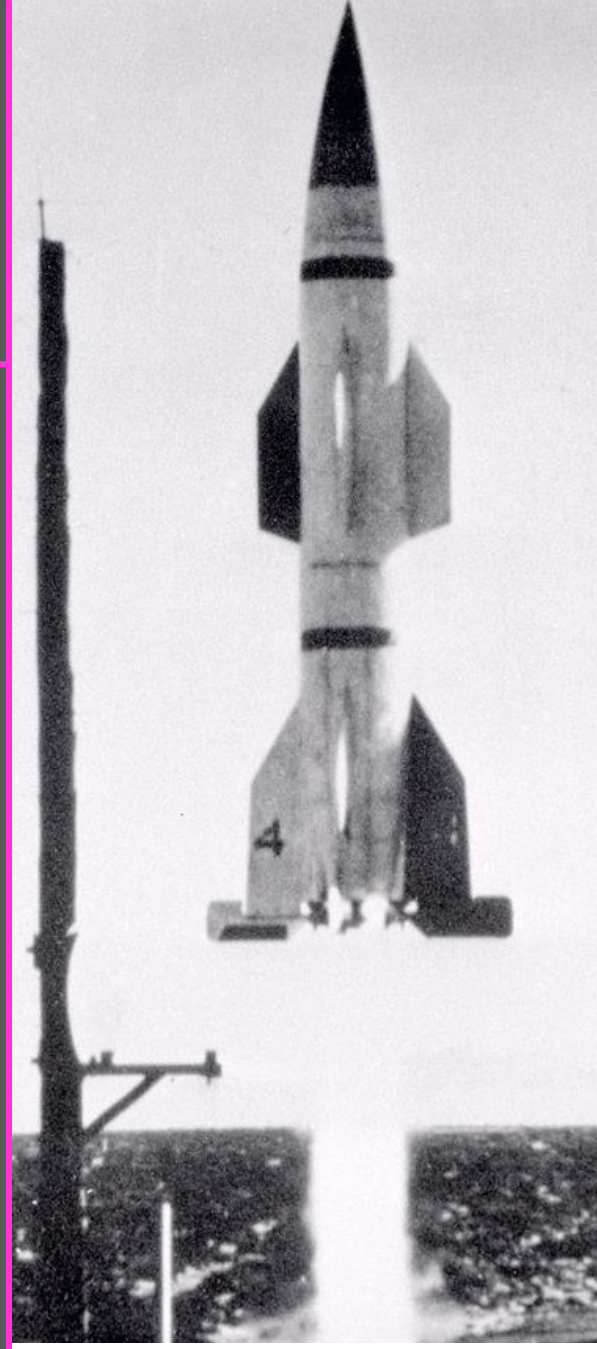
Technologie Proche:

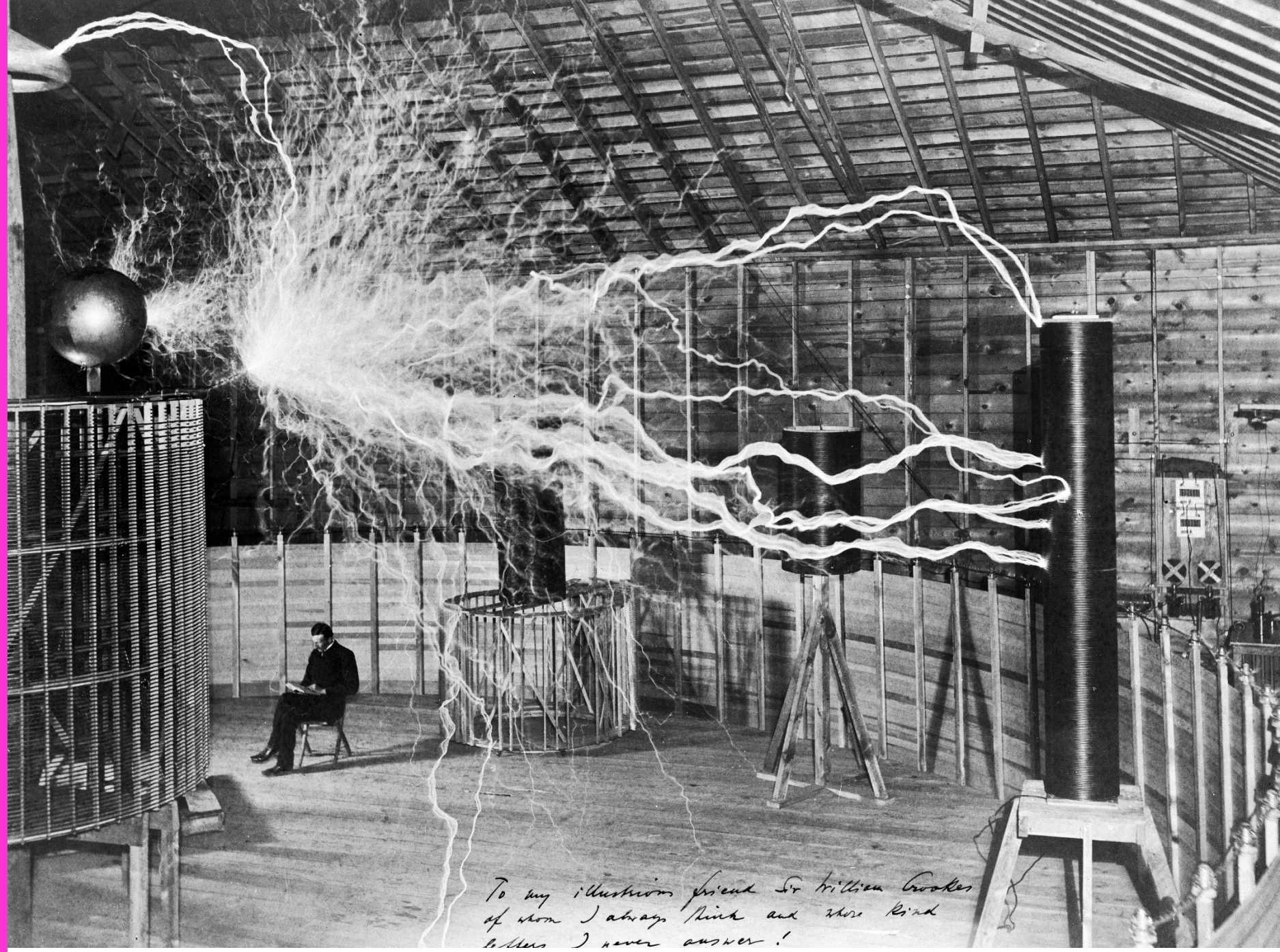
Envisioning the Archival Systems of Tomorrow with the Tools of Today

Dan Gillean – Artefactual Systems
ACA Conference 2016 – Montréal, QC



https://en.wikipedia.org/wiki/Hermes_%28missile_program%29#/media/File:Hermes_A-1_Test_Rockets_-_GPN-2000-000063.jpg

This is going to be a bit of an experiment ….



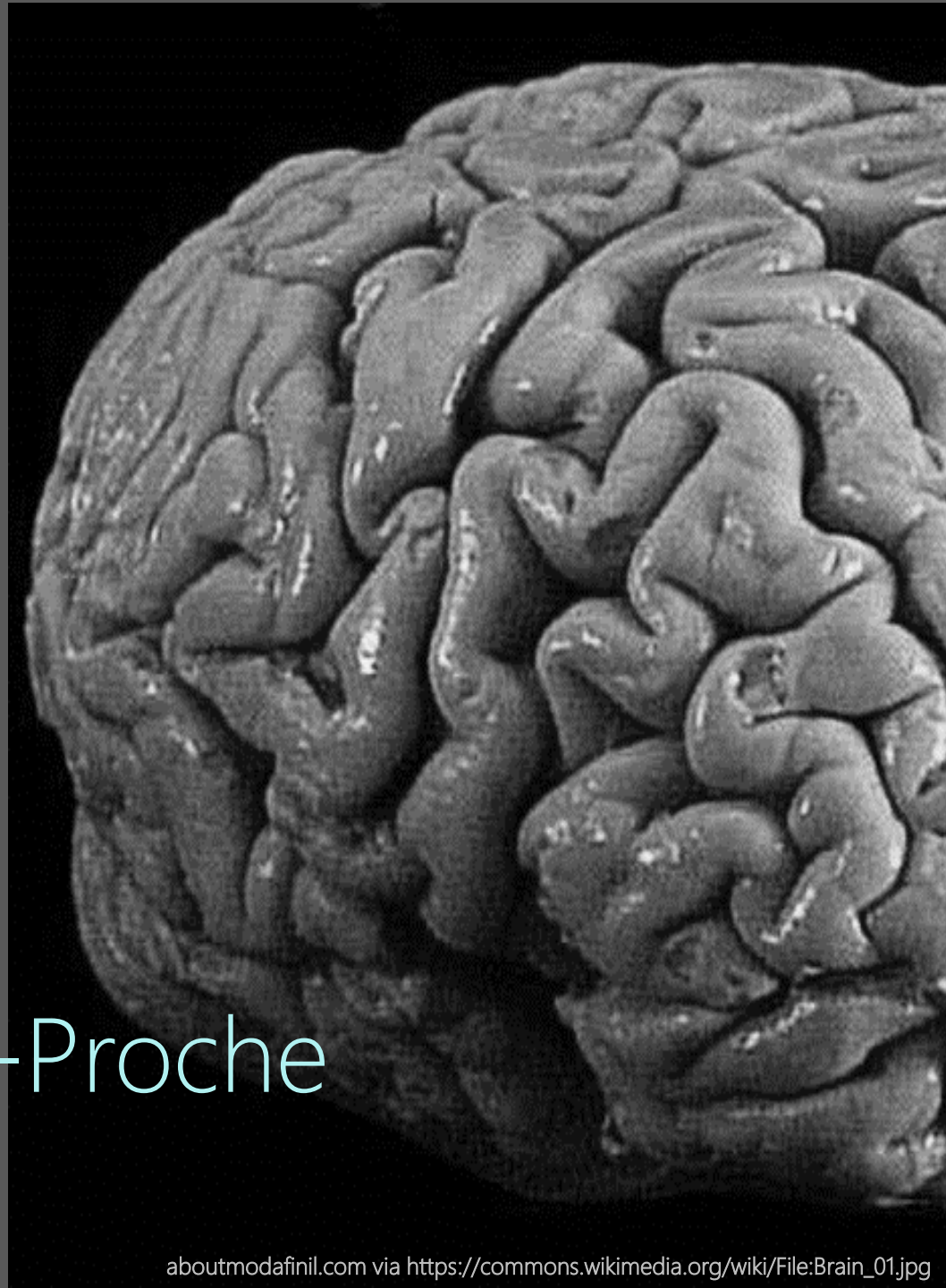
https://en.wikipedia.org/wiki/Nikola_Tesla#/media/File:Nikola_Tesla,_with_his_equipment_Wellcome_M0014782.jpg

Group Brainstorm:

What projects, technologies, or initiatives from the last couple of years have **inspired you** around archives and technology? What are you anticipating?

Google doc: http://bit.ly/tech-Proche

Twitter: #techProche



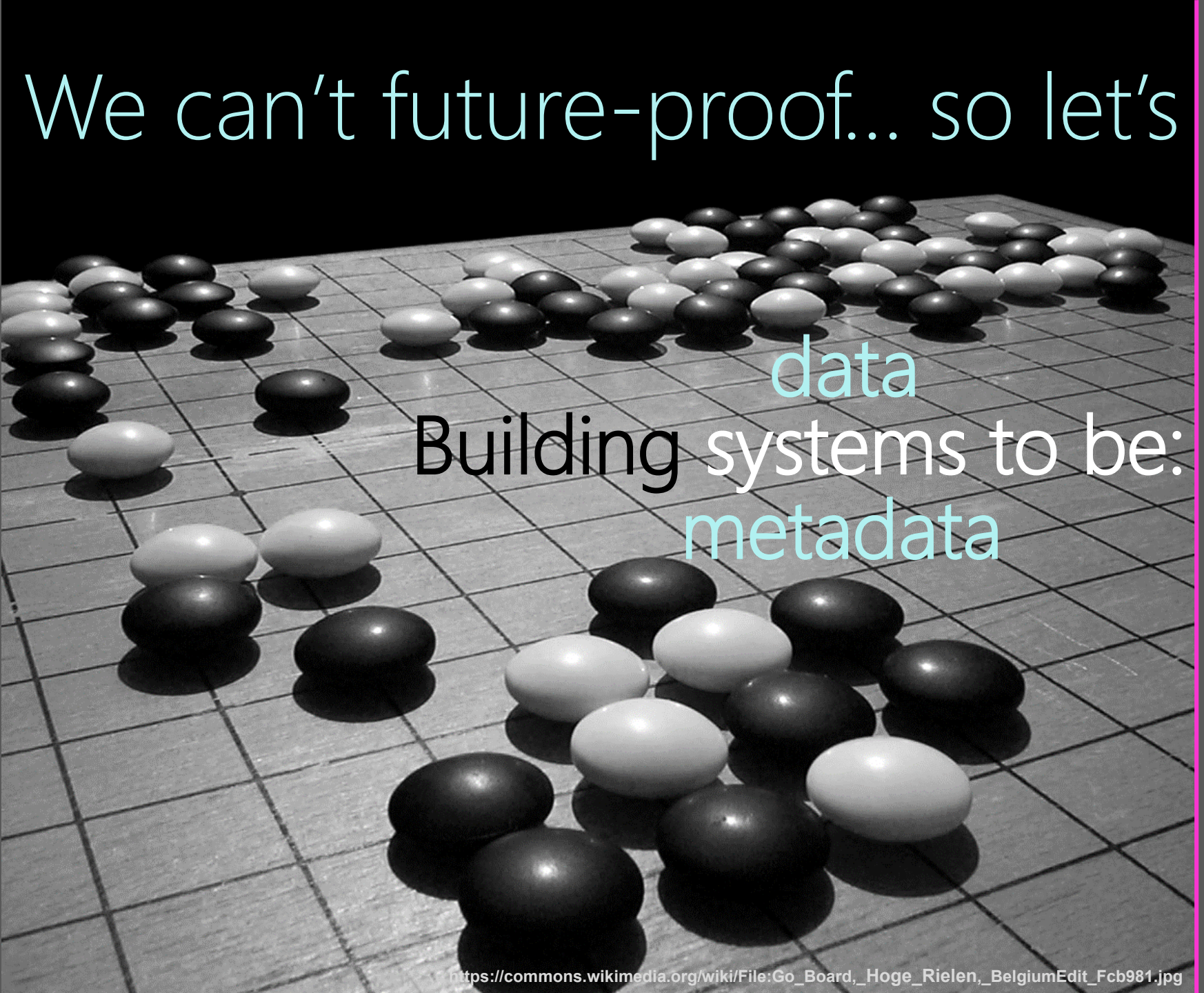
aboutmodafinil.com via https://commons.wikimedia.org/wiki/File:Brain_01.jpg


I HAVE NO IDEA
WHAT I'M DOING
http://knowyourmeme.com/memes/i-have-no-idea-what-i-m-doing

Change is inevitable

https://en.wikipedia.org/wiki/Analog_computer#/media/File:Analog_Computing_Machine_GPN-2000-000354.jpg

We can't future-proof… so let's be strategic



Building systems to be:

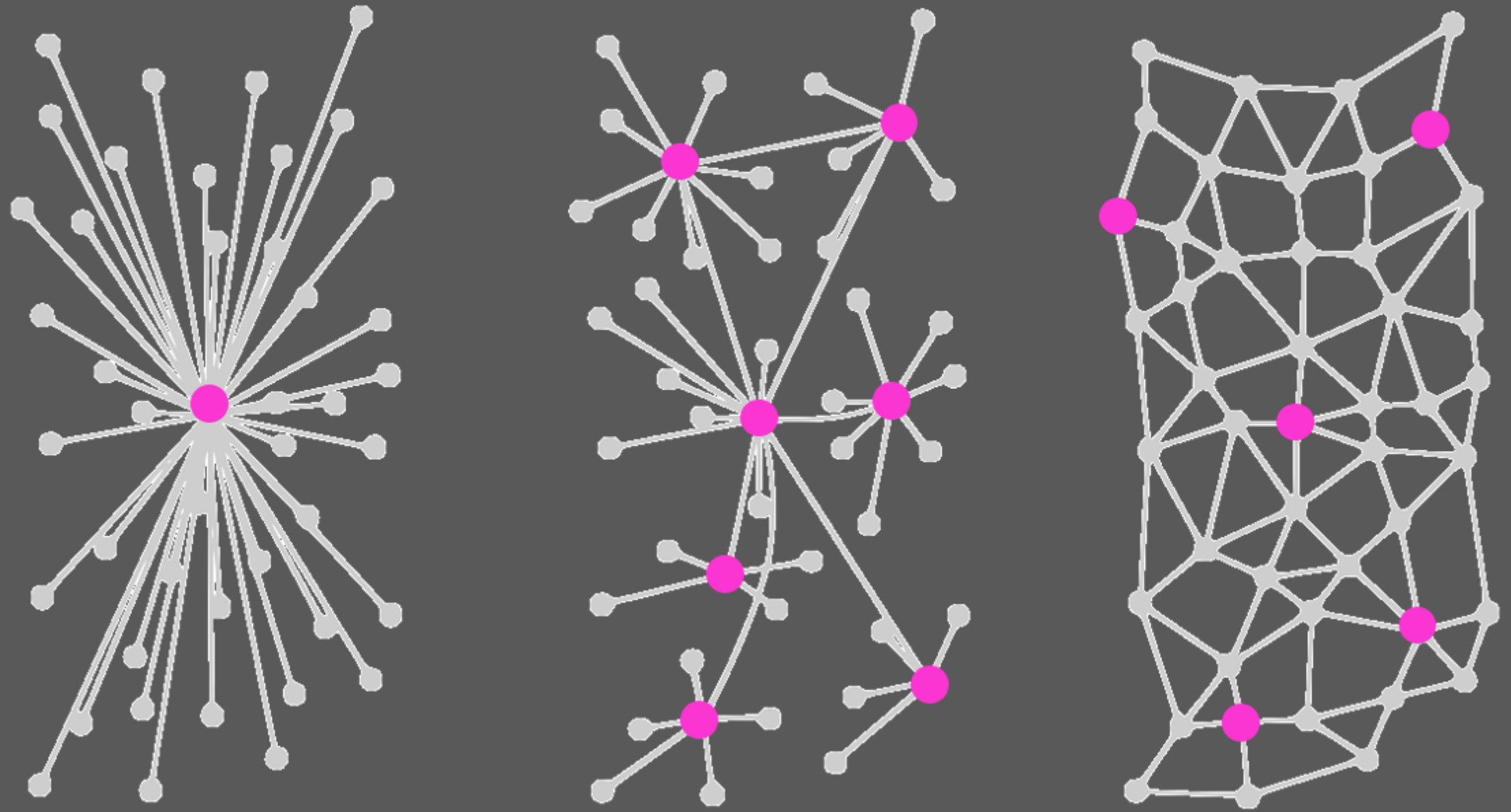
data

metadata

- Open
- Collaborative
- Decentralized
- Agnostic
- Interoperable

https://commons.wikimedia.org/wiki/File:Go_Board,_Hoge_Rielen,_BelgiumEdit_Fcb981.jpg

Centralized, Decentralized, and Distributed



http://www.truthcoin.info/blog/measuring-decentralization/

We need to leverage what already exists

- Open source
- Open standards
- Existing communities

http://www.totallylocalvc.com/art-of-teamwork/

The Tools of Today

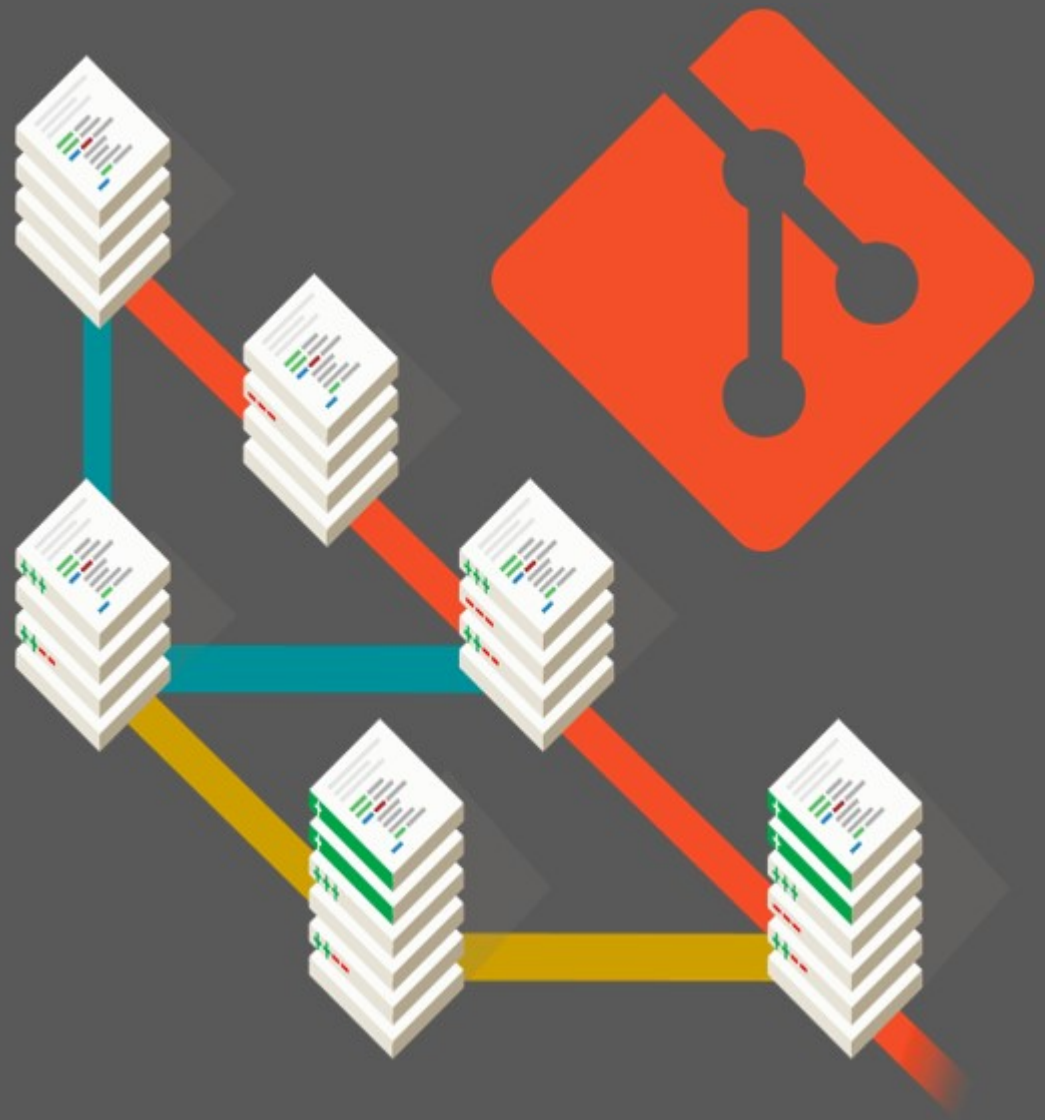
Rusty tools; Biser Todorov via https://commons.wikimedia.org/wiki/File:Rusty_tools.JPG

Distributed version control and git

Version control:

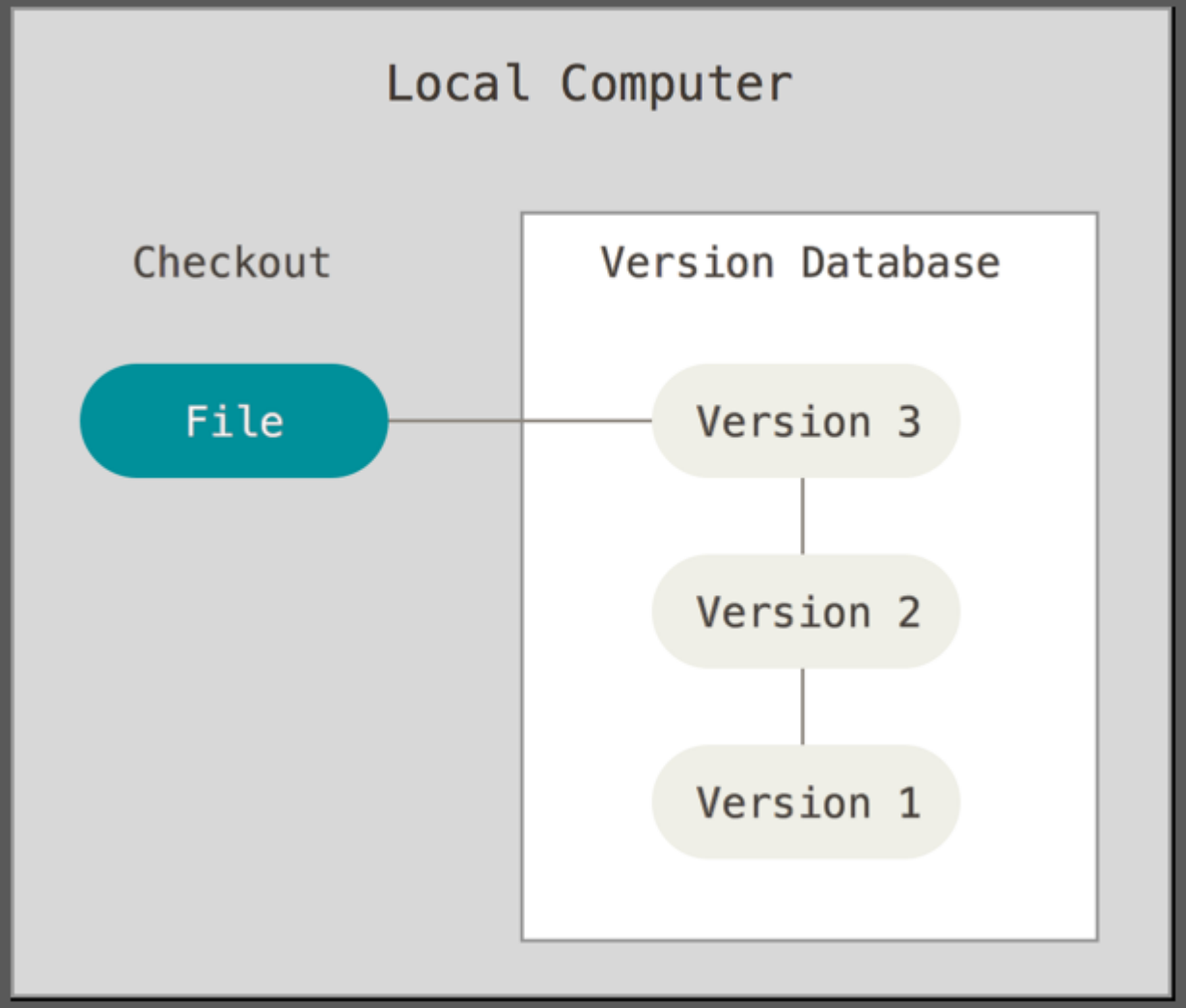
A system for managing changes to source data (documents, code, etc) over time.

Version control systems help you track who made what change when, and why. They also allow you to roll back changes to your data to a previous state.



https://git-scm.com/

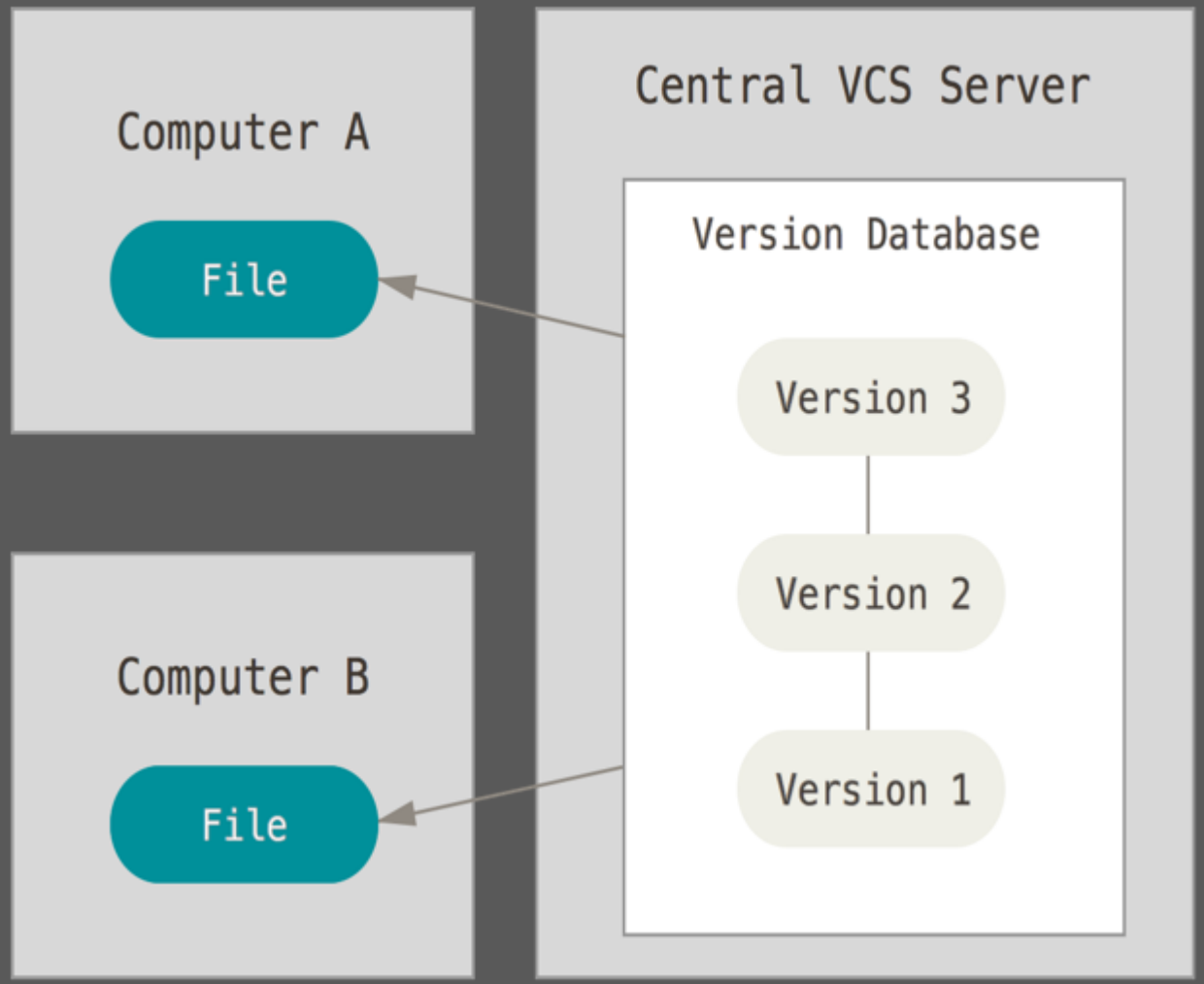
Local version control



- Changes are tracked on a local computer
- Simple – great for working alone
- Not effective for collaboration

https://git-scm.com/book/en/v2/Getting-Started-About-Version-Control

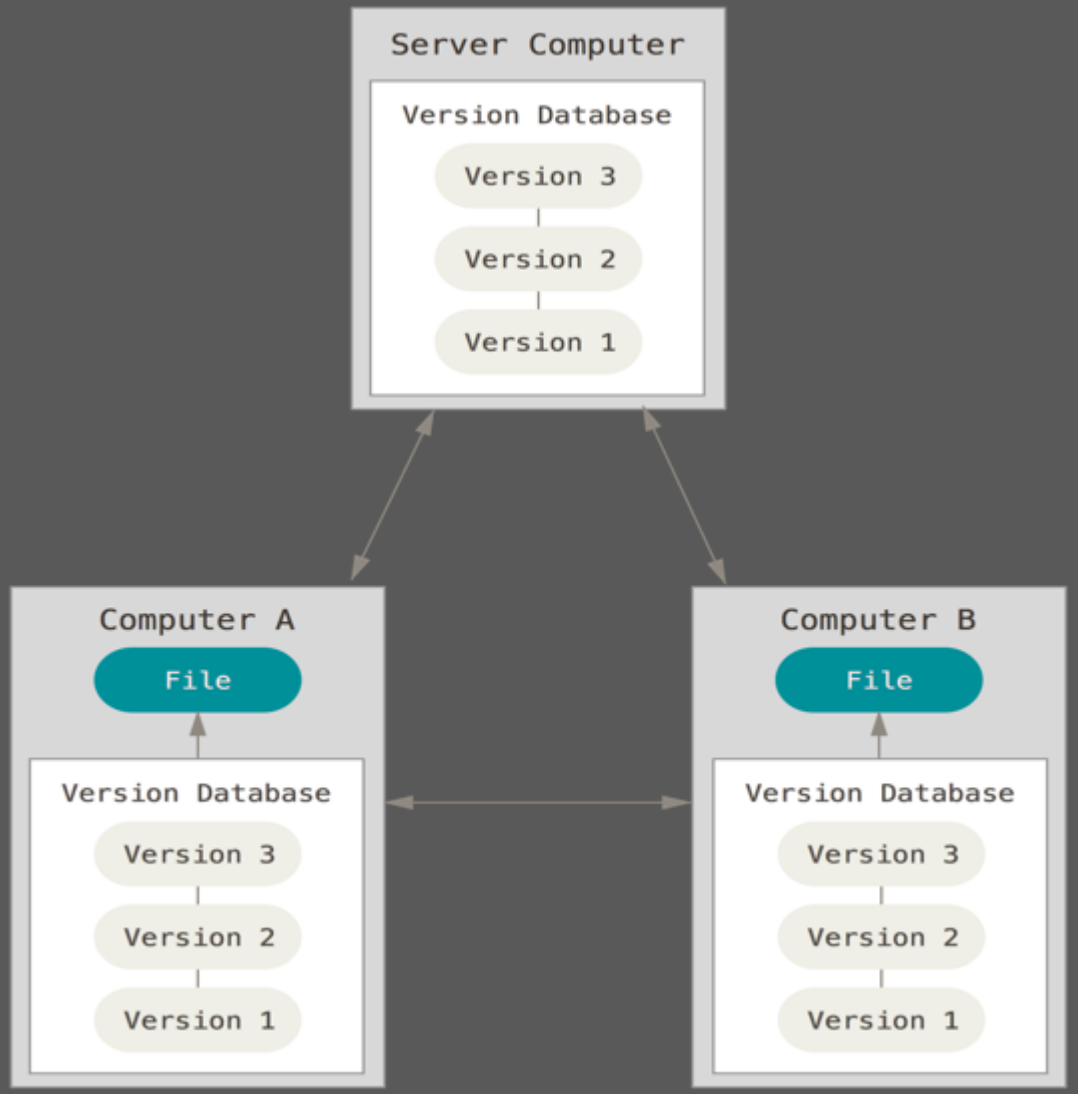
Centralized version control



- Single server with all versioned files; users "check out" files
- Allows for collaboration
- Single point of failure

https://git-scm.com/book/en/v2/Getting-Started-About-Version-Control

Distributed version control

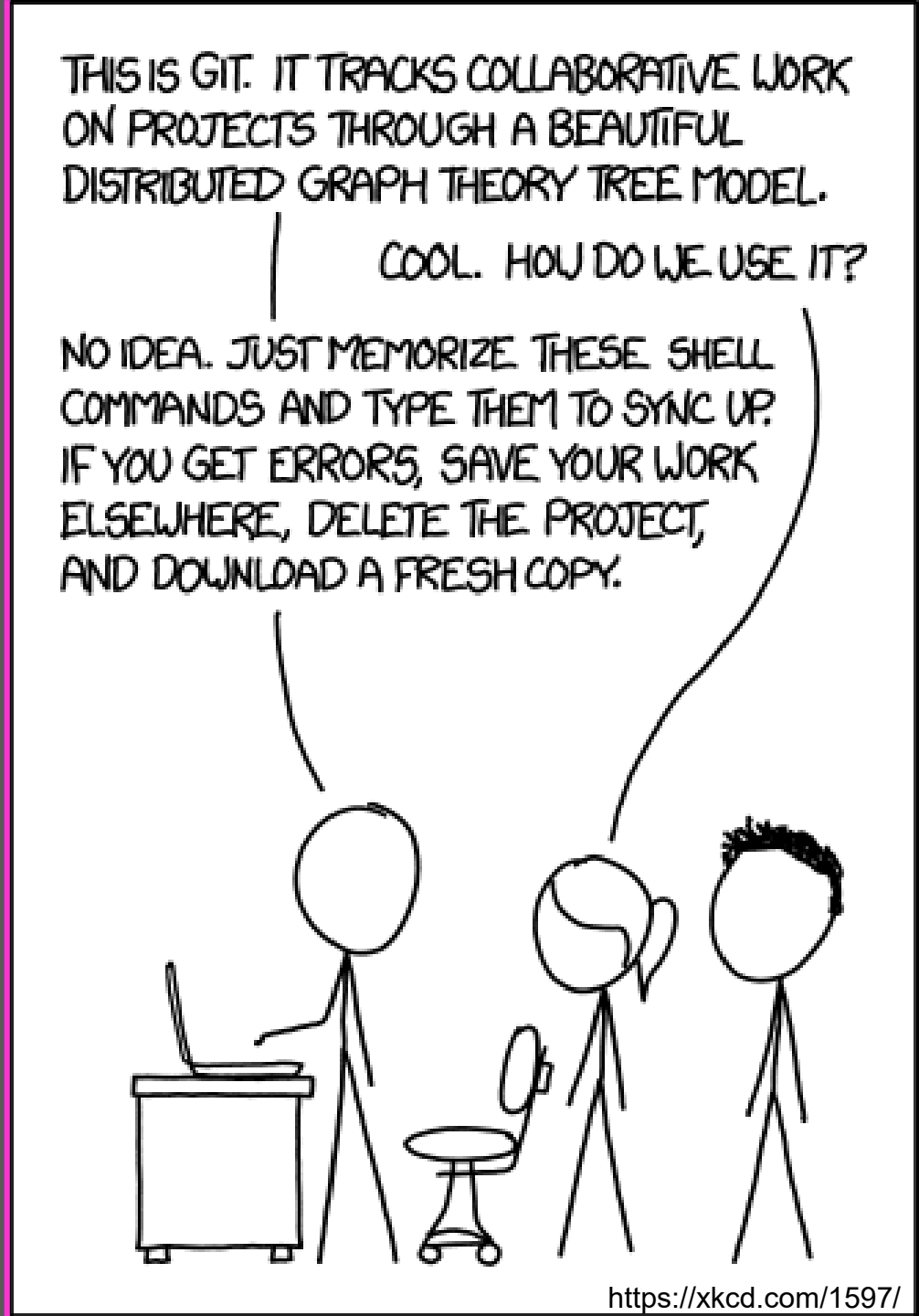


- Full mirroring of source with each user
- Broadest potential for collaboration
- Complex merging, branching, etc. possible

https://git-scm.com/book/en/v2/Getting-Started-About-Version-Control

Git

- Open source
- Distributed model
- Strong User Community, lots of free documentation
- Widely used and supported across platforms



Git

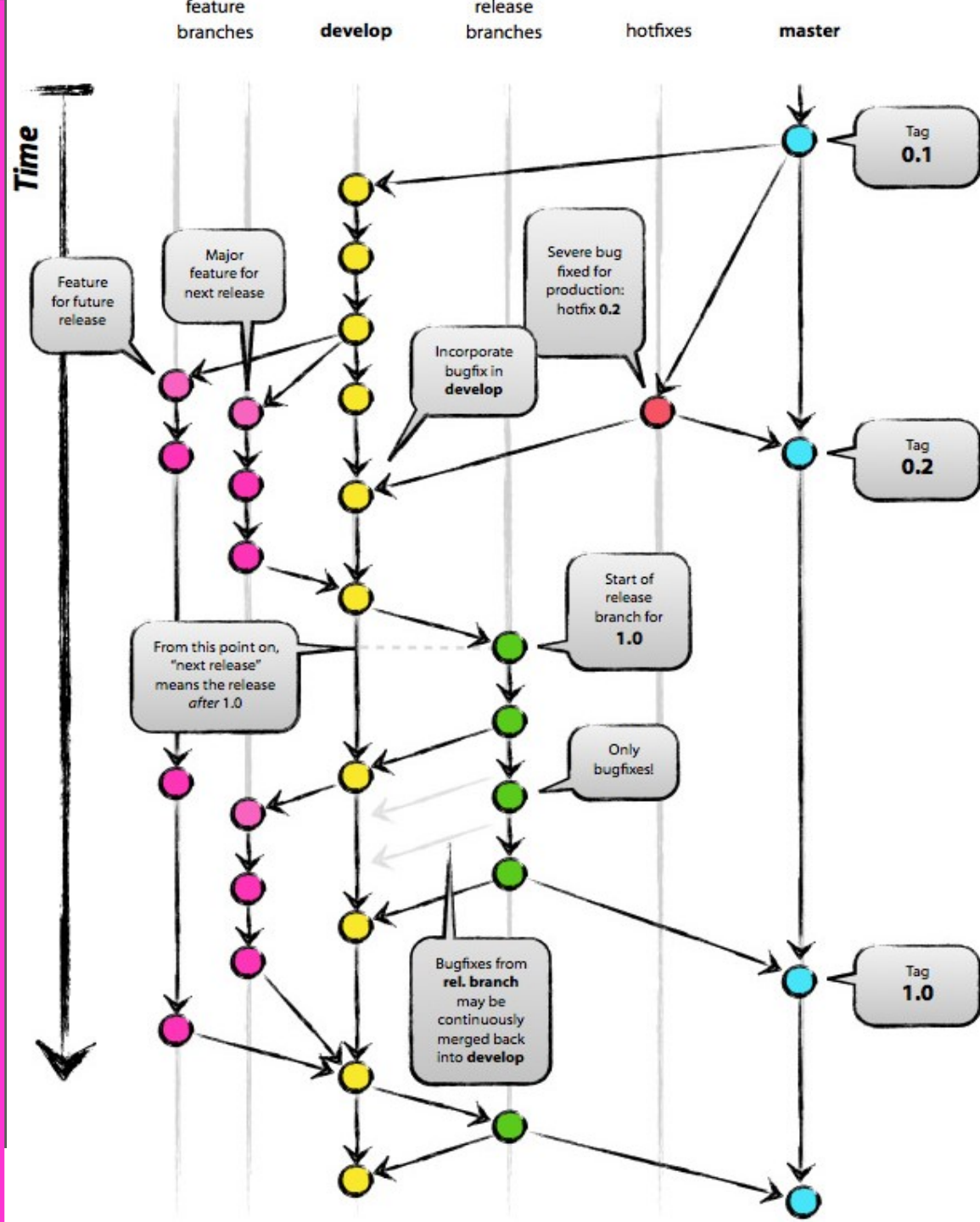
- Hashes all files in the repository
- Produces a diff of changes →
- Commit messages explain who did what when and why
- Supports parallel workflows with a method for resolving conflicts

```
584       -* AtoM looks for creator names in the *creators* column in the RAD and ISAD CSV
585       -  import templates, as well as :term:`access point` names (used as subjects) in
586       -  the *nameAccessPoints* column during a CSV import of
      622 +* AtoM looks for creator names in the *eventActors* column in the RAD and ISAD
      623 +  CSV import templates, as well as :term:`access point` names (used as subjects)
      624 +  in the *nameAccessPoints* column during a CSV import of
587   625     :term:`archival descriptions <archival description>`.
588   626   * Similarly, any Administrative / biographical history data in an archival
589       -  description CSV import (i.e. data contained in the *creatorHistories* CSV
      627 +  description CSV import (i.e. data contained in the *eventActorHistories* CSV
590   628     column will be mapped to the "History" :term:`field` (ISAAR-CPF 5.2.2) in the
591   629     related :term:`authority record` (generated from the data contained in the
592   630     *creators* column of the CSV), and then is presented in AtoM in any related
593   631     descriptions where the entity is listed as a creator.
594   632   * Where multiple creator names and histories are included in an import,
595       -  *creators* and *creatorHistories* elements are matched 1:1 in the  order they
596       -  appear in the CSV, divided by pipe elements (e.g. ``|`` ). For example, if the
597       -  *creators* column contains ``name 1|name 2``, the *creatorHistories* should
598       -  also include ``history 1|history 2`` to match on import.
599       -* If a creator history element is included in a CSV import, but no creator
      633 +  *eventActors* and *eventActorHistories* elements are matched 1:1 in the
      634 +  order they appear in the CSV, divided by pipe elements (e.g. ``|`` ). For
      635 +  example, if the *eventActors* column contains ``name 1|name 2``, the
      636 +  *eventActorHistories* should  also include ``history 1|history 2`` to match on
      637 +  import. If there is **no** history for the first actor, you can include
      638 +  ``NULL``, and AtoM will ignore the imput - e.g. ``name 1|name 2`` should be
      639 +  matched with ``NULL|history 2`` to include only a history for name 2.
      640 +* This same ``NULL`` approach can be used for any matched date values where
      641 +  multiple actor names are included for import - ``eventDates``,
      642 +  ``eventStartDates``, ``eventEndDates`` can all include ``NULL`` if you wish
      643 +  to leave these blank when associating multiple actors with an event. An
      644 +  example, using the RAD template:
      645 +
      646 +.. image:: images/csv-creatorDates-2.*
      647 +   :align: center
```

https://github.com/artefactual/atom-docs/commit/6d8fe4ab9803ec3421b04bcabe215e3dbb4a77dc

Git

- Essentially a (NoSQL) database – a key/value store
  - Objects DB and Refs DB
- Libgit2: git implementation in C
  - Supports different backends
- Might git provide a model for auditable, versioned data management?



Vincent Driessen via http://nvie.com/posts/a-successful-git-branching-model/

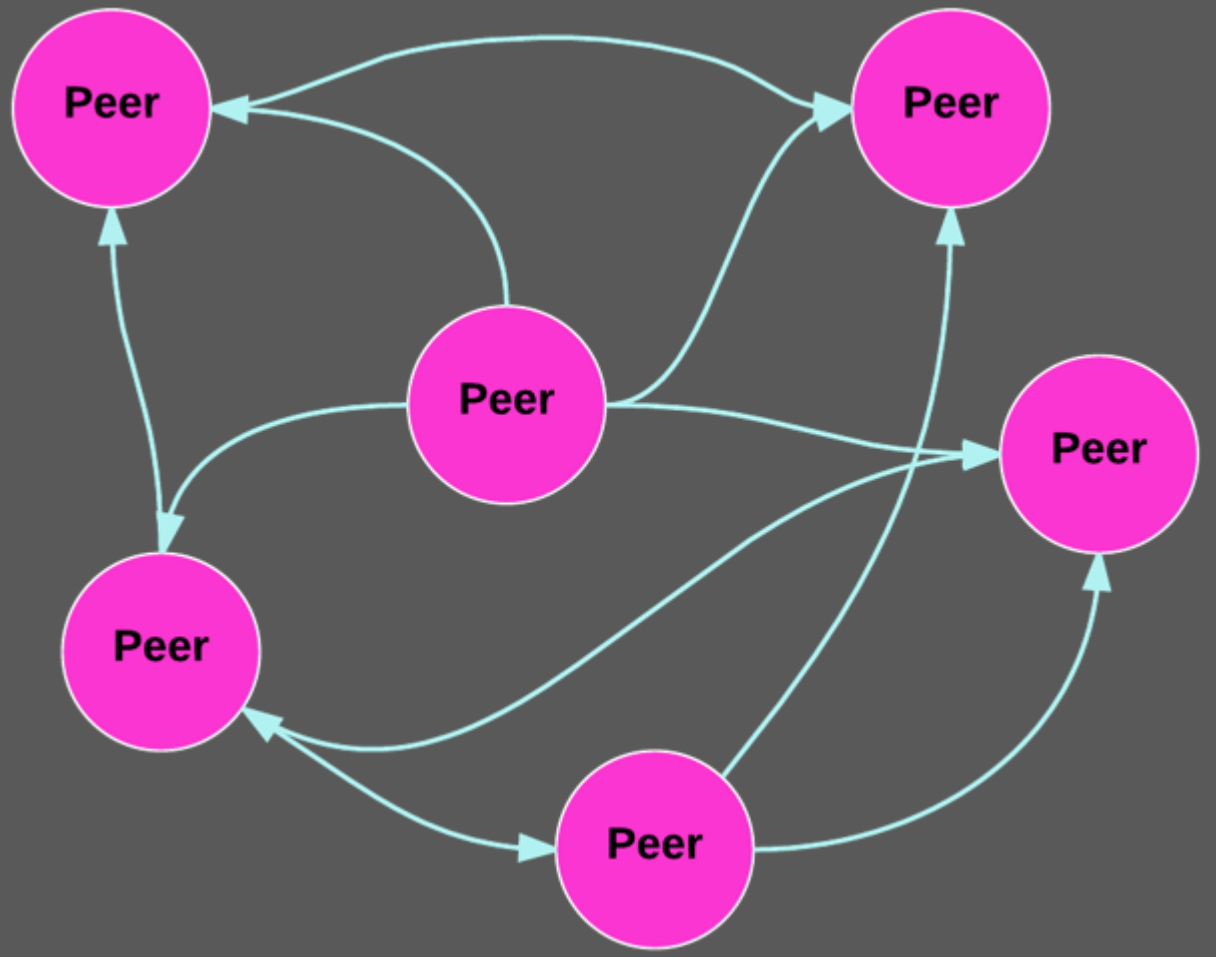
Git for Archives?

- Description systems:
  - Maintain transparent provenance and manage collaboration and concurrency – e.g. authority records
  - Collaborative preservation environments
  - Track transfer/ingest, preservation actions
- Distributed, version-controlled backups
  - Reconcile changes via branch management

Peer to Peer computing and BitTorrent

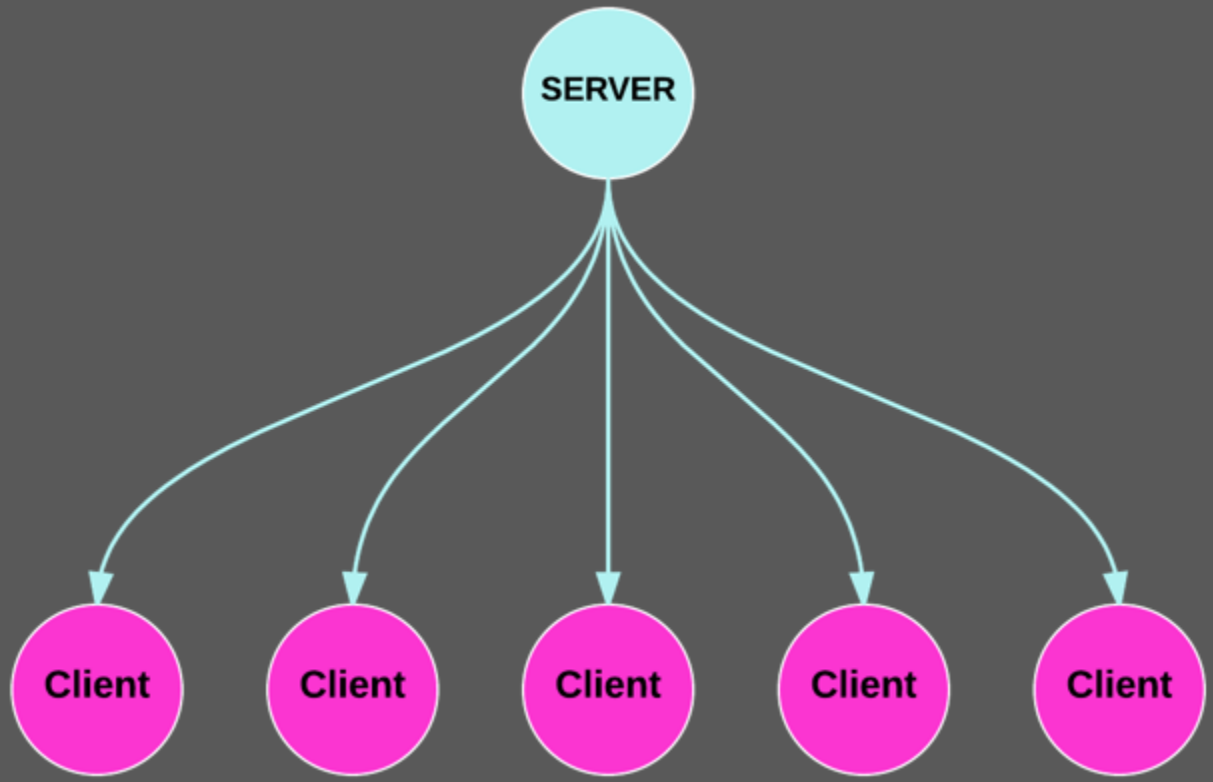
P2P:

A **distributed** application architecture that directly links users as peers with equal privileges, instead of relying on a centralized server for communication and exchange

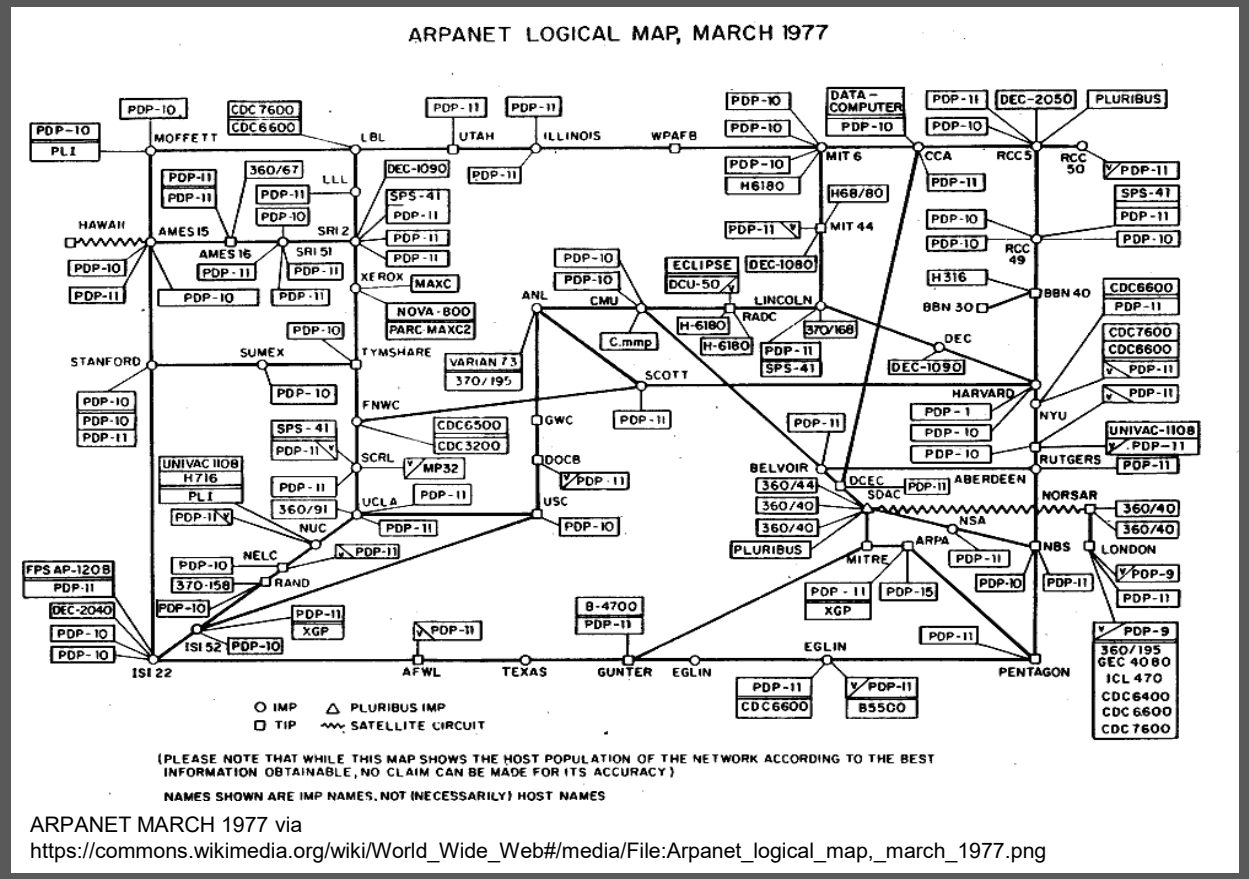


Early mainframe computing

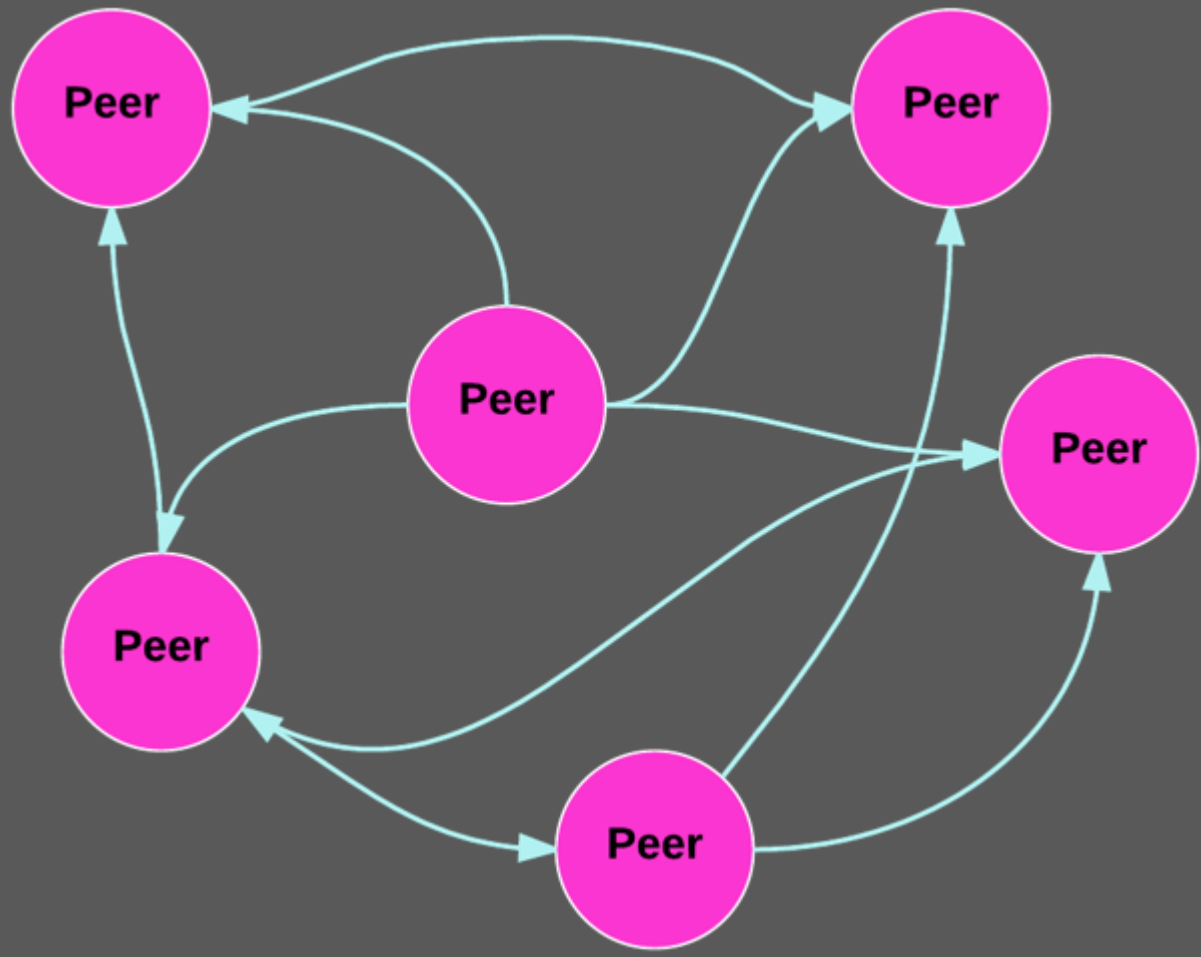
Lawrence Livermore National Laboratory. IBM 704 Mainframe, via https://en.wikipedia.org/wiki/Mainframe_computer#/media/File:IBM_704_mainframe.gif



The promise of the web

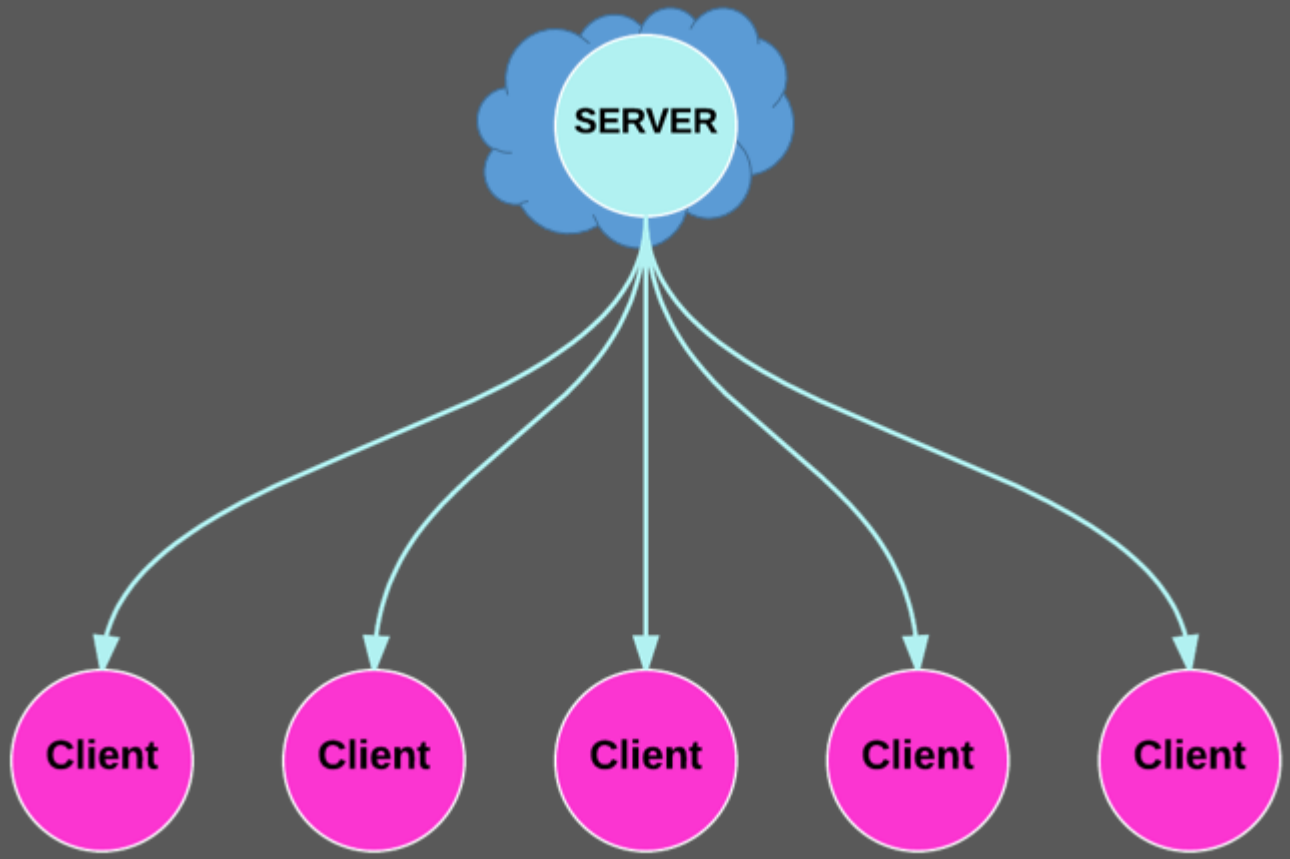


ARPANET MARCH 1977 via
https://commons.wikimedia.org/wiki/World_Wide_Web#/media/File:Arpanet_logical_map,_march_1977.png

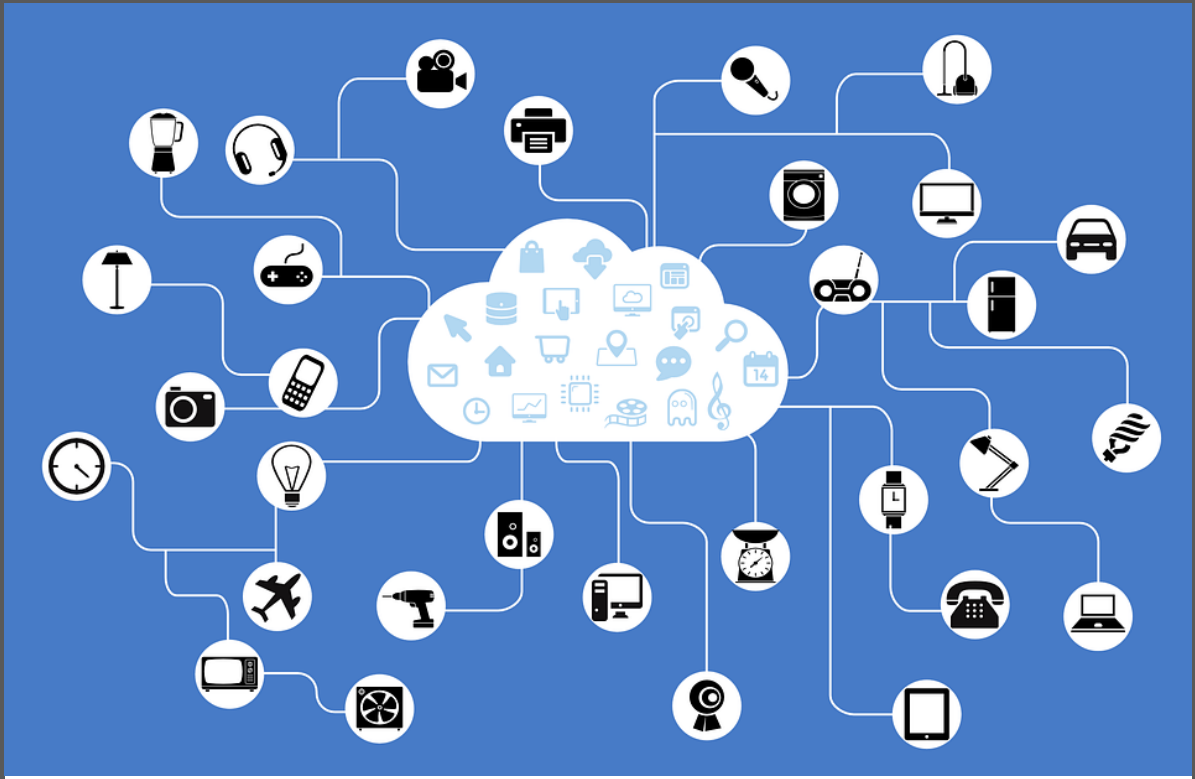


The rise of the cloud

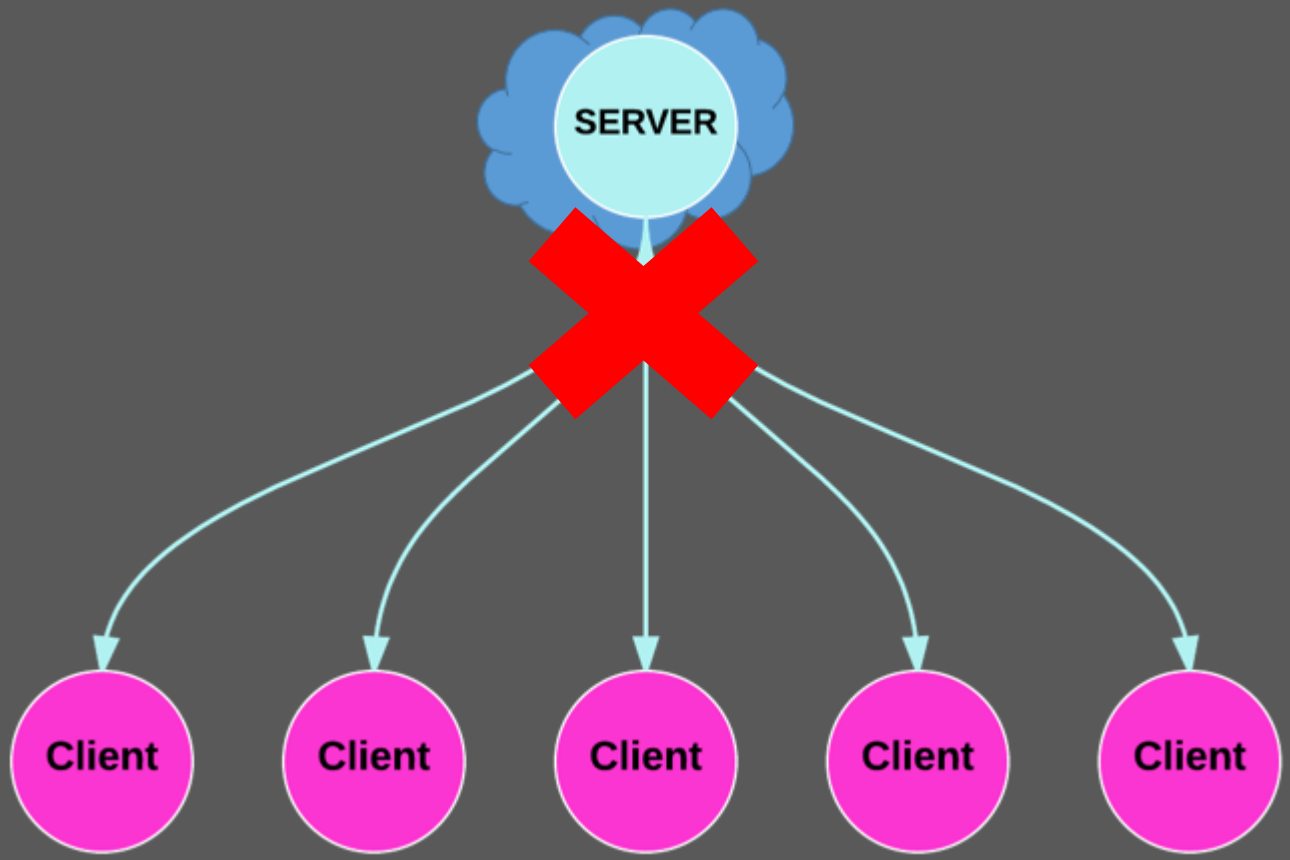
https://pixabay.com/en/network-iot-internet-of-things-782707/



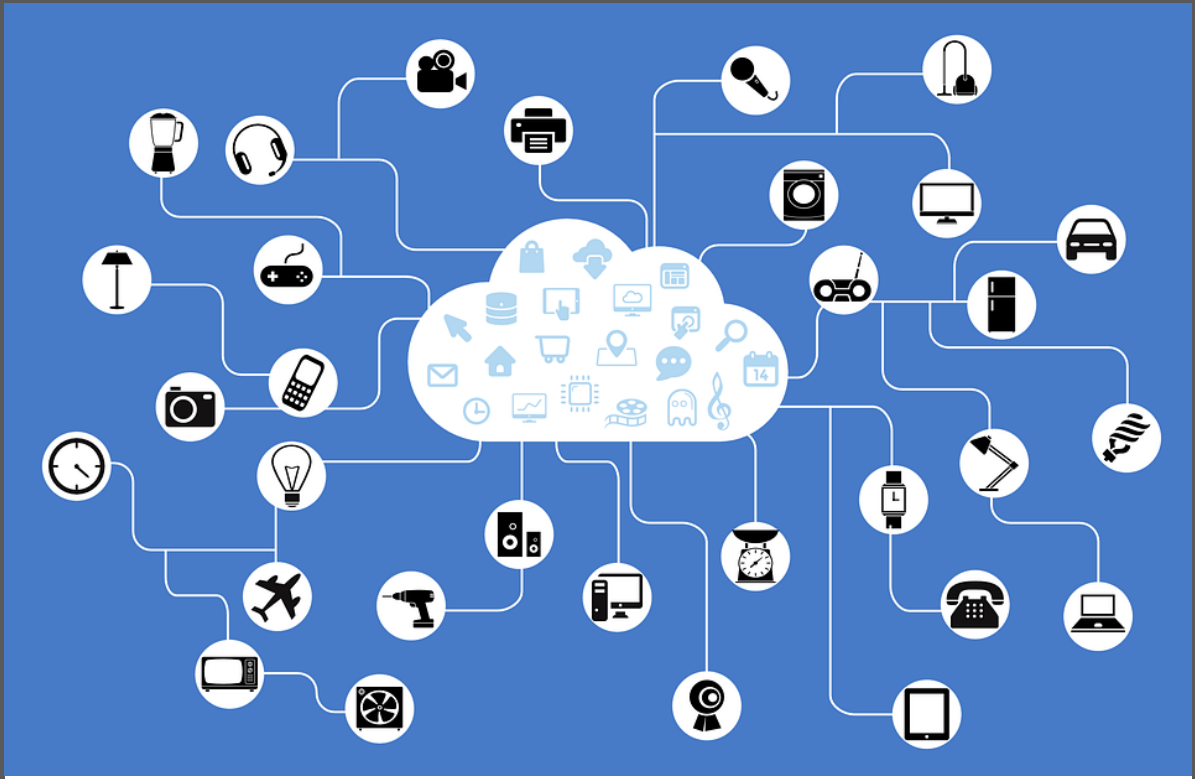
The rise of the cloud



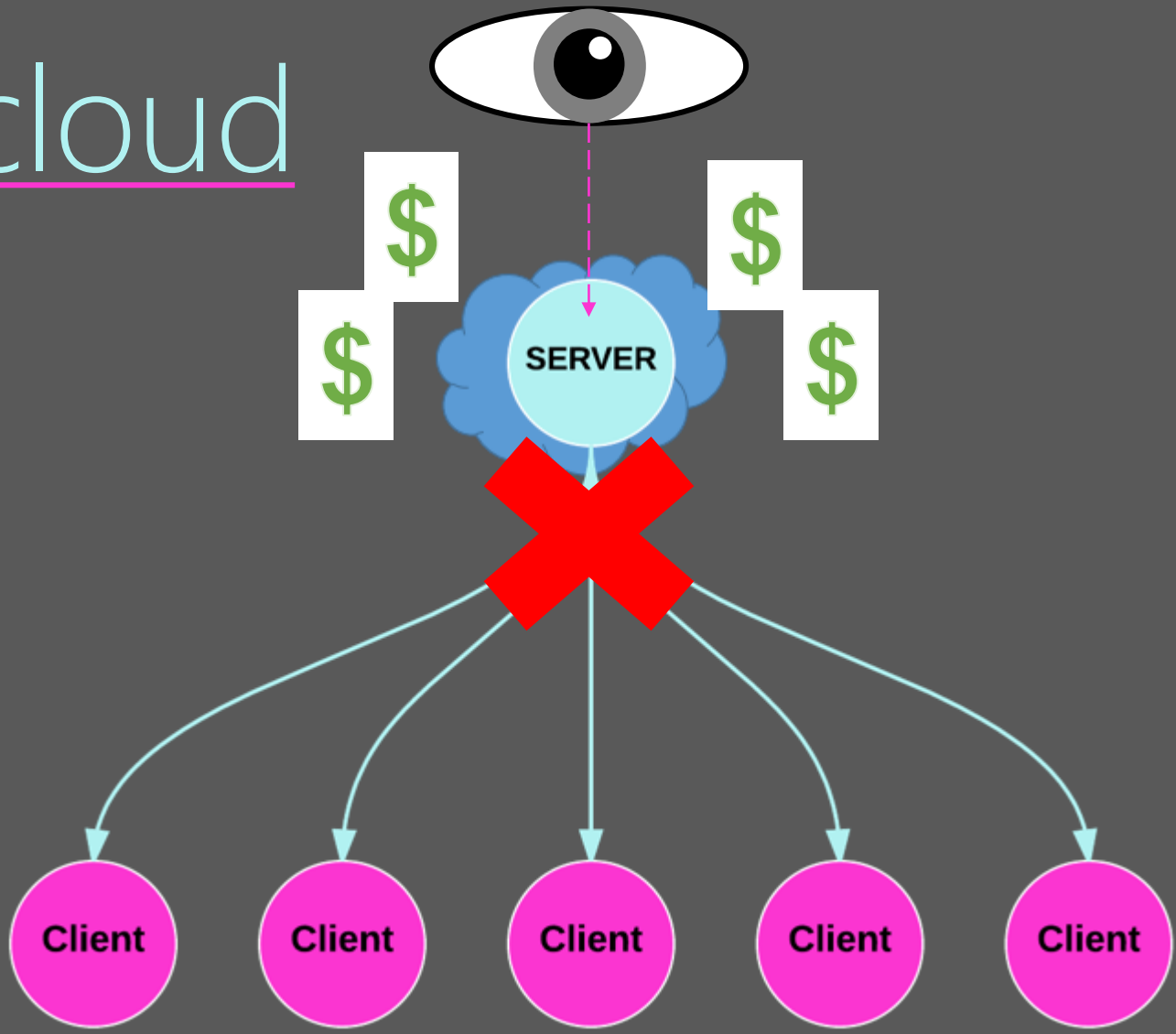
https://pixabay.com/en/network-iot-internet-of-things-782707/



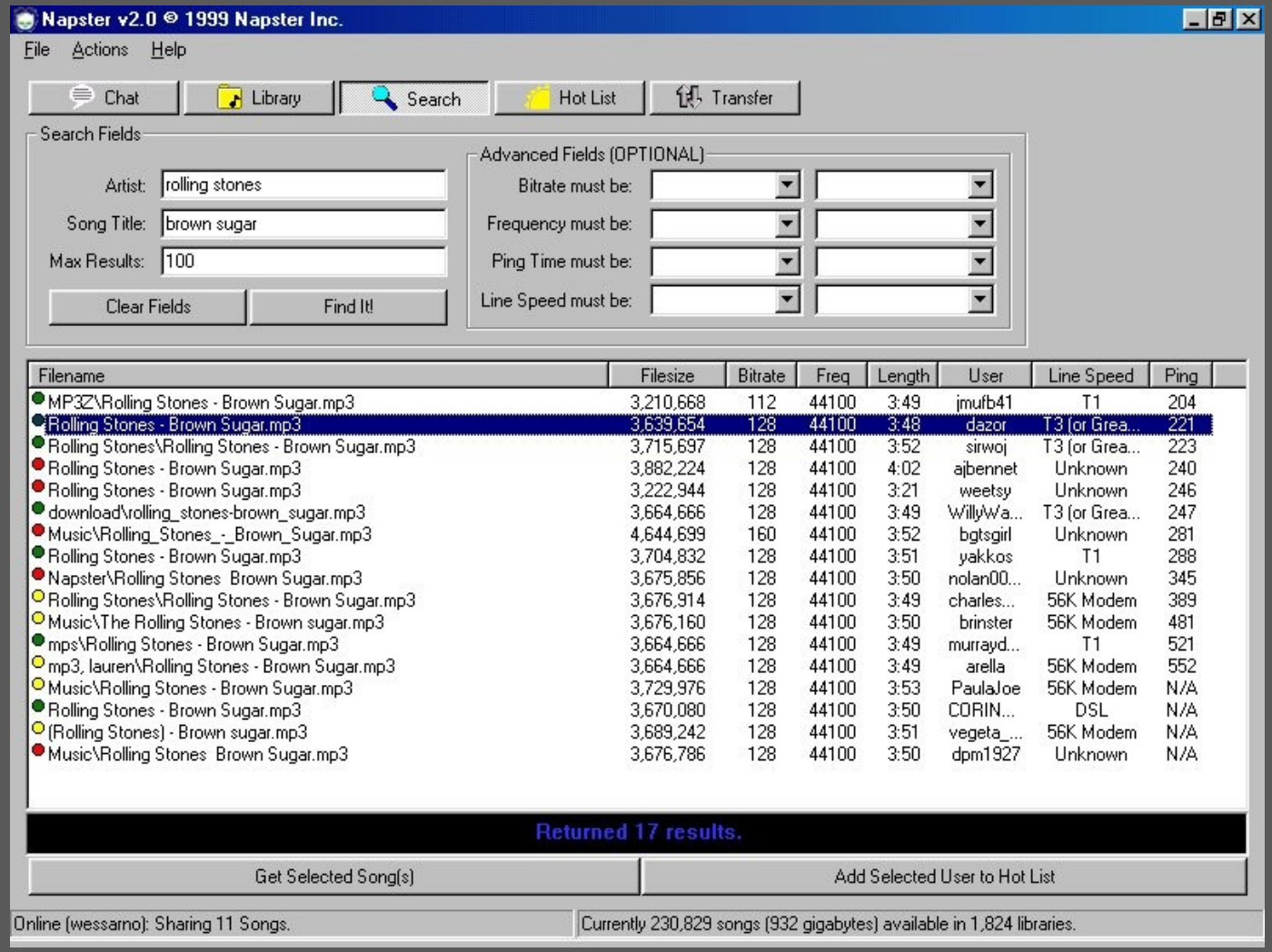
The rise of the cloud



https://pixabay.com/en/network-iot-internet-of-things-782707/



P2P file sharing



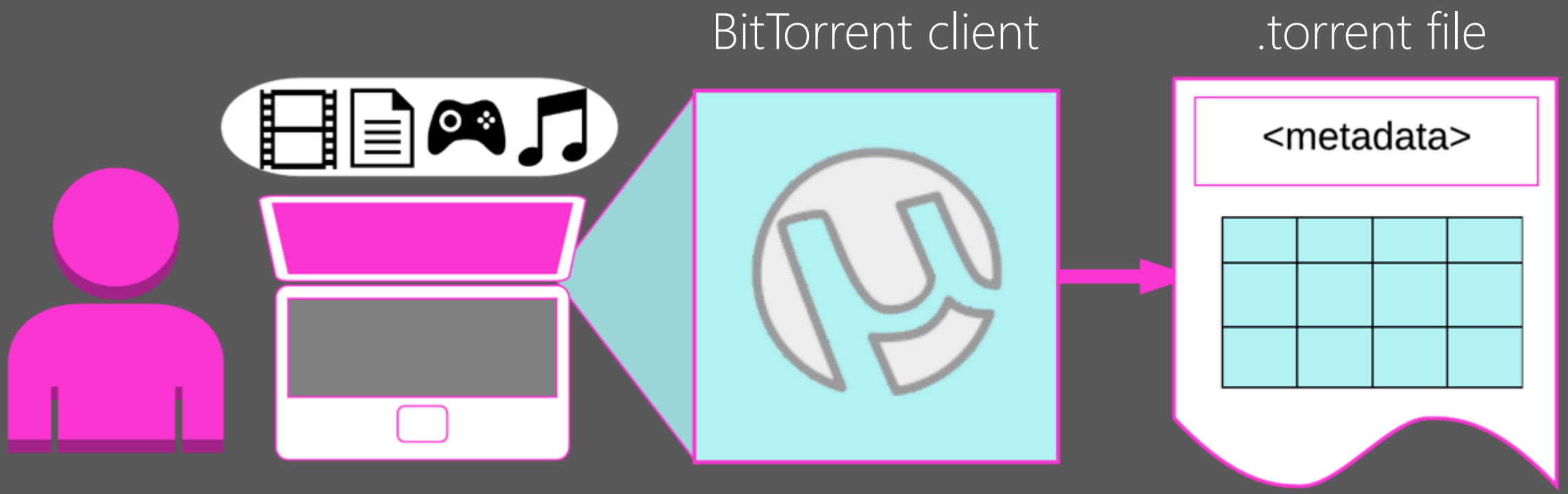
http://www.nextinpact.com/dossier/hadopi-nicolas-sarkozy-sacd-snep/1.htm

http://popjoust.com/jousts/400

BitTorrent Protocol

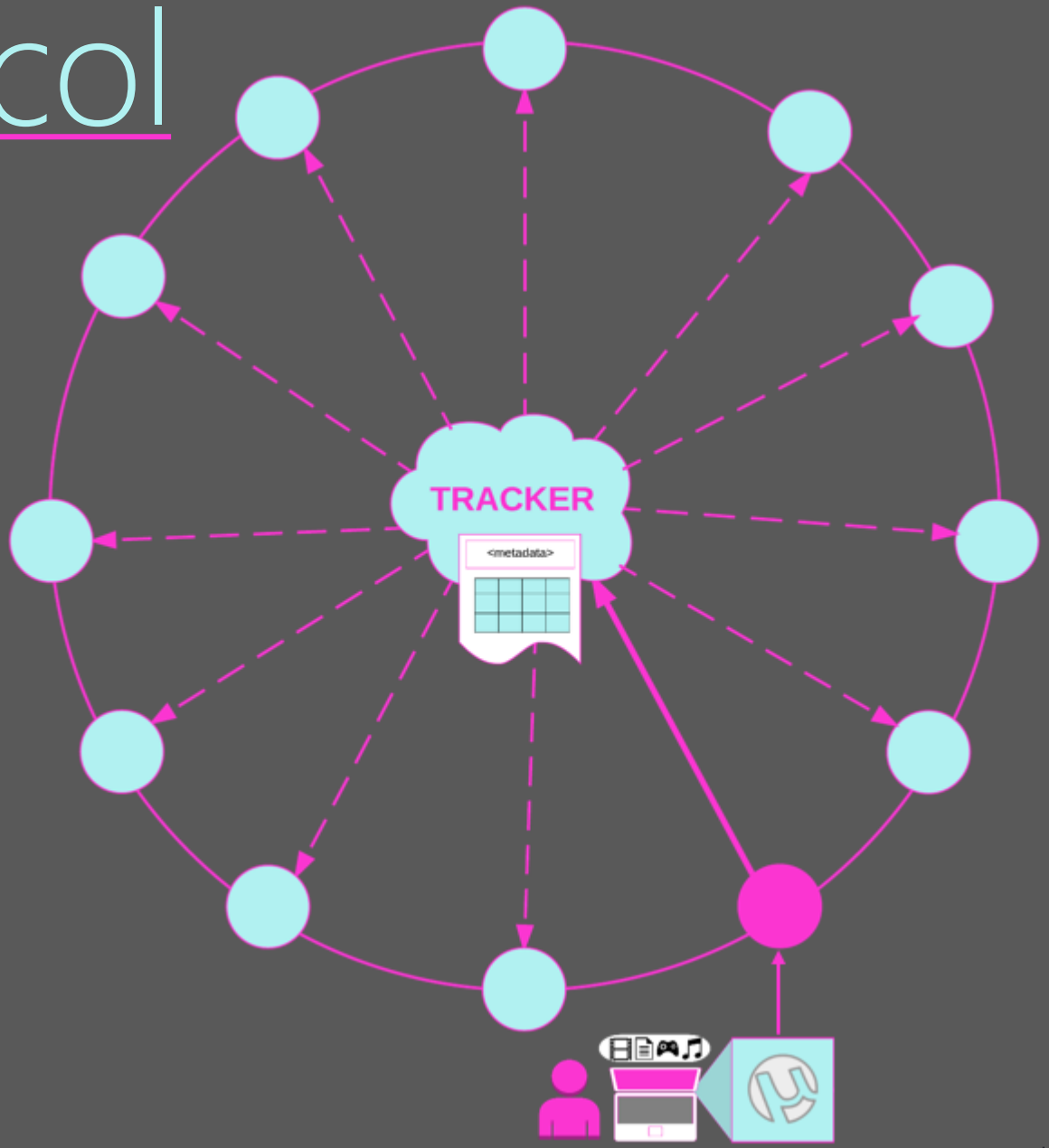
- A communications protocol for peer-to-peer file sharing
- Works with any kind/size of files
- Increases speed with increases in users – each user becomes a potential download source
- Uses cryptographic hashing to segment files into pieces, so different pieces can be downloaded from different users and verified against the torrent tracker

BitTorrent Protocol



BitTorrent Protocol

- Torrent users who share content are called peers
- A group of peers is called a swarm
- Users who download without also sharing are called leeches
- Early BT implementations used a centralized tracker to discover and connect peers, before they could start sharing directly

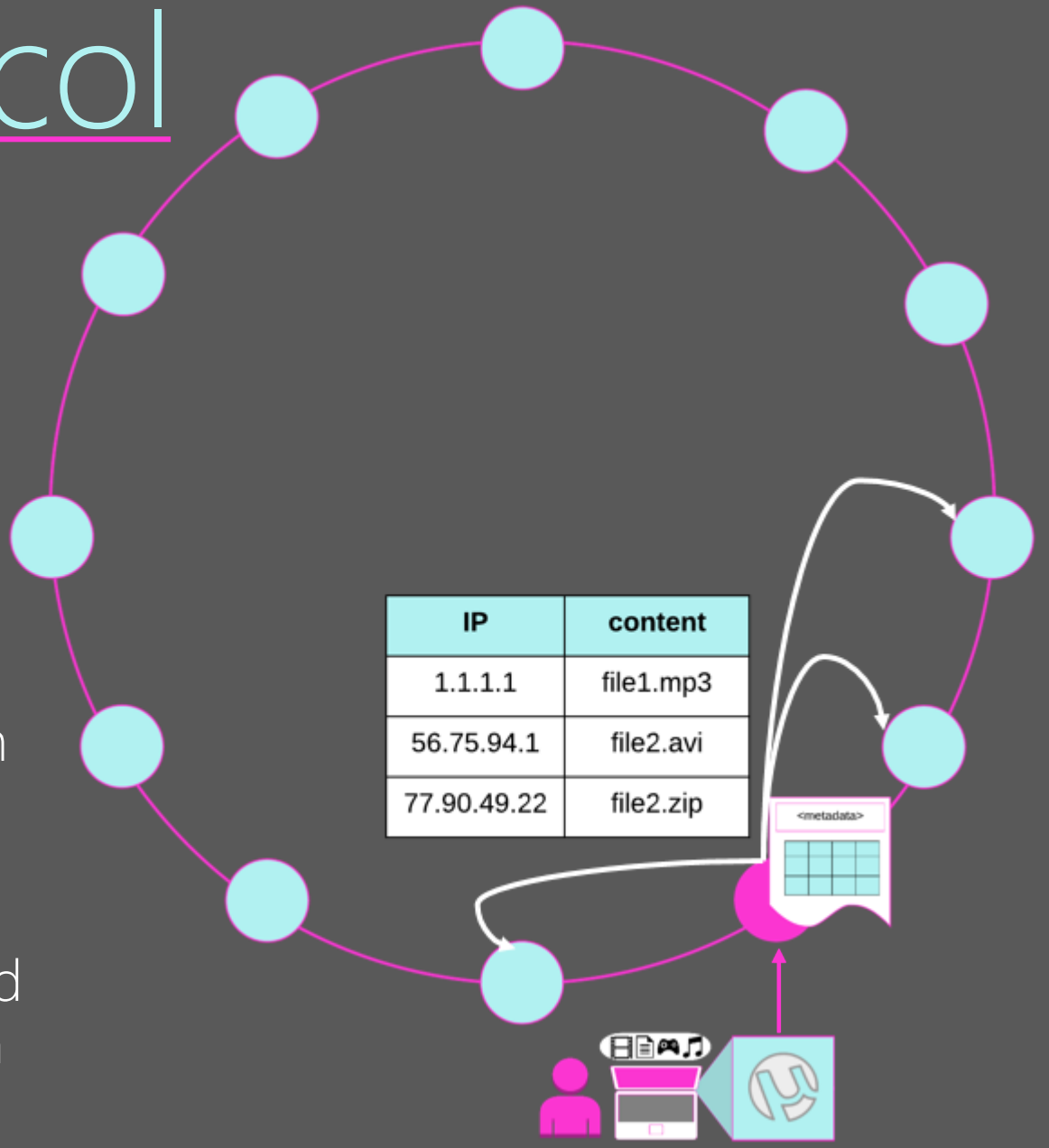


BitTorrent Protocol

DHT:

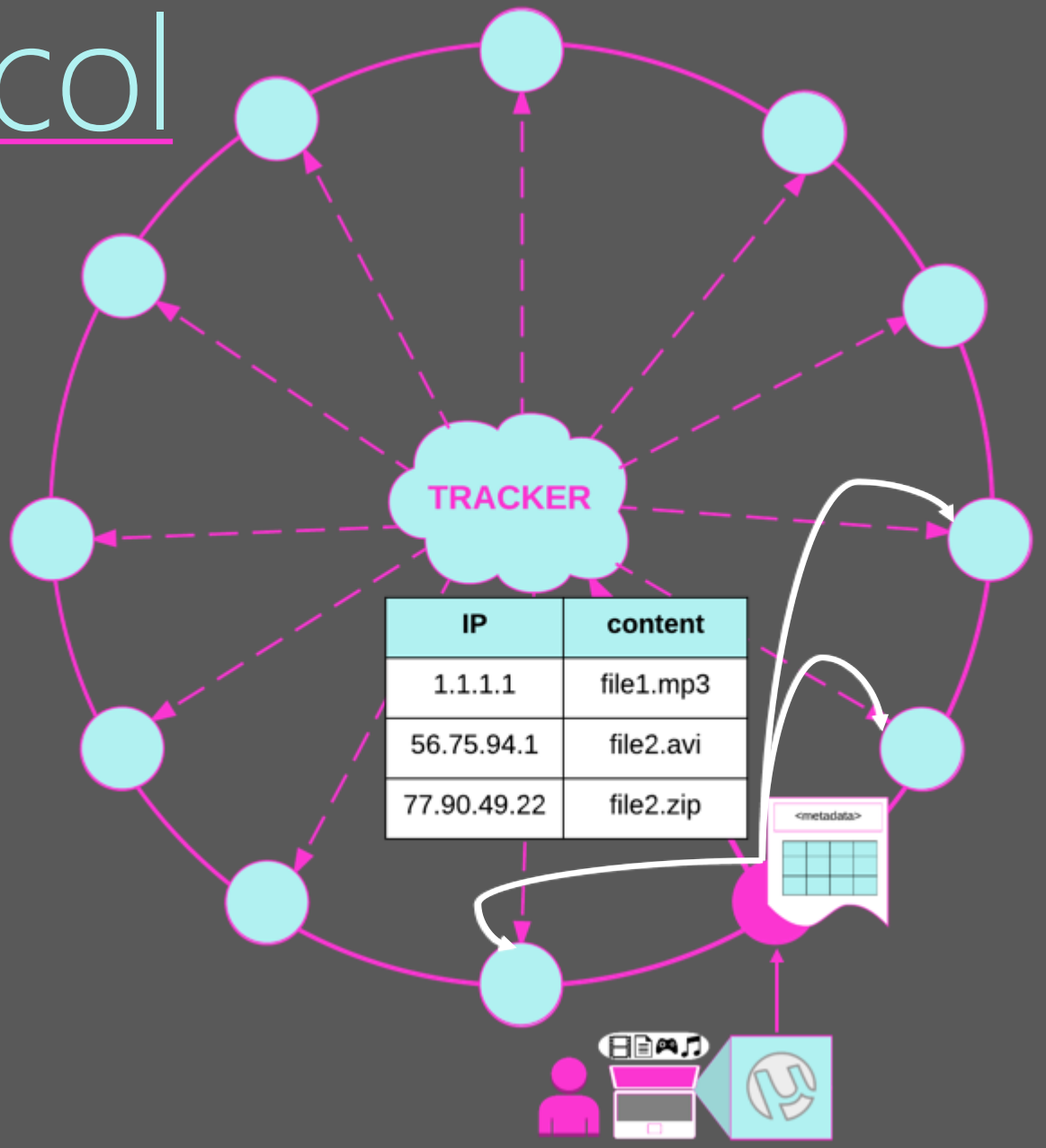
Distributed Hash Table. A decentralized lookup system (similar to a hash table) with key/value pairs.

- Pairs are stored in the DHT
- Any peer can look up the value associated with a key
- Responsibility for maintaining the mapping from keys to values is algorithmically distributed among the nodes to minimize disruption when nodes appear/depart/fail

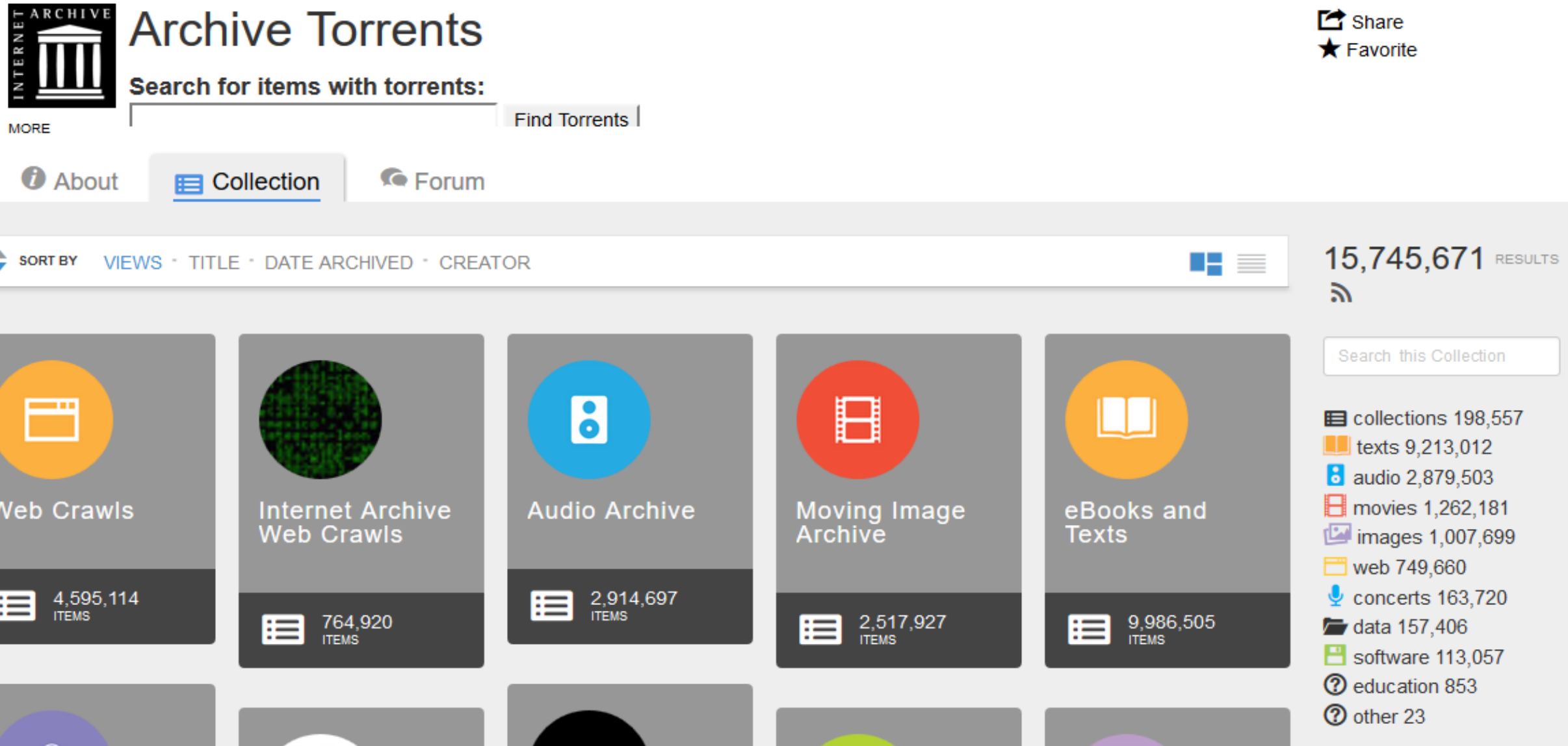


BitTorrent Protocol

- Central trackers and DHT can be used together with other methods to increase efficiency
- Other features (e.g. encryption) can be added on top

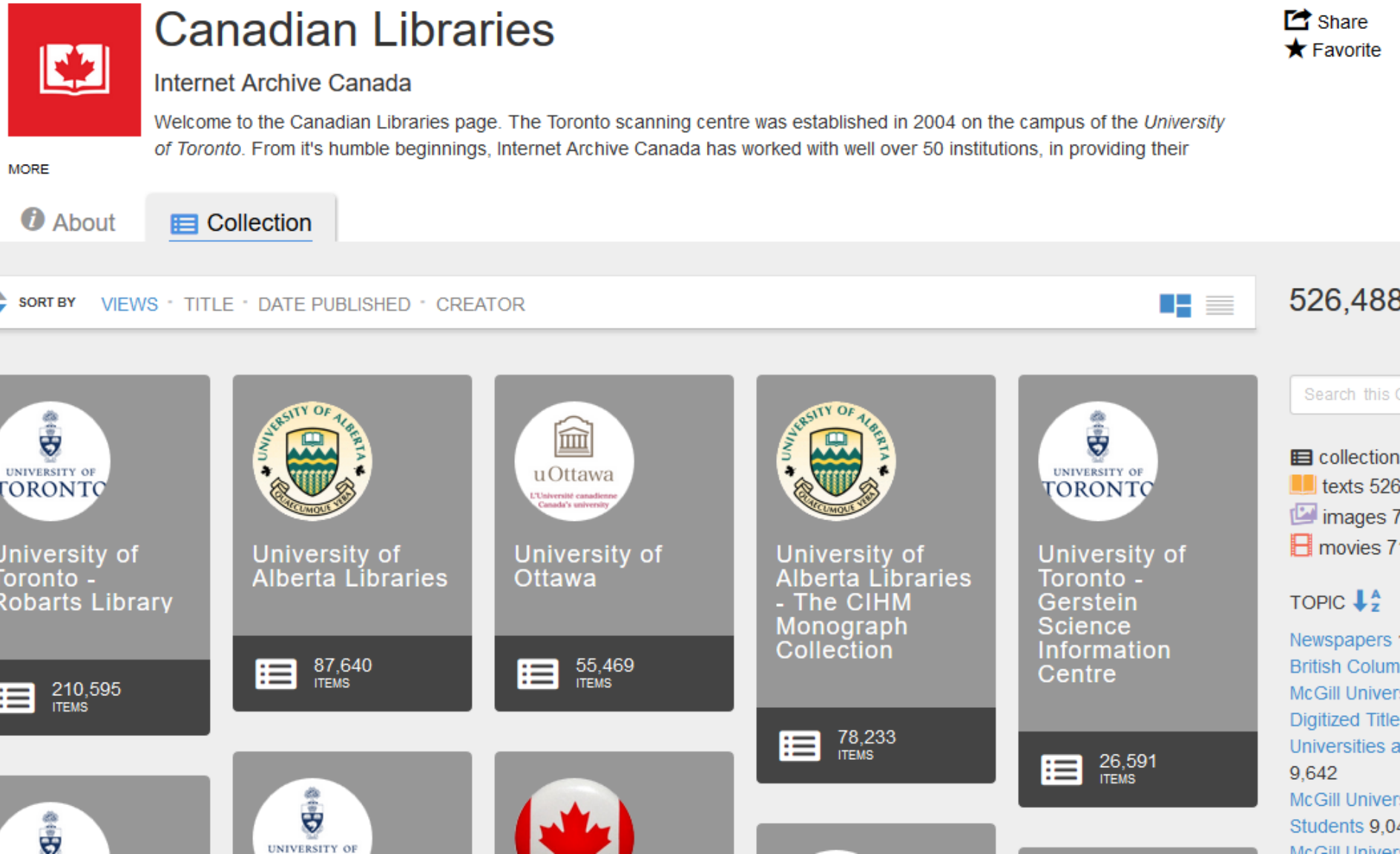


Archival uses for BitTorrent?



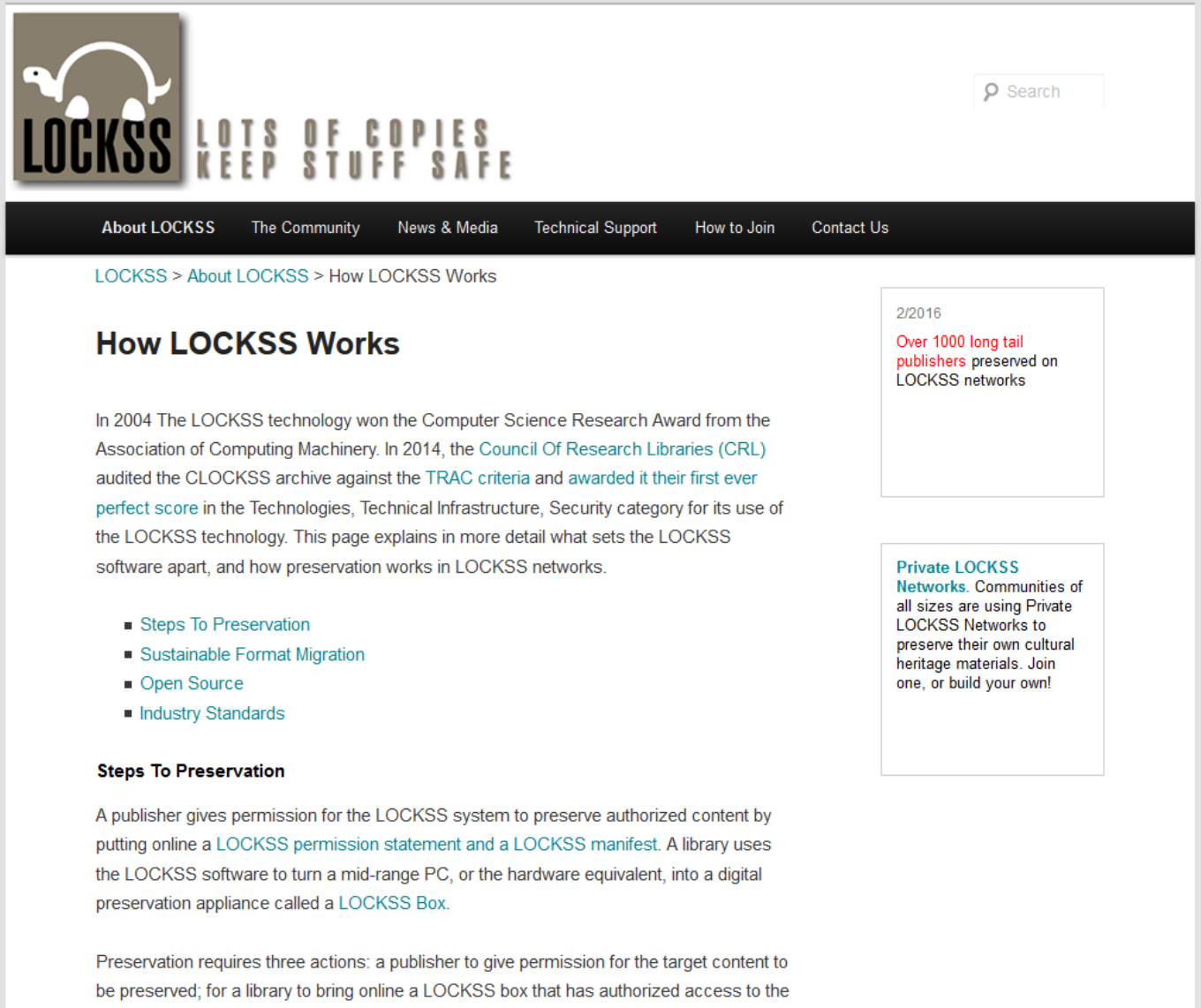
https://archive.org/details/bittorrent

Archival uses for BitTorrent?



https://archive.org/details/toronto

Archival uses for P2P and data?



LOCKSS LOTS OF COPIES KEEP STUFF SAFE

Search

About LOCKSS | The Community | News & Media | Technical Support | How to Join | Contact Us

LOCKSS > About LOCKSS > How LOCKSS Works

How LOCKSS Works

In 2004 The LOCKSS technology won the Computer Science Research Award from the Association of Computing Machinery. In 2014, the Council Of Research Libraries (CRL) audited the CLOCKSS archive against the TRAC criteria and awarded it their first ever perfect score in the Technologies, Technical Infrastructure, Security category for its use of the LOCKSS technology. This page explains in more detail what sets the LOCKSS software apart, and how preservation works in LOCKSS networks.

- Steps To Preservation
- Sustainable Format Migration
- Open Source
- Industry Standards

Steps To Preservation

A publisher gives permission for the LOCKSS system to preserve authorized content by putting online a LOCKSS permission statement and a LOCKSS manifest. A library uses the LOCKSS software to turn a mid-range PC, or the hardware equivalent, into a digital preservation appliance called a LOCKSS Box.

Preservation requires three actions: a publisher to give permission for the target content to be preserved; for a library to bring online a LOCKSS box that has authorized access to the

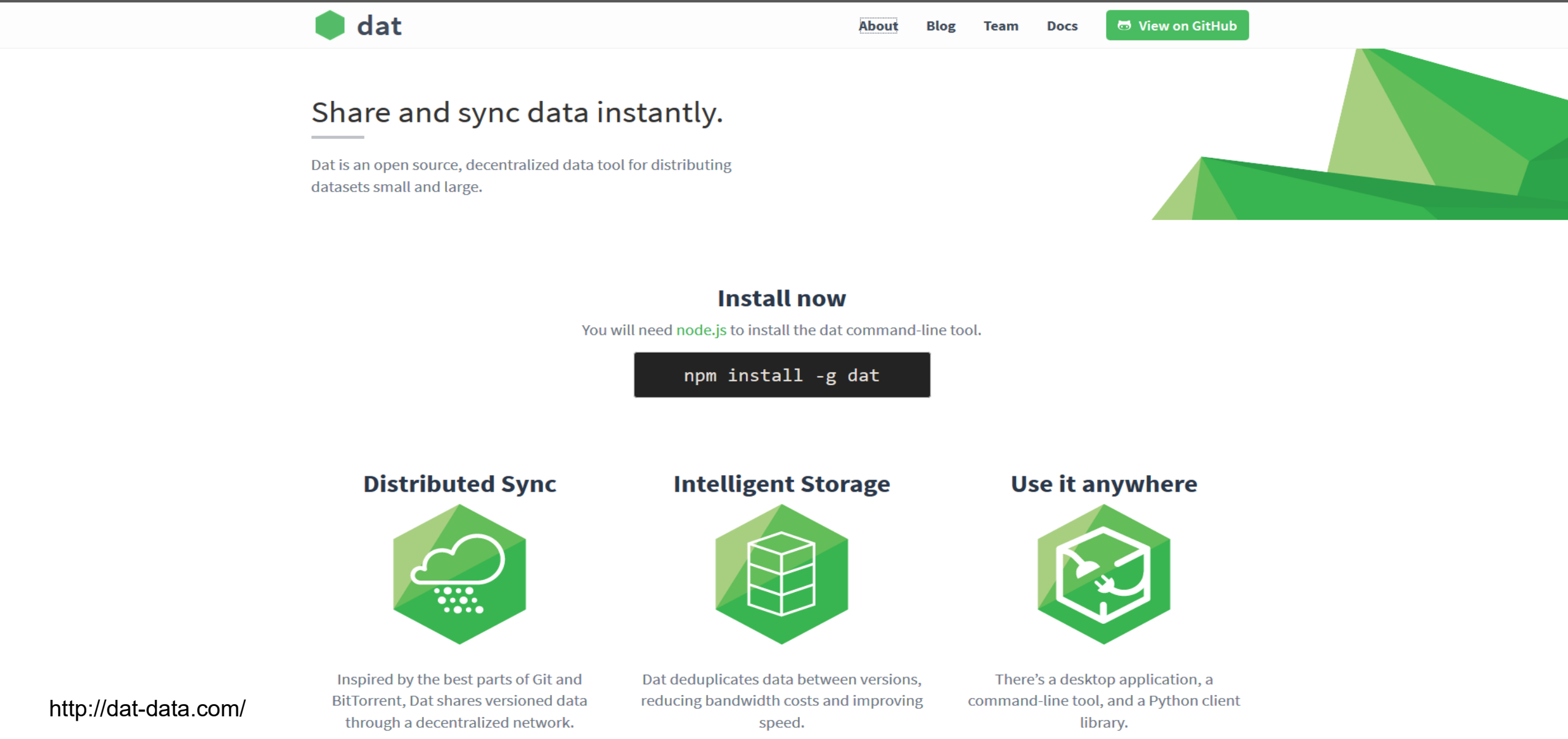
2/2016

Over 1000 long tail publishers preserved on LOCKSS networks

Private LOCKSS Networks. Communities of all sizes are using Private LOCKSS Networks to preserve their own cultural heritage materials. Join one, or build your own!

https://www.lockss.org/about/how-it-works/

Archival uses for P2P and data?



http://dat-data.com/

Archival uses for P2P and data?

README.md

dat jawn: 'Git for Tabular Data'

build passing

NPM

Jawn is a node.js module that allows *distributed version control of Tabular Data*. It's connected to the dat project. It allows you to import tabular data (rows and columns like CSV or TSV) and track how those data change over time. *Do you have non-tabular data? read this:* What about Non Tabular Data?

The key features for jawn are to:

- **manage and track change history** in tabular data
- **create historical checkpoints** with metadata (e.g., message, timestamp, author)

Jawn relies on hypercore to handle the core functions around creating merkle chains, which allows us to

- **supply access points to data** across the network with a peer-to-peer model
- **sync incrementally** between machines

This is where jawn connects with the current work of the dat team, who created hypercore and are using it to do the same things with directories of files. For more background info, read our Technical Background and Reference Code Bases wiki page.

Project Team

jawn is maintained by a Code for Philly project that aims to be a model for mentorship and collaborative learning. For full information about the project go to the jawn project page

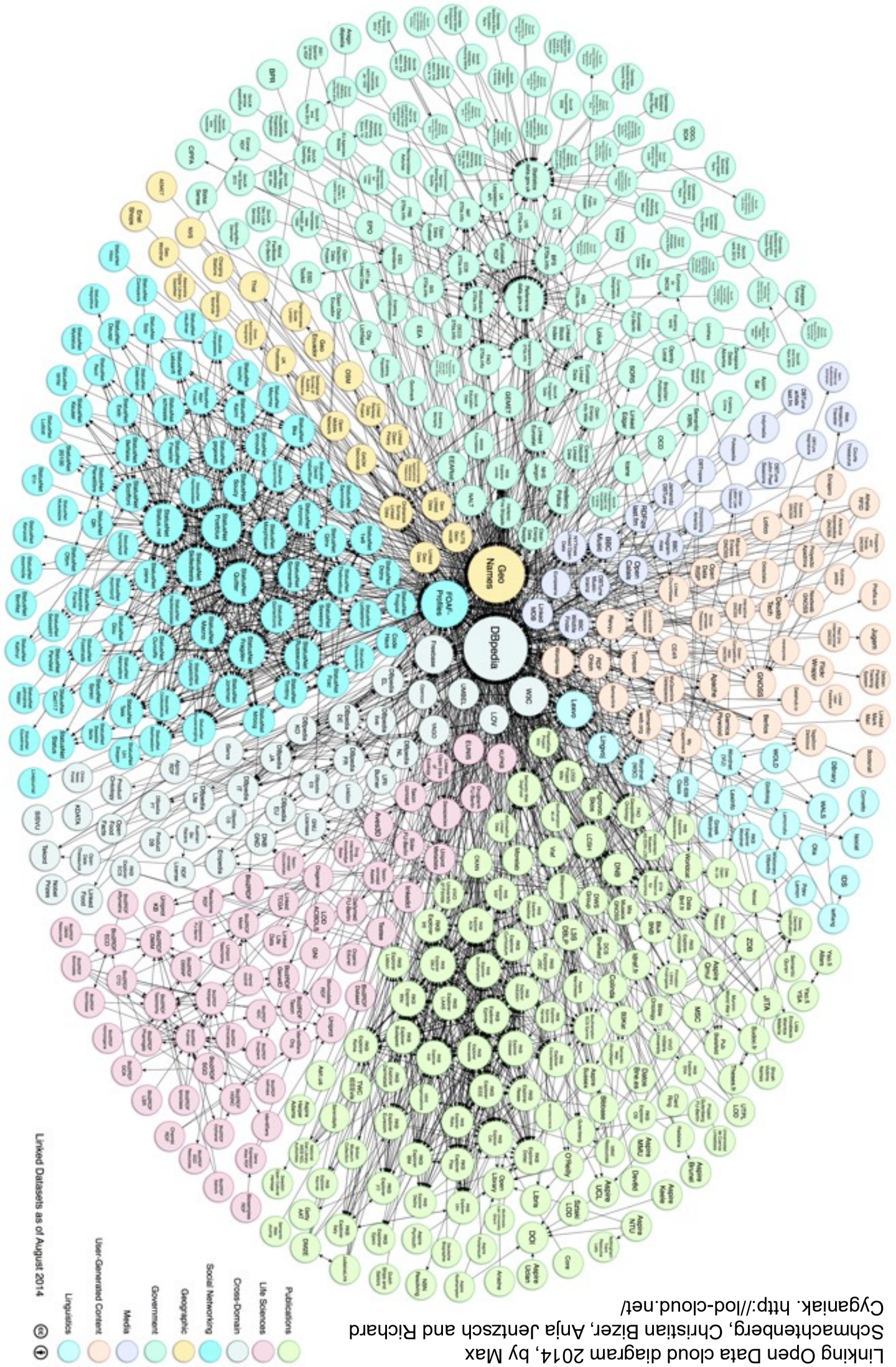
We welcome contributions from anyone.

https://github.com/CfABrigadePhiladelphia/jawn

Archival uses for P2P and data?

- Description and access: a networking and data exchange protocol
  - Exchanging authority records between repositories
  - Updating portal sites
  - Offering content to the public
- Preservation: Sharing supporting technology images
  - Operating systems, codecs, tools
  - LOCKSS-style approach to preservation and real-time access despite service interruptions at any one point – distributed copies

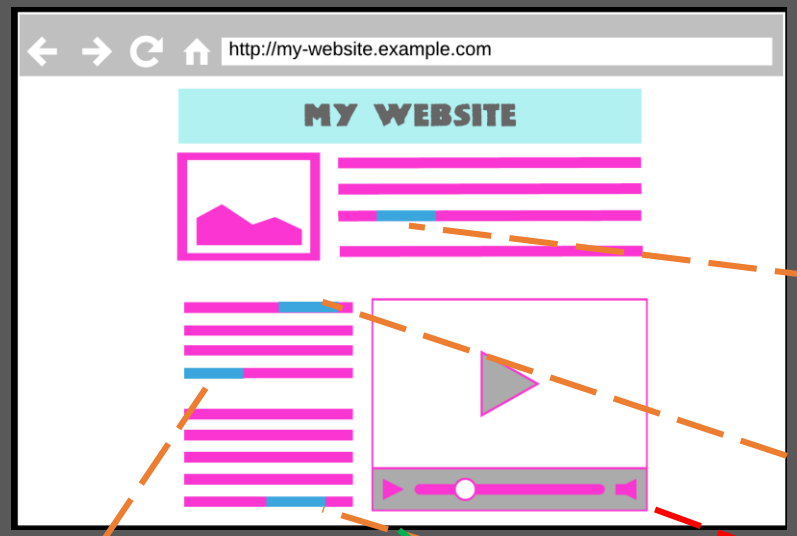
Linked open data (LOD)

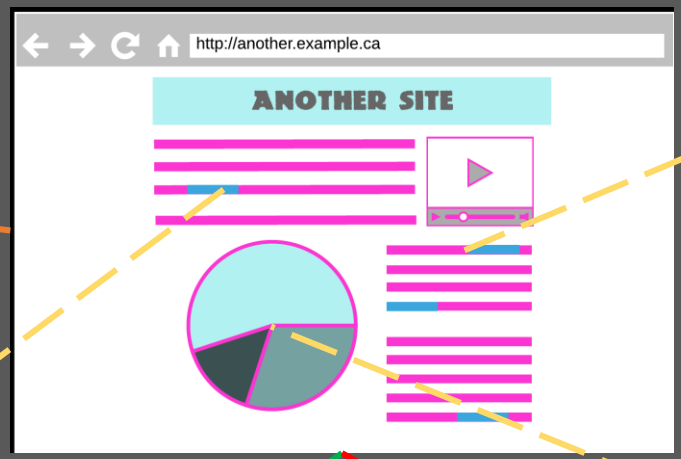


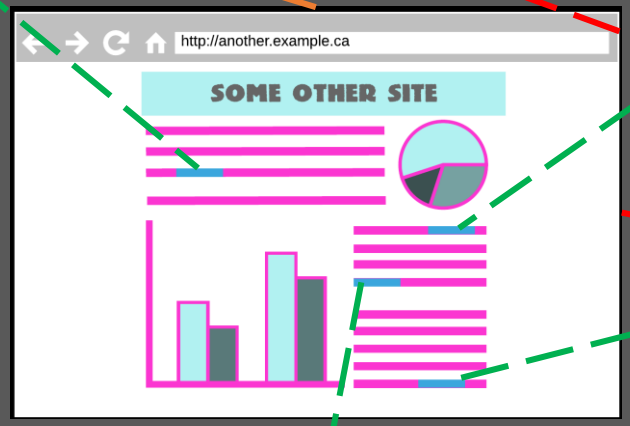
Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

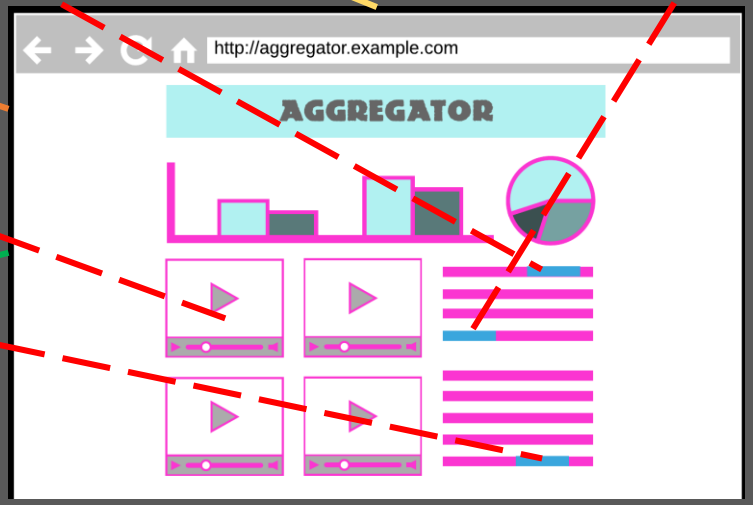
The WWW

The WWW

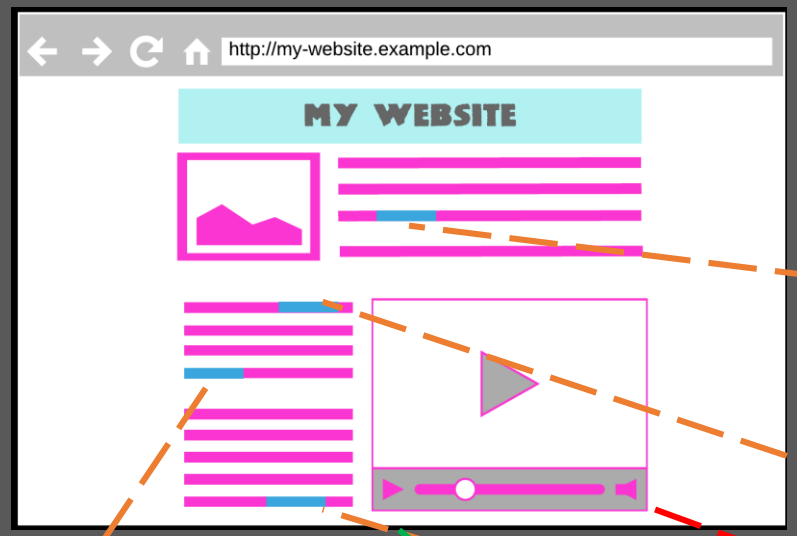

http://my-website.example.com
MY WEBSITE

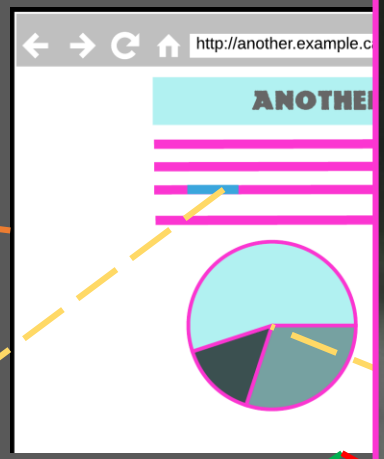

http://another.example.ca
ANOTHER SITE

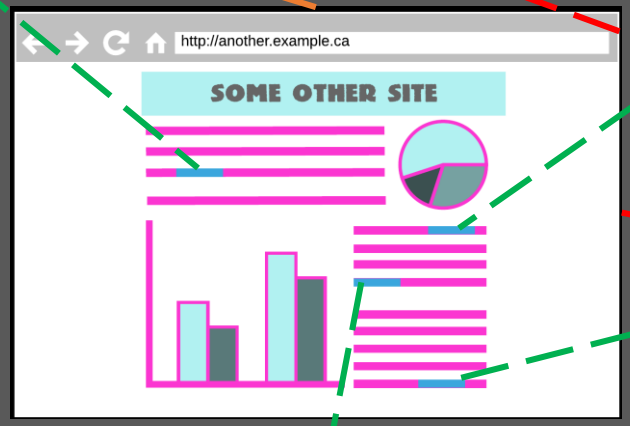

http://another.example.ca
SOME OTHER SITE

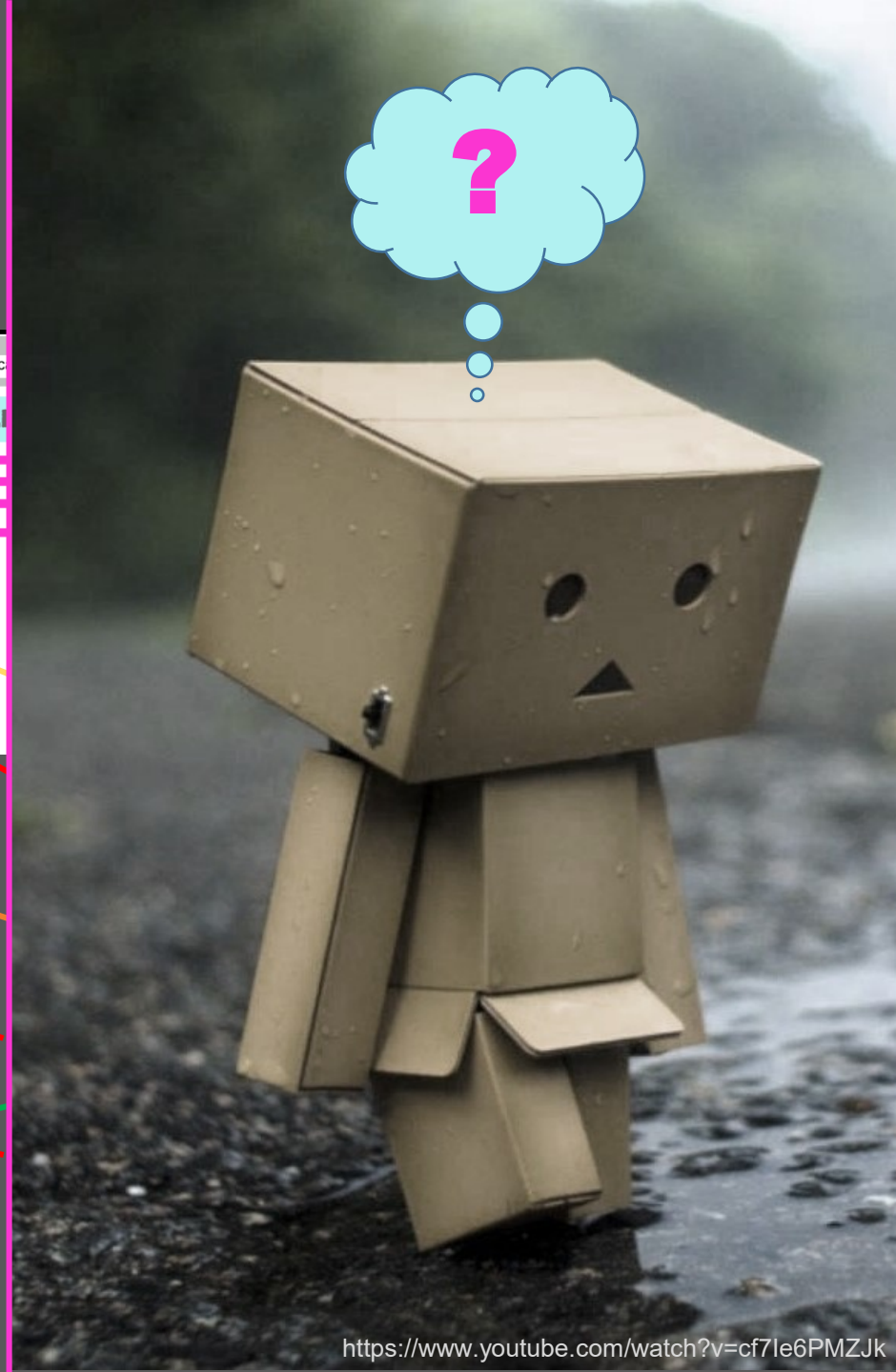

http://aggregator.example.com
AGGREGATOR

The WWW







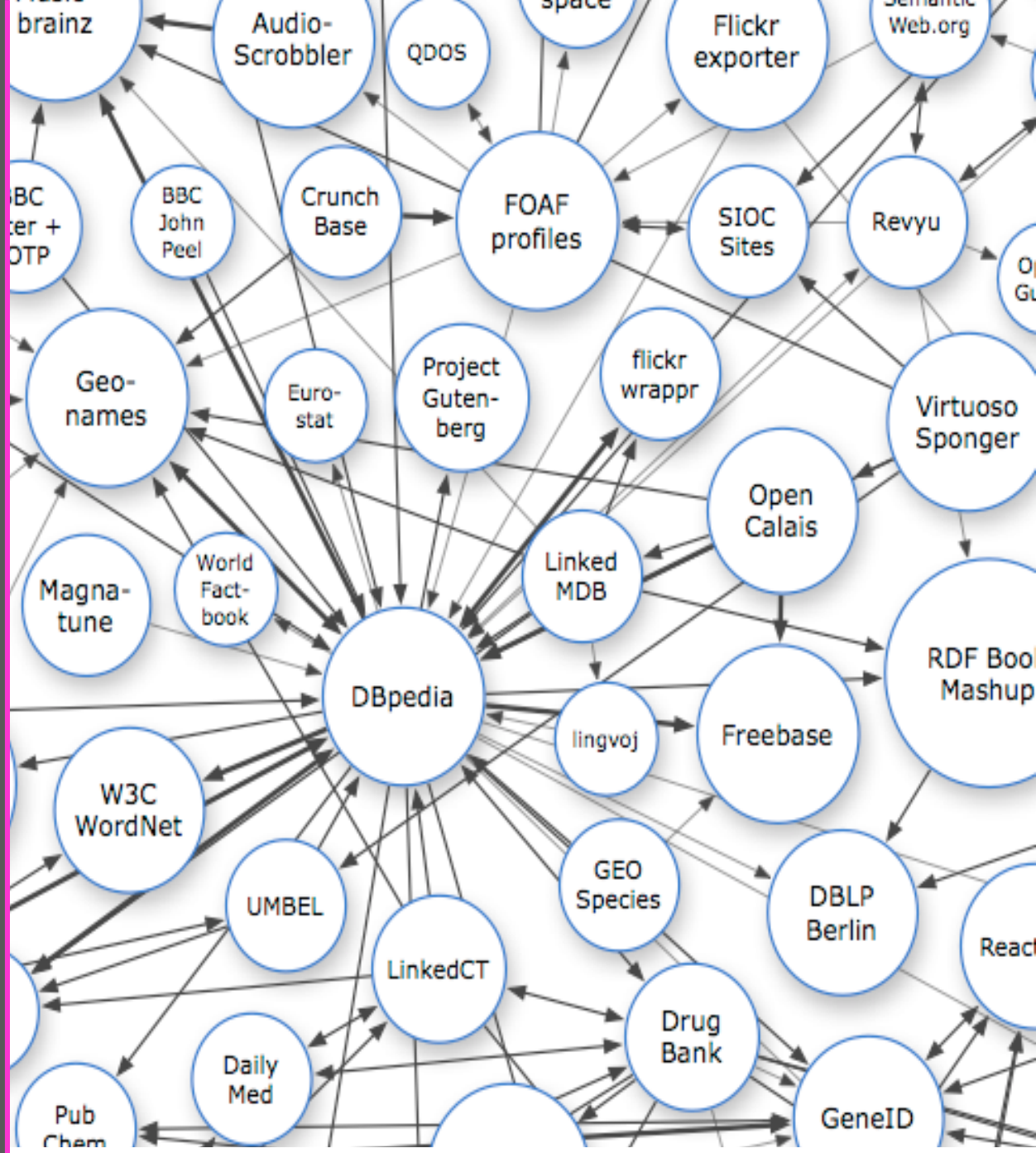


https://www.youtube.com/watch?v=cf7Ie6PMZJk

Linked data

A method of expressing and publishing structured data so that it can be interlinked and queried semantically by both humans and machines.

- Turning the web of documents into a machine-readable web of data
- From the World Wide Web to the Giant Global Graph

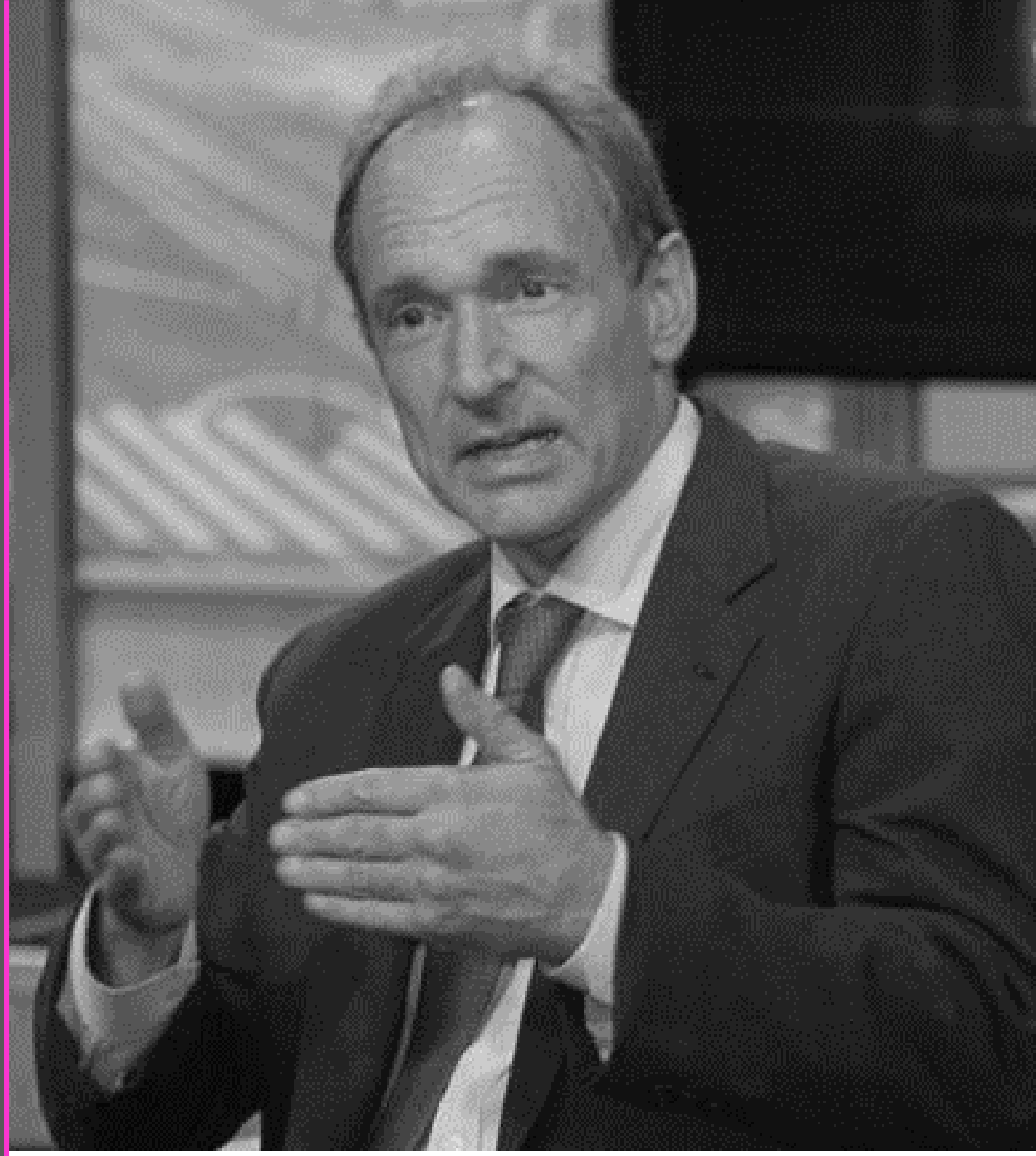


http://linkeddata.org/

Linked data

4 Principles:

- **Use** URIs to name (identify) things.
- **Use** HTTP **URIs** so that these things can be looked up (interpreted, "dereferenced").
- **Provide useful information** about what a name identifies when it's looked up, using open standards such as RDF, SPARQL, etc.
- **Refer to other things** using their HTTP URI-based names when publishing data on the Web.



Tim Berners-Lee. John S. and James L. Knight Foundation via https://commons.wikimedia.org/wiki/File:Tim_Berners-Lee-Knight-crop.jpg

RDF

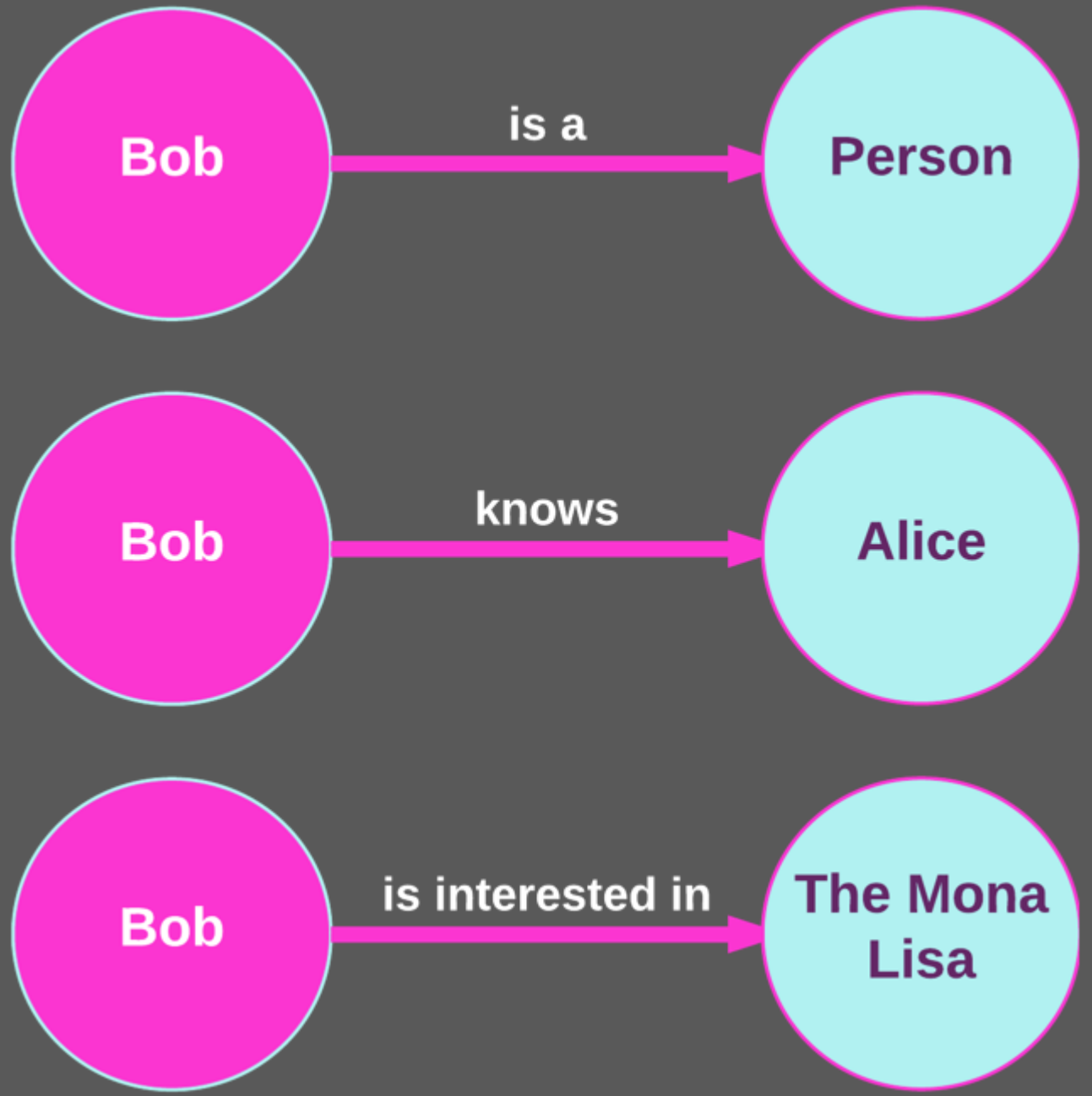
A standard model for data interchange on the web

- Makes statements about resources as expressions
- Uses subject → predicate → object structure to form triples

RDF

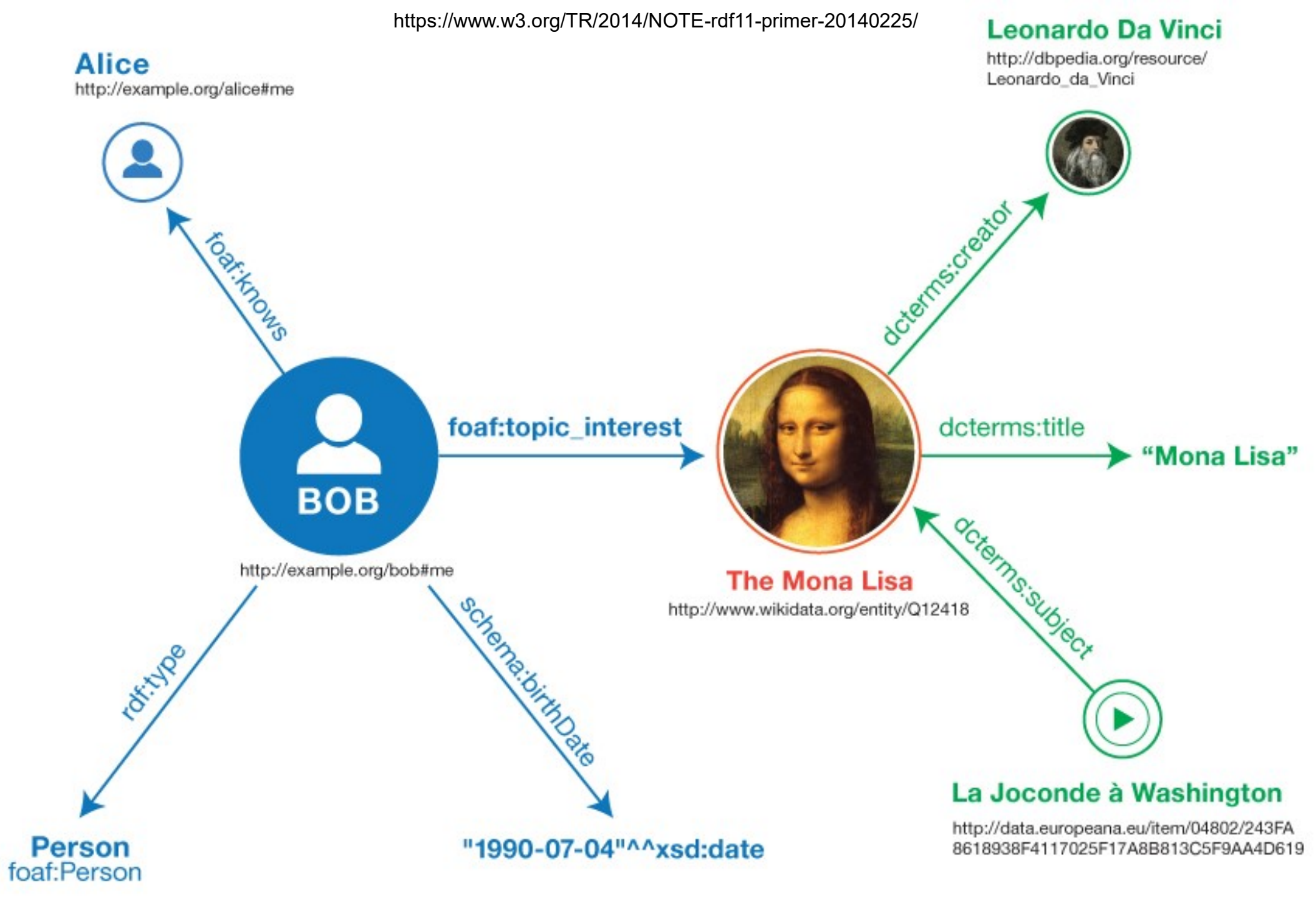
A standard model
for data interchange
on the web

- Uses URIs to express each element of the triple
- Can be serialized into many different formats



RDF

https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/



Linked open data

★ Make your stuff available on the Web (whatever format) under an open license

★★ Make it available as structured data (e.g., Excel instead of image scan of a table)

★★★ Make it available in a non-proprietary open format (e.g., CSV instead of Excel)

★★★★ Use URIs to denote things, so that people can point at your stuff

★★★★★ Link your data to other data to provide context

5-star deployment scheme for open data

http://5stardata.info/en/

Linked open data and archives

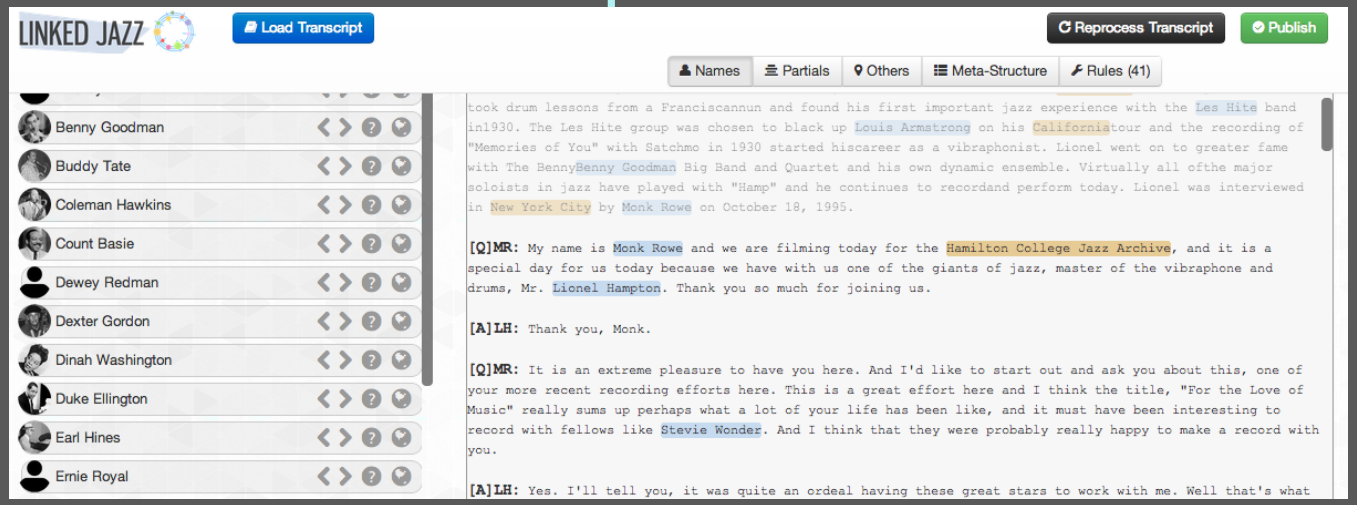
Lots of discussion, many existing useful ontologies, and some implementation in larger portal sites. Few institutional implementers out there.

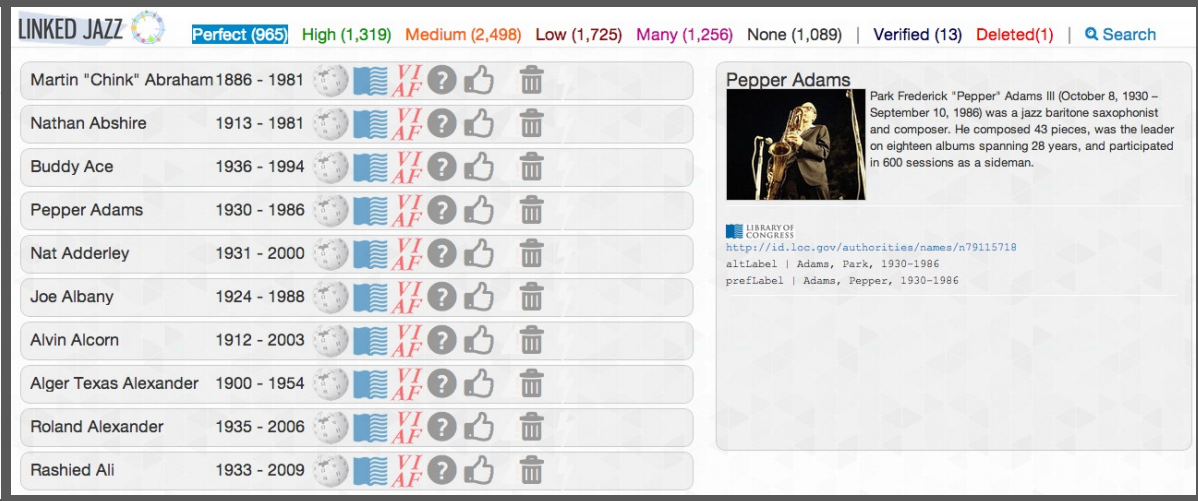
- Europeana Data Model (EU), Archives Hub (UK)
- LoC, W3C, Getty; SKOS, ViAF, GeoNames, etc.
- PREMIS → creating an OWL ontology
- PRONOM ontology? Work on hold
- ICA EGAD developing an entity relationship model that can be expressed semantically?

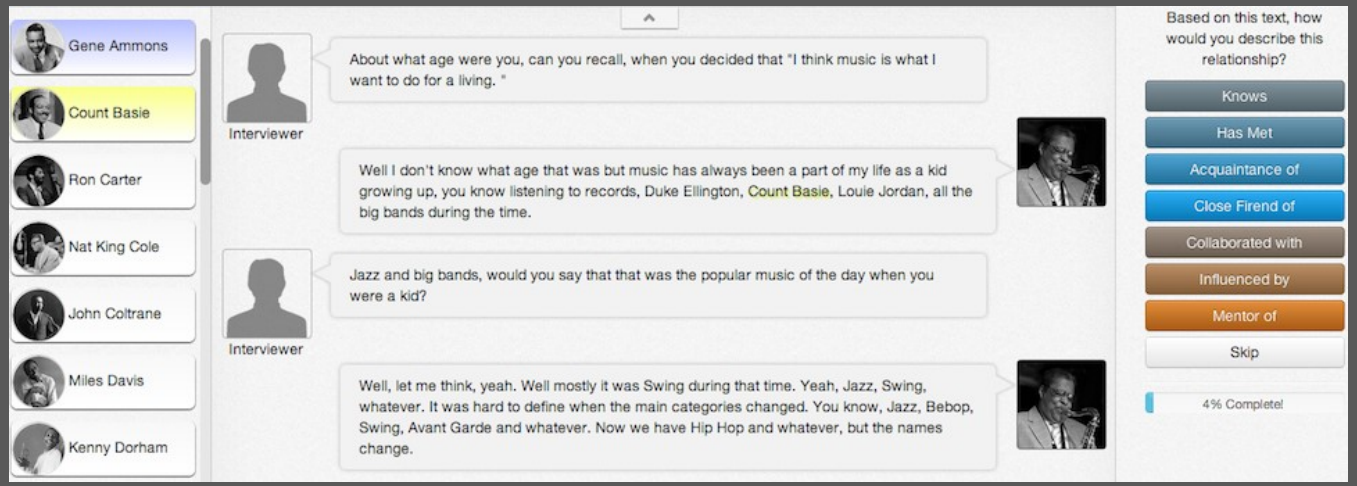
Linked open data and archives

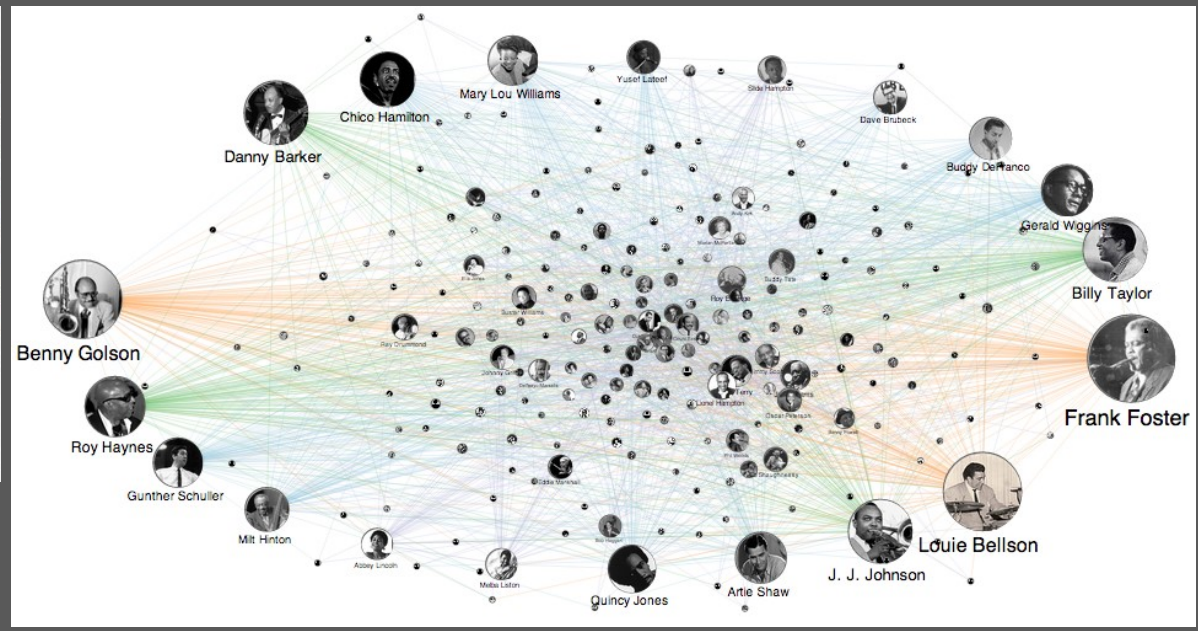
https://linkedjazz.org/

Linked open data and archives









Combining linked open data, machine learning, and crowdsourcing into a suite of user friendly tools

https://linkedjazz.org/tools/

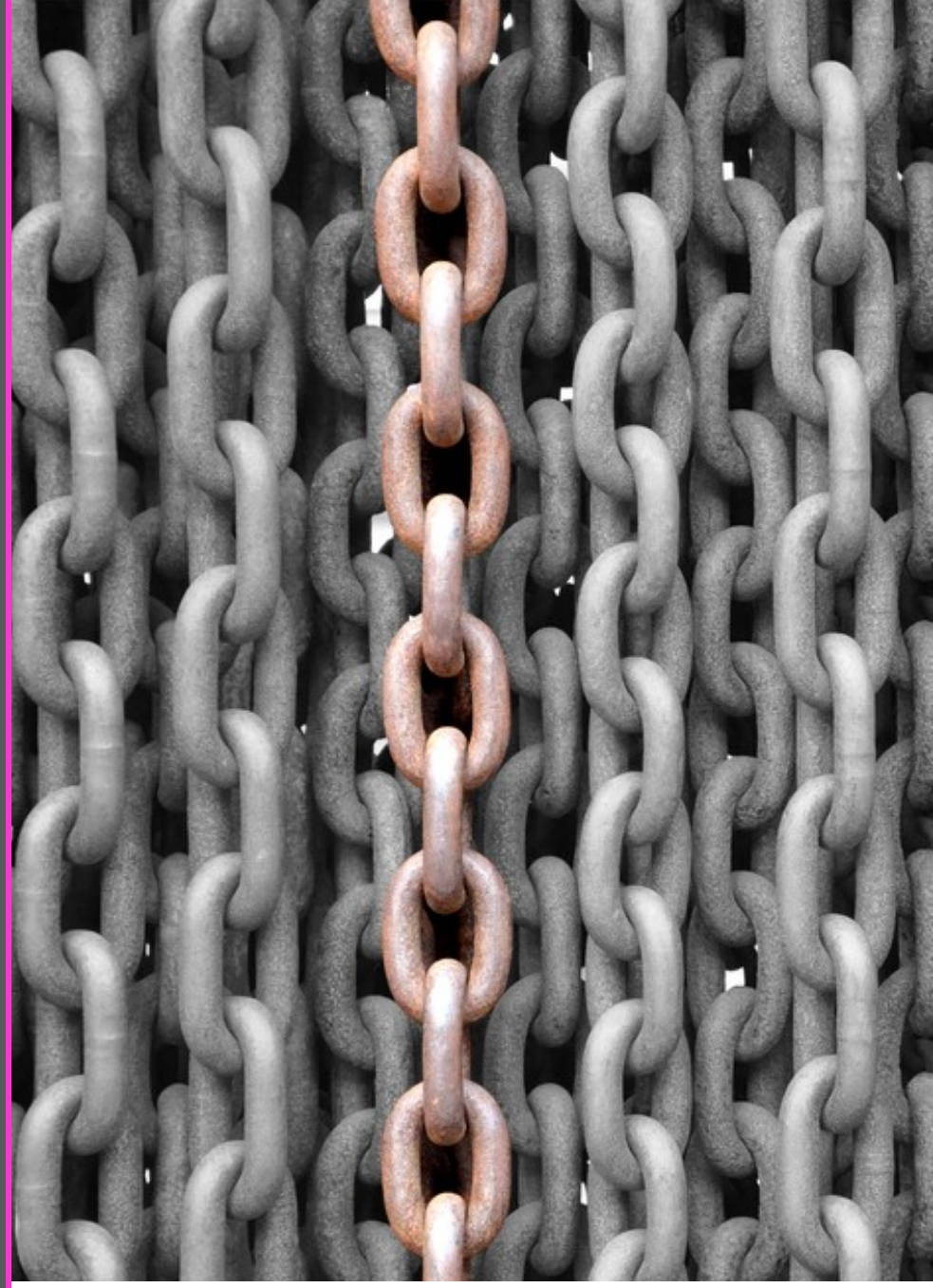
Linked open data and archives

- Description and Access
  - Our systems should connect users to other related information – linked data is the way
  - Relationship visualizations, queries across institutions, domains
  - Crowd-sourced description and reconciling from communities of interest
- Preservation
  - PREMIS ontology + PRONOM ontology + W3C PROV= building blocks for a community Format Policy Registry (FPR)
  - Ability to generate community-wide preservation stats

Blockchain technology

Blockchain:

A distributed ledger that uses cryptographic hashing to maintain a continuously growing record of transactions, divided into linked, time-stamped blocks.



Jorge González, “Chains reloaded.”July 29, 2011.
https://www.flickr.com/photos/aloriel/6292261464

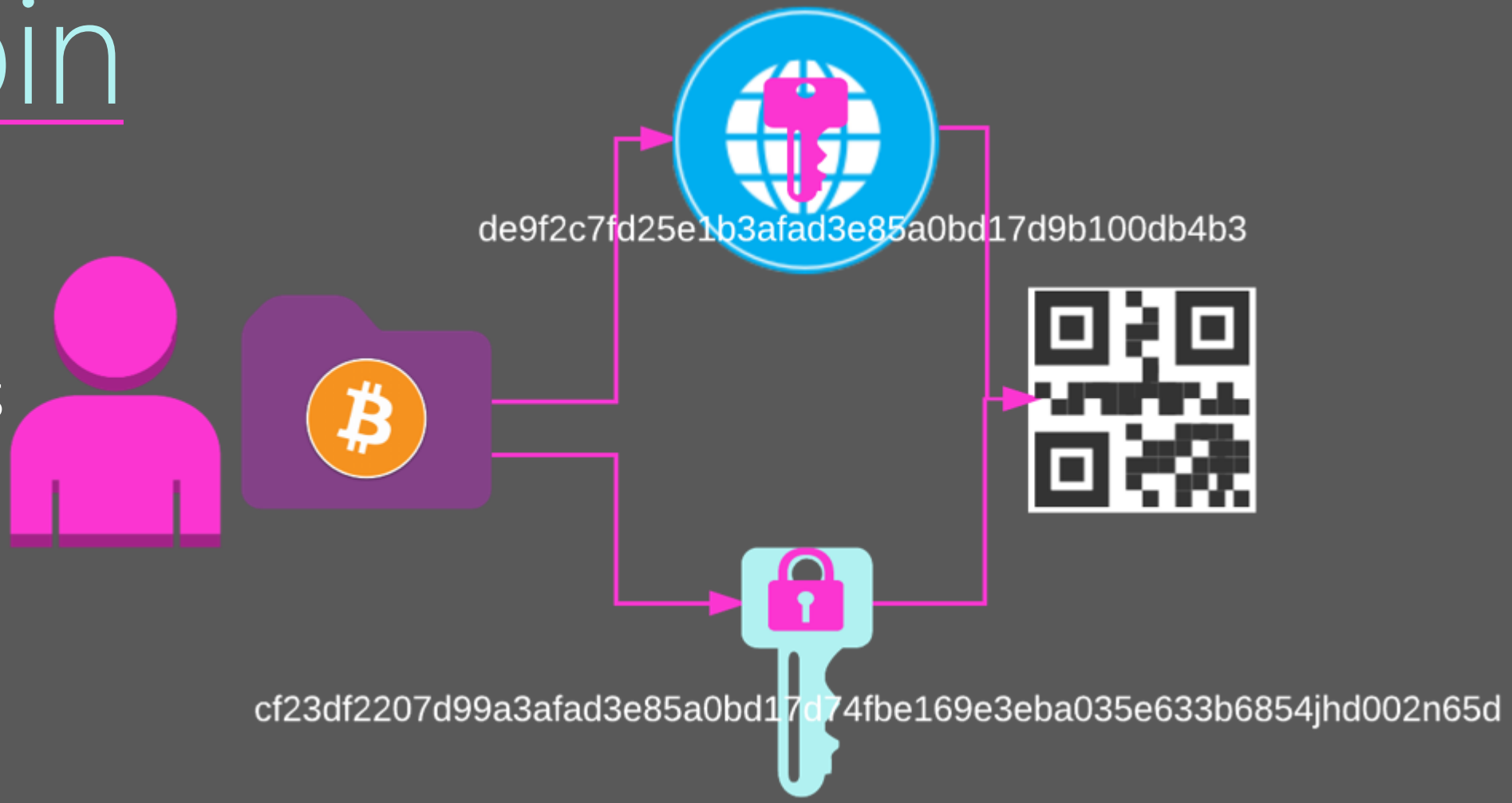
Bitcoin

- An online, anonymous cryptocurrency
- Allows peer-to-peer transactions without an intermediary
- Created 2008, released open-source in January 2009
- Relies on the blockchain:
  - Transactions are verified by nodes on the network
  - Nodes are rewarded for verification
  - Once transactions are verified, they are added to a block every 10mins

https://www.flickr.com/photos/jason_benjamin/12795733984

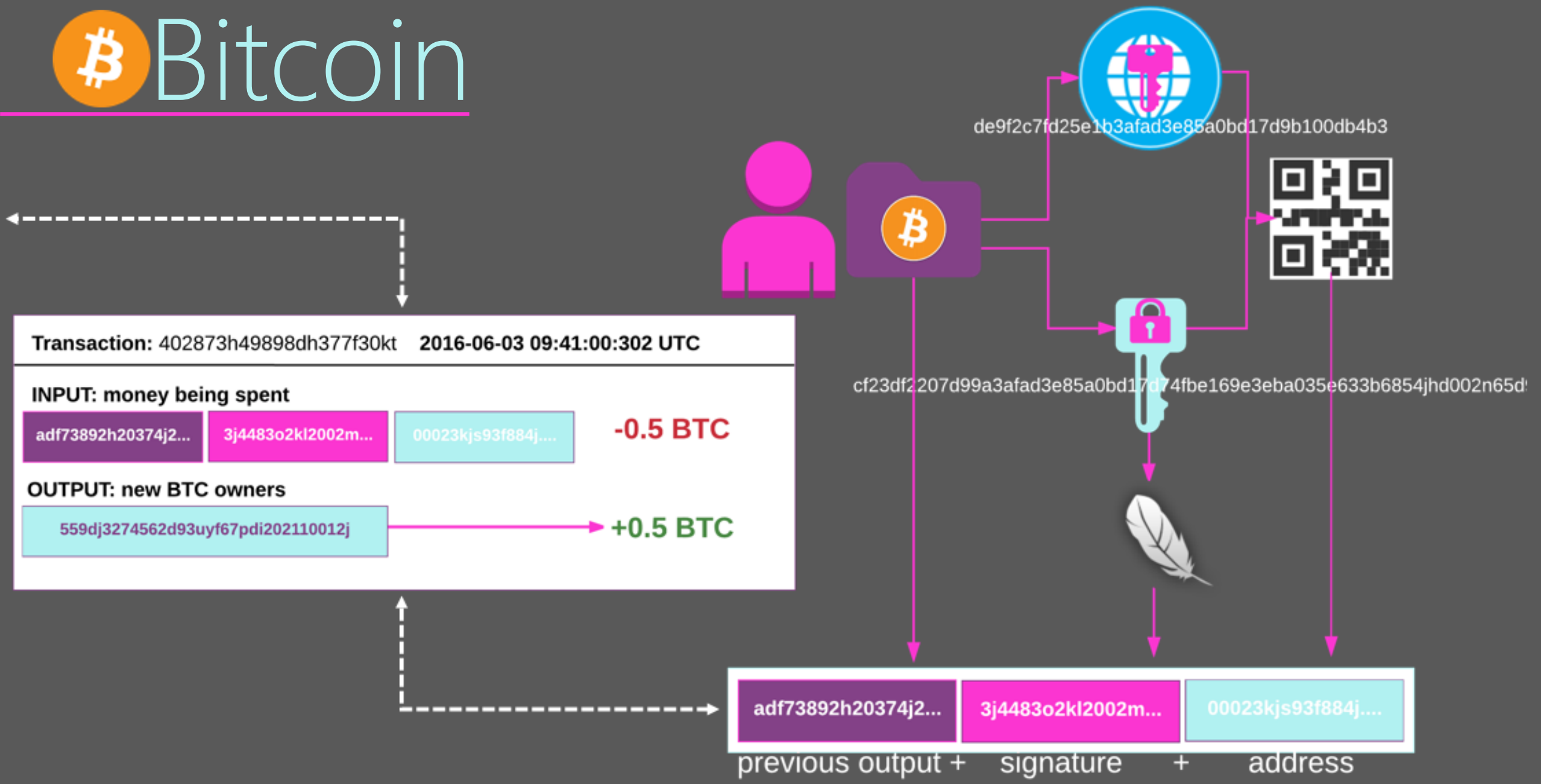
Bitcoin

User installs a Bitcoin client (a wallet), and uses it to create a bitcoin address for a transaction

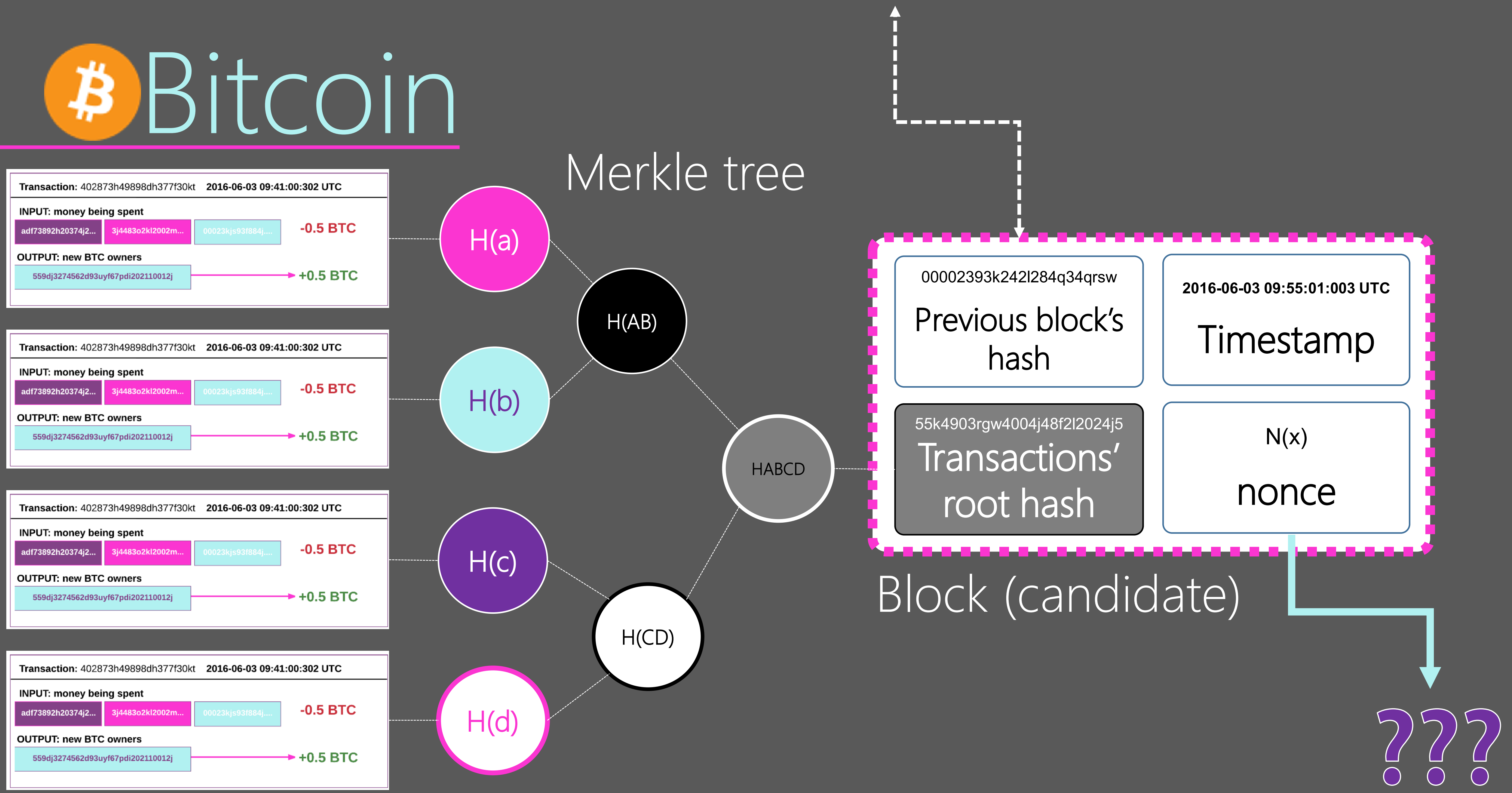


- A wallet can have as many BTC addresses as desired
- An address is created via public/private key cryptography

Bitcoin


de9f2c7fd25e1b3afad3e85a0bd17d9b100db4b3
Transaction: 402873h49898dh377f30kt 2016-06-03 09:41:00:302 UTC
INPUT: money being spent
adf73892h20374j2...
3j4483o2kl2002m...
00023kjs93f884j....
-0.5 BTC
OUTPUT: new BTC owners
559dj3274562d93uyf67pdi202110012j
+0.5 BTC
cf23df2207d99a3afad3e85a0bd17d74fbe169e3eba035e633b6854jhd002n65d
adf73892h20374j2...
3j4483o2kl2002m...
00023kjs93f884j....
previous output + signature + address

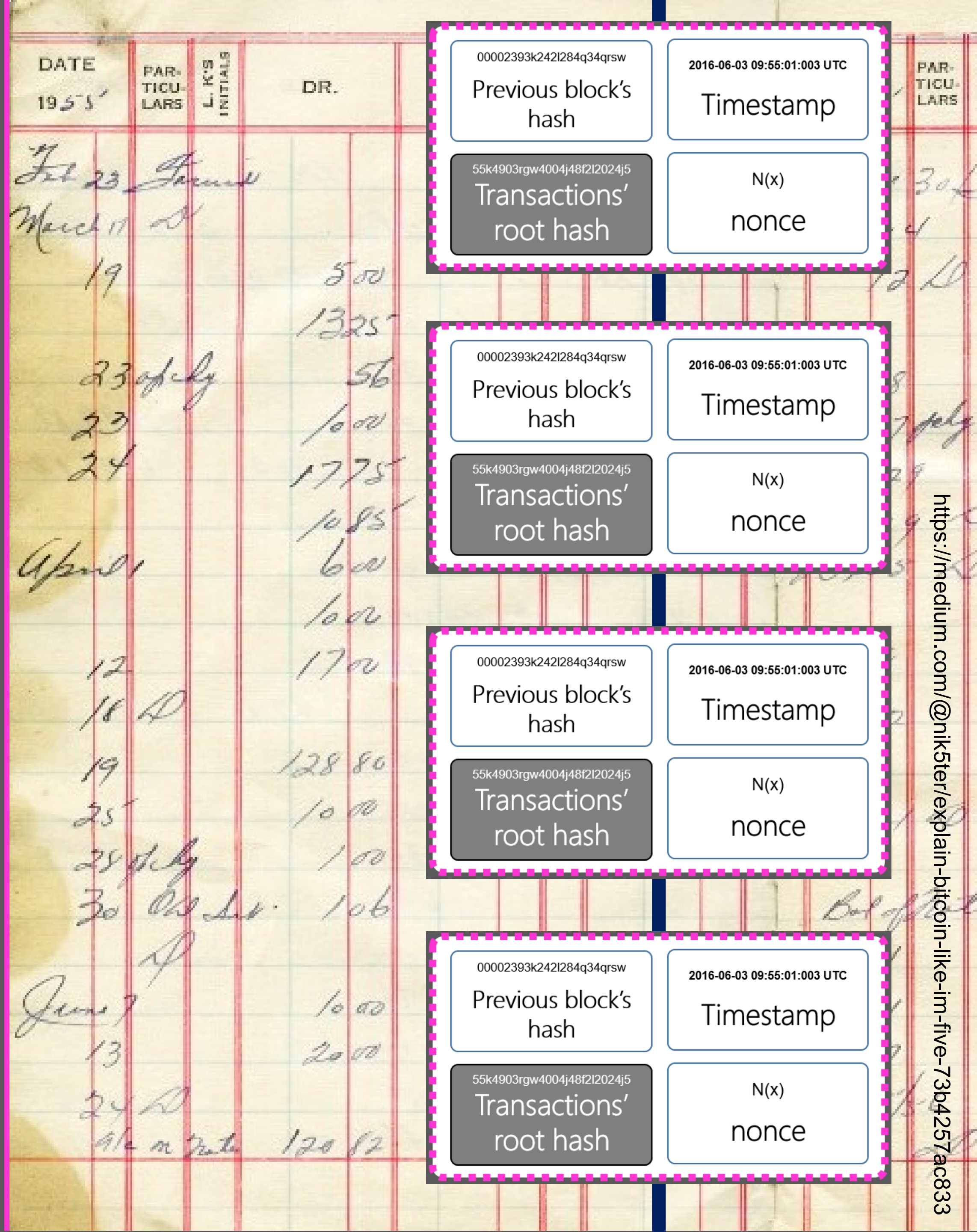
Bitcoin



Previous hash + transaction hash + nonce = new block hash (0000000…)

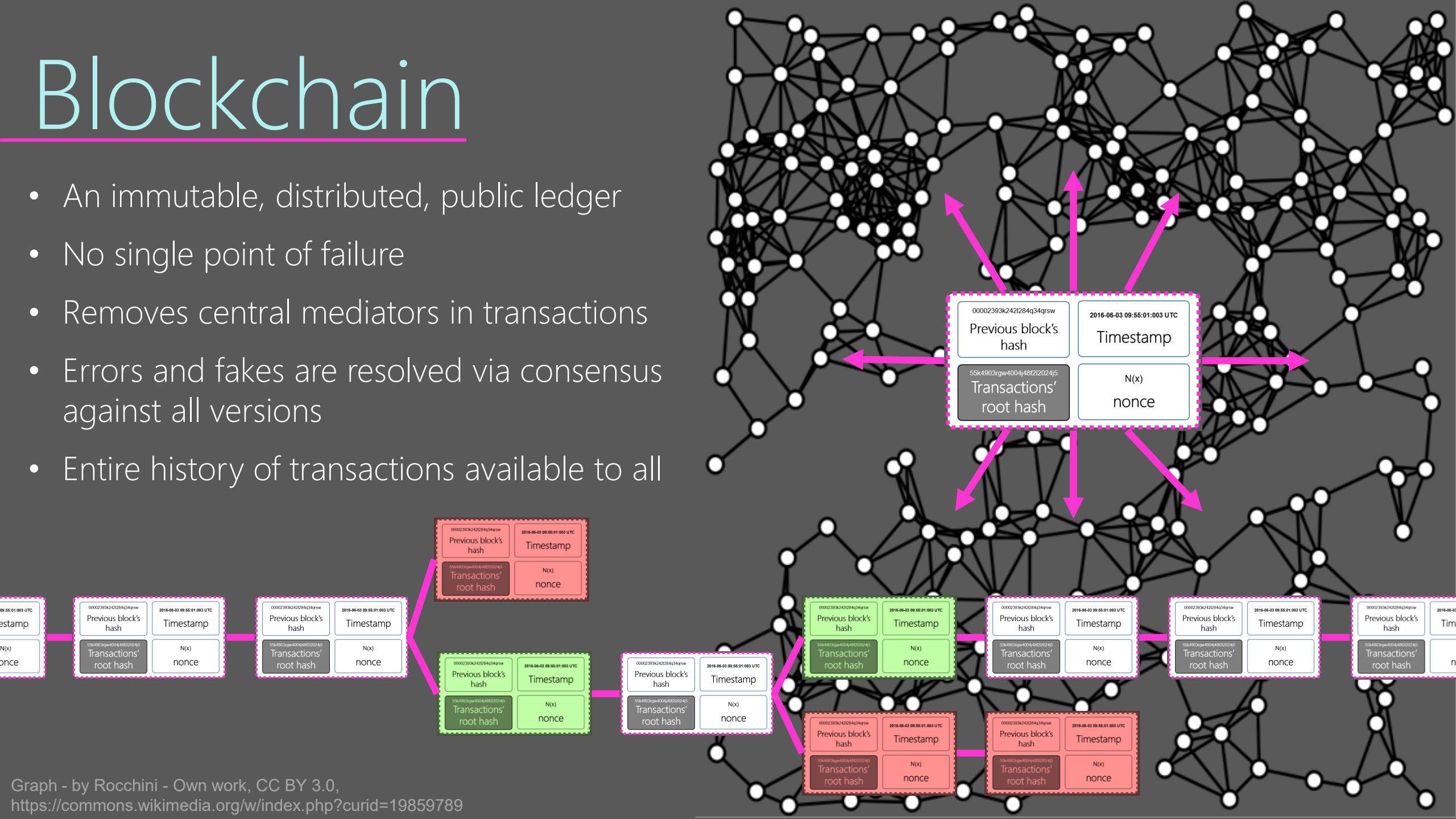
Blockchain

- Miners (peers) try to solve the cryptographic puzzle of the nonce
- First to solve is rewarded with Bitcoin
- Other miners verify the solution
- When there is consensus, the block is added to the blockchain
- New blocks are created every 10 minutes

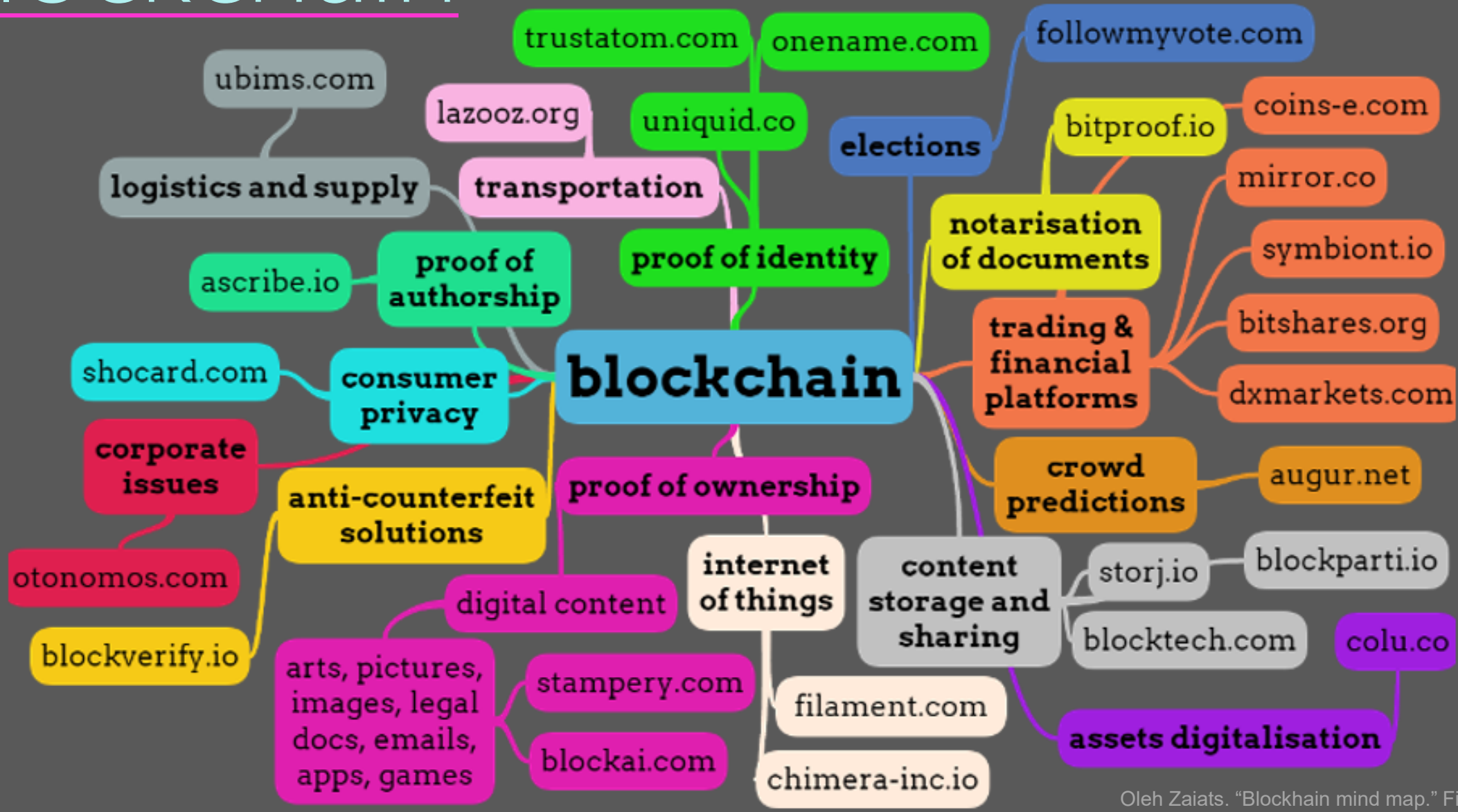


Blockchain

- An immutable, distributed, public ledger
- No single point of failure
- Removes central mediators in transactions
- Errors and fakes are resolved via consensus against all versions
- Entire history of transactions available to all



Blockchain



Oleh Zaiats. “Blockhain mind map.” FinTech, Jan 12, 2016. http://fintechinfo.com/blockchain-mind-map/

Blockchain

https://www.ethereum.org/

Blockchain

Smart contracts

- Contracts are essentially agreements – much like traditional law – often with economic impacts
- Terms and conditions, time frames and consequences, all lend themselves well to code
- Ethereum provides digital currency and a programming language to create distributed, enforceable agreements

http://etherscripter.com/what_is_ethereum.html

Blockchain

STATE OF THE ĐAPPS

http://dapps.ethercasts.com/

Search

219 dapps listed

Sort: Updated

Trustery
Mustafa Al-Bassam
Public Key Infrastructure and identity management system
Working Prototype 2016-05-19

EtherDesign
Ed
A.I. Design Dapp
Live 2016-05-18

Zonafide
Paul Worrall
Trusted Zones: stopping loss due to identity theft
Working Prototype 2016-05-17

KPCS Ethereum
Eric Schulte
The Kimberley Process certificate issuing process
Working Prototype 2016-05-17

ETH Digger
ETH Digger
Investment
Live

Proof of Phone
Igor Barinov
Live 2016-05-16

Crypted RPS
WhySoS3rious
Uncheatable RPS Duels with encrypted hands
Live 2016-05-14

elcoin
Pavel Usenkov & Sergey Primachik
Multifunctional platform based on elCoin - ethereum token
Live 2016-05-14

CO•AKT
Nick Barba , Kevin Kriss
The Future of Crowdsourcing: Rapid Business and Group Formation
Work in Progress 2016-05-12

Decibel.LIVE
Shane Loomb
Real time Noise Monitoring
Work in Progress 2016-05-11

games
2016-05-09

FarmShare
William E Bodell III
Distributed community-supported agriculture (CSA) platform
Work in Progress 2016-05-08

Matching Ethers
WhySoS3rious
Strategy and luck game. Round of 2 teams, Bet upside or flipped
Live 2016-05-07

Etheroll
James Britt
Ether dice game casino / gamble ether
proprietary
Stealth Mode 2016-05-05

Eaterra
Elliot Yeo
Planetary Food Sovereignty
Stealth Mode 2016-05-05

Live

Status Color Key

The background of each Dapp shows a particular color depending on it's state:

- Live
- Working Prototype
- Demo
- Work In Progress
- Concept
- Stealth Mode
- On Hold
- Abandoned
- Unknown

Archival Blockchain uses?

Chain of custody

- An archival blockchain could support the maintenance provenance and authenticity
- Register acquisitions, donor agreements, transfers, preservation actions
- LOCKSS-style P2P backups – use ledger + smart contracts to track who is preserving what where, under what terms – no need for centralized coordination

Archival Services Network

- Institutions donate storage space, compute power, or services
- Are rewarded with "ArchCoin" for use on the network
- These are used to cover network costs, use of other services
- DAOs and smart contracts allow automation of decentralized preservation services

Bringing it all together

Thomas Hawk via https://commons.wikimedia.org/wiki/File:Fireworks_4.jpg

Bringing it all together

https://ipfs.io/

Bringing it all together

Brewster Kahle's Blog

← How about 3 billion people, all living the good life?
Divertissement for Warming Orchestra #D4 →

Locking the Web Open: A Call for a Distributed Web

Posted on August 11, 2015 by jeff kaplan

(Short form article, Short lecture, Long lecture, demo of a fraction of the idea of a distributed website (or paste this link in maelstrom))

Over the last 25 years, millions of people have poured creativity and knowledge into the World Wide Web. New features have been added and dramatic flaws have emerged based on the original simple design. I would like to suggest we could now build a new Web on top of the existing Web that secures what we want most out of an expressive communication tool without giving up its inclusiveness. I believe we can do something quite counter-intuitive: We can lock the Web open.

One of my heroes, Larry Lessig, famously said "Code is Law." The way we code the web will determine the way we live online. So we need to bake our values into our code. Freedom of expression needs to be baked into our code. Privacy should be baked into our code. Universal access to all knowledge. But right now, those values are not embedded in

Search

Recent Posts

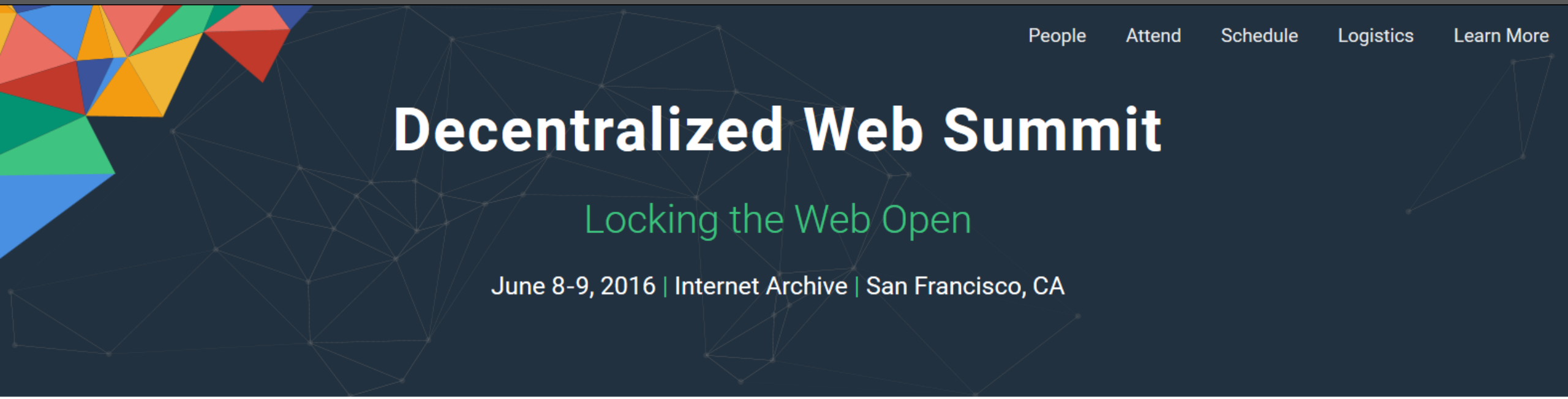
- Successful German Model for Permanently Tenant-Friendly Housing
- Custom School Short Video Description and Short Video by Kid having done it
- Paper on systems like Foundation Housing
- Divertissement for Warming Orchestra #D4
- Locking the Web Open: A Call for a Distributed Web

Recent Comments

- brewster on Successful German Model for Permanently Tenant-Friendly Housing
- Federico on Successful German Model for Permanently Tenant-

http://brewster.kahle.org/2015/08/11/locking-the-web-open-a-call-for-a-distributed-web-2/

Bringing it all together



We invite you to join us at the first Decentralized Web Summit!

New Connections

Call to action

Sharing vision & Tech know-how

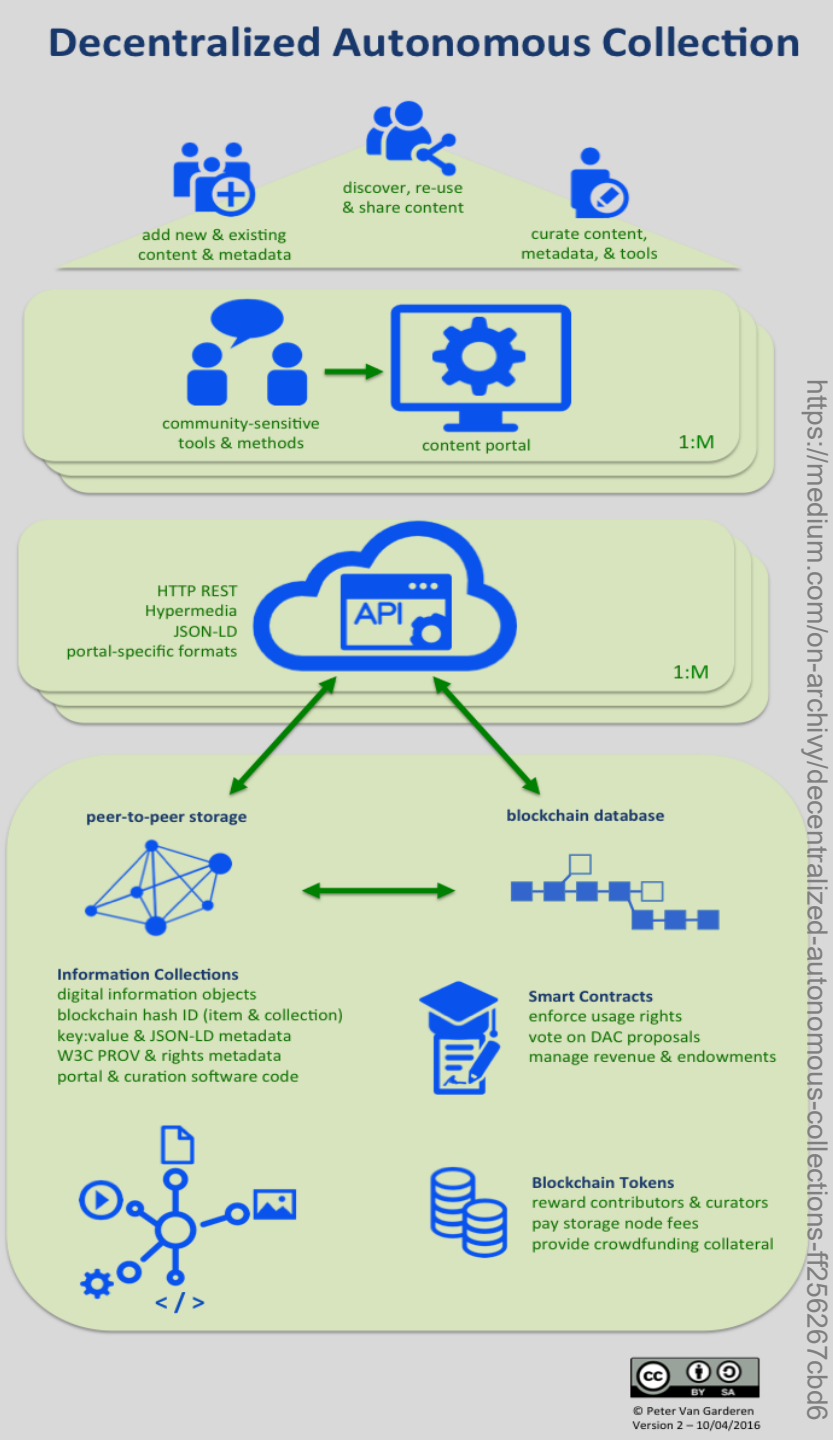
http://www.decentralizedweb.net/

Bringing it all together

"A Decentralized Autonomous Collection is a set of digital information objects stored for ongoing re-use with the means and incentives for independent parties to participate in the contribution, presentation, and curation of the information objects outside the control of an exclusive custodian."

> "...I believe that the emergence of blockchain technology and the concept of Decentralized Autonomous Organizations, alongside the maturation of peer-to-peer networks, open library technology architectures, and open-source software practices offers a new approach to the issues of control, privilege, and sustainability that are inherent to many centralized information collections."

-- Peter Van Garderen, Decentralized Autonomous Collections



https://medium.com/on-archivy/decentralized-autonomous-collections-ff256267cbd6

Will we be a part of this future?

https://commons.wikimedia.org/wiki/File:GT056-Antigua_C-ruins.jpeg

Group questions

- What is the most interesting part of this technology in terms of archives? What potential do you see?
- Where do you think these ideas fall short? What might be some of the challenges?
- How might the ideas shared around this technology impact your institution if it existed now? How might it change the way you carry out these workflows/activities?
- If it **doesn't** seem like it would solve a challenge you face, why not? What commonalities can you find in terms of challenges with others in your group?

Google doc: http://bit.ly/tech-Proche
Twitter: #techProche

Thanks!

Dan Gillean – Artefactual Systems
ACA Conference 2016 – Montréal, QC