital media works

Socio-cultural context is not in a VM or container

Hardware peripheral and displays are not replicated







Overview of Workshop

1. Why Software Preservation?
2. What is Software Curation?
 - a. Scope and Description
 - b. Migration
 - c. Reproduction and Access
 - d. Outreach and Culture
3. Trends
4. Efforts at CMU

Outreach

Locating significant materials around CMU community

Implement reproducible practices inside research labs
and departments

Sustainable software development

Overview of Workshop

1. Why Software Preservation?
2. What is Software Curation?
 - a. Scope and Description
 - b. Migration
 - c. Reproduction and Access
 - d. Outreach and Culture
3. Trends
4. Efforts at CMU

Sustainable Software Development

Testing oriented

Well documented

Instrumentation

- Continuous integration
- Package management
- Configuration management
- Version Control

Tools and Organizations

Reproducibility Tools Supporting Software

- Code Ocean - <http://codeocean.com>
- Occam - <http://occam.cs.pitt.edu>
- Collective Knowledge - <http://cknowledge.org/>
- Umbrella - <http://ccl.cse.nd.edu/software/umbrella/>
- ReproZip - <https://www.reprozip.org/>

Organizations supporting software preservation

- Software Sustainability Institute (SSI) - UK Organization - <http://www.software.ac.uk>
- Data and Software Preservation for Open Science (DASPOS) - CERN - <http://daspos.org>
- Software Preservation Network (SPN) - US Memory Institutions
 - <http://www.softwarepreservationnetwork.org>

ACM Reproducibility

- Repeatability (Same team, same experimental setup)
- Replicability (Different team, same experimental setup)
- Reproducibility (Different team, different experimental setup)

ACM Badging Levels



Artifacts Available

- Software and data present in publication are available for download and investigation

Artifacts Evaluated - Functional

- Software and data have been audited and validated as working

Artifacts Evaluated - Reusable

- Software and data have been audited by a third party, are functional, and significantly oriented toward reusability through documentation, code / software organization, etc.

Results Replicated

- Software and data have been used to validate results

Results Reproduced

- Different software and data have been used to validate results

Overview of Workshop

1. Why Software Preservation?
2. What is Software Curation?
 - a. Scope and Description
 - b. Migration
 - c. Reproduction and Access
 - d. Outreach and Culture
3. Trends
4. Efforts at CMU

Tools and Services at CMU

Kilthub Repository

- Source code
- Research data and software executable binaries

Code Ocean (Beta)

- Targeted for August 2019
- Reproducibility platform

Emulation as a Service (EaaS)

- Currently in research beta
- Distributed containerized execution contexts

Software and Data Carpentries

- Two day courses on basic research support tools like Python, R, and Git

History of Science and Technology at CMU (HOST@CMU)

- Interdisciplinary initiative to locate and celebrate CMU technical history

KiltHub

Discover research from Carnegie Mellon University ▼



+ Follow

NEW

POPULAR

CATEGORIES

SEARCH Q

19,165 posts

1,750,937 views

7,196,680 downloads

more stats...



Spiritual Narratives in Beethoven's Quartet, Op. 132

John Ito

22/03/2019



Koch's Metrical Theory and Mozart's Music: A Corpus Study

John Ito

22/03/2019



Income Mobility in America

Manu Navjeevan

21/03/2019




Monocular Facilitation Implicates Subcortical Involvement in Holistic...

Rebeka C. Almasi

21/03/2019



Code Ocean

 CODE OCEAN
BETA

DASHBOARDEXPLOREHELP

C++FreeType and PDF testsMetadataCodeInterfaceCollaborate

Code<

pdfcode.c

>Results

makeaname.cc

pcg_basic.c

pcg_basic.h

pdfcode.c

random.c

README.md

run.sh

Data

amiri-bold.ttf

596.7 KB

```
17 #define MIN_HEIGHT 500
18
19 /*Random number functions contained in random.c*/
20 extern char* make_random_filename(void);
21 extern char* make_fname(char* fpath, char*fname, char* extensi
22 extern void make_random_integer(unsigned int* radius, unsigned
23 extern void make_random_point(int*x, int*y, int pdfwidth, int
24 extern void make_random_colour(unsigned int* red, unsigned int
25 extern void make_bounded_random_integer(unsigned int* value, u
26
27 void makepdf (int pdfwidth, int pdfheight);
28
29 void makepdf (int pdfwidth, int pdfheight)
30 {
31     unsigned int i;
32     unsigned int j;
```

Output - Run 5817394

Run environment setup complete, running algorithm...

Search

M.png

40.09 KB

m.png

33.47 KB

N.png

32.18 KB

n.png

21.84 KB

O.png

24.06 KB

o.png

15.73 KB

output.pdf

P.png

p.png

20.69 KB

Q.png

28.1 KB

q.png

23.99 KB

R.png

28.4 KB

r.png

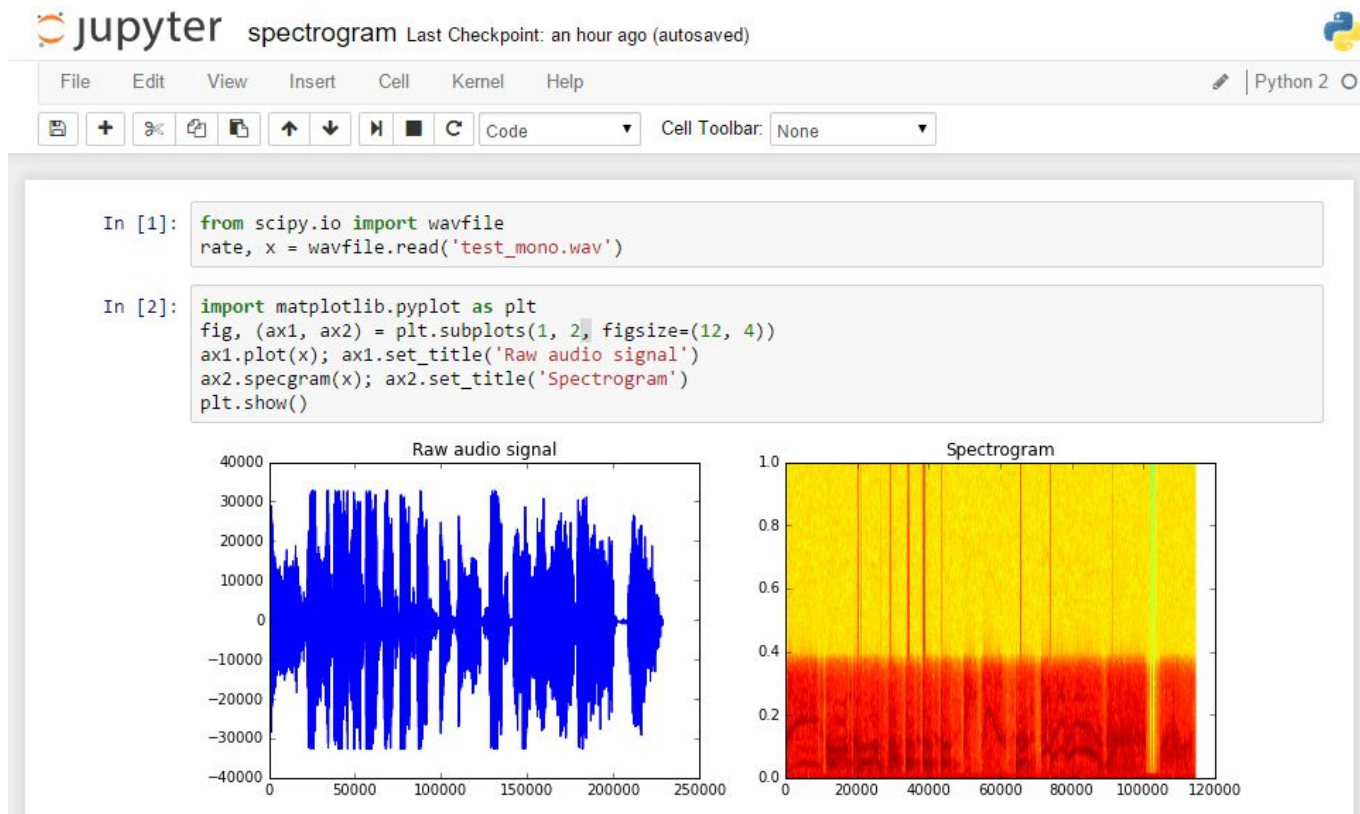
12.29 KB

S.png

20.94 KB

Download

Jupyter Notebooks



Emulation as a Service Infrastructure

Interuniversity Distributed Network of
Emulation Nodes

Six partner institutions

Environment Contexts with full software
installation

Access to legacy files and objects



Questions/Comments

Eric Kaltman

Data Curation Fellow

ekaltman@andrew.cmu.edu



**Carnegie
Mellon
University**