Content Addressable Stores

Thomas Kluyver

⚡ Lightning Talks
PyData Edinburgh, October 2019

```
r2pBADkfbgiFZNMepTngfUep9wTm74NghwZh9
NOaBToa1bDiyVkme58rsOpJA7Vs3e7bRHqsSa
6mjFeWvdYbcq88pjRbLSg0Tmpfjy5JDVrb9EP
bbIz3enqmQmxCSZp1zTN2THWNQCDAcIa4ZZew
jbRWOEh0An3K1Ts5AmcQh3sLKbGfZs5JJ7SQb
p8v2mHf4cQkjM8Gl4LubFmoKSwqScmcUPKJ7b
bCiCXpFvdy5Zj8GIUgJqQGQ29peMNDg9tjJTS
fFD36HjGoK9i5oN3UP7fKPoTa1nCSdMUHyVyI
sh0XPE0VyrO18ql2AJkGBegC522RJVFuIH4FP
dxzUTl5k0lpAIY0puQqbhhCgc8LpNP1ez9m2y
O8V6F1YrP3oTkHD4awVBbk8z2HqMmzBrHnCf5
XyKYdLMxQ5zqdh4m9dJXXkGSaYkMuuyXa27Fu
NErbDGzJiz2UxZjUJlurhRzEdkmS60vCh305b
xDR2d46un0wznZmPtlT7gO7mjTudlHQkjcTKn
CgvjC5u2zyWgr3m0aK84sxePFiRzcSIWaCb3K
HBBKYUuWWJ8Hjc6CbSPtWlh9Duk6Xeprl9aAz
ZB4StK6MI9YUgsq9htG81oaAnzMvQD9TQ8myH
v3W4EgQQGvJwWhrkZyAn97KQlKQTyvRia0iDZ
VJiruTeK32dAZZjvk5SbYXX7GQDMlmYJqKlbe
6ooF80O2W3PPsg8NYzKFXaG33Fd5hqQa85lXS
EDuOCoBLZIbwxRXi7kPPRxqWAVI6hxsSAyUGf
yS1UBVucP45onIeVeO4cYnXCcNqY10JQrjI85
x376jYYSaPc1SIUTV3O2Z8ZEYwR3AcssLuMaE
mSIk2p8sQURa0Gidys8nCoE4rdzOuYp4mXHHw
86zQ3sn4SzEOlwYQmyCyZNqcCtInAAB4jPJXX
3AdmO0NCHKJRoJbr7xbh7YaRPdds1XOHpfCat
HFl6aJ2PC6eOlLxZpzopbZU3mtdOo4cu8AbZA
rw1nzjIrZ122PQOctIhT0YdLmlHqRVStvOjLu
UcF8ZjhJAaRapwhouYFRczbbZWOOGRMds5Kic
```

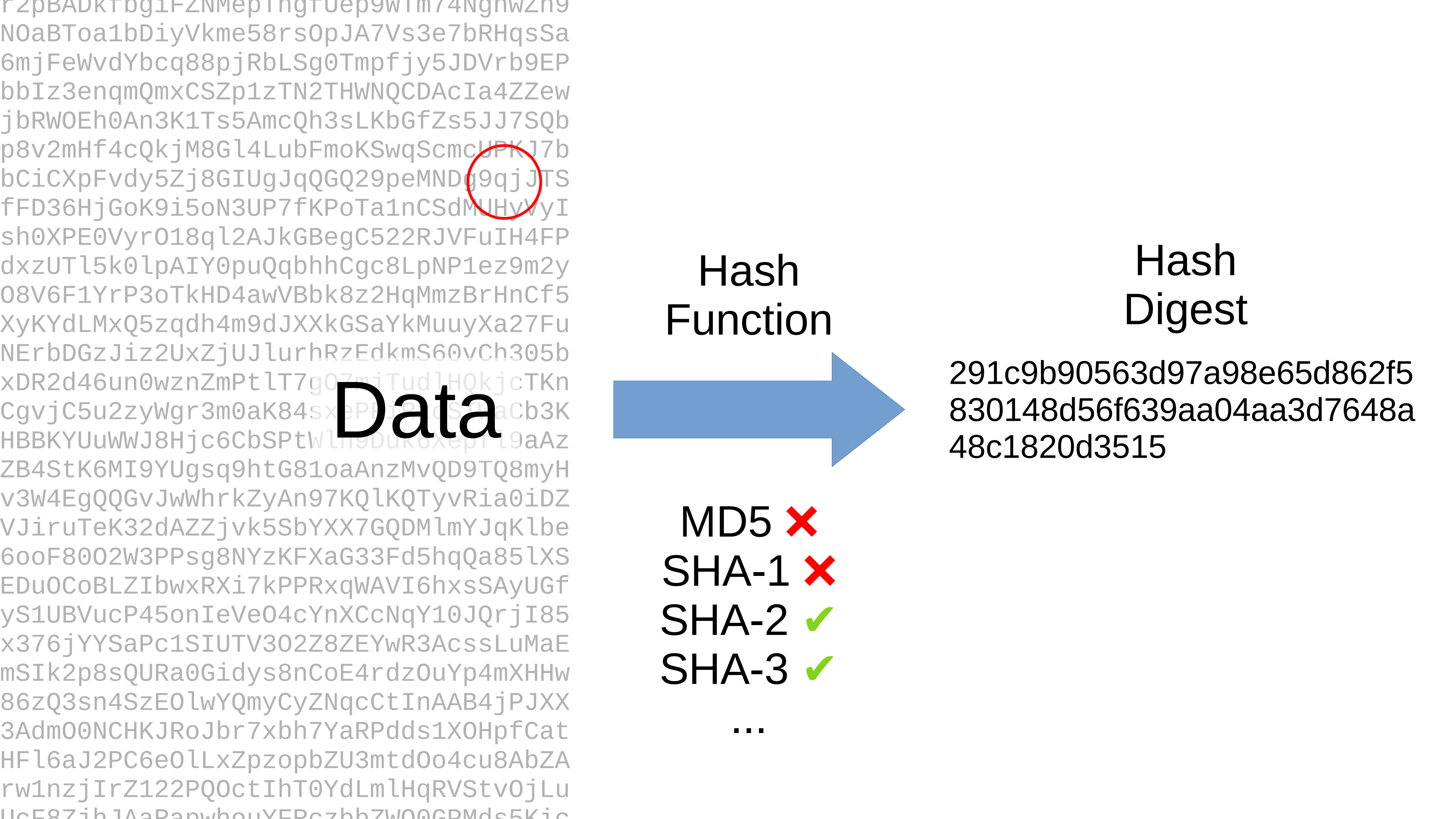
Data

Hash Function

Hash Digest

f125b340930c3d3be26209c7ed
44a2b83059dec7a684ce5b066f
2b5a59abfe0d

MD5 ✗
SHA-1 ✗
SHA-2 ✓
SHA-3 ✓
...



Data

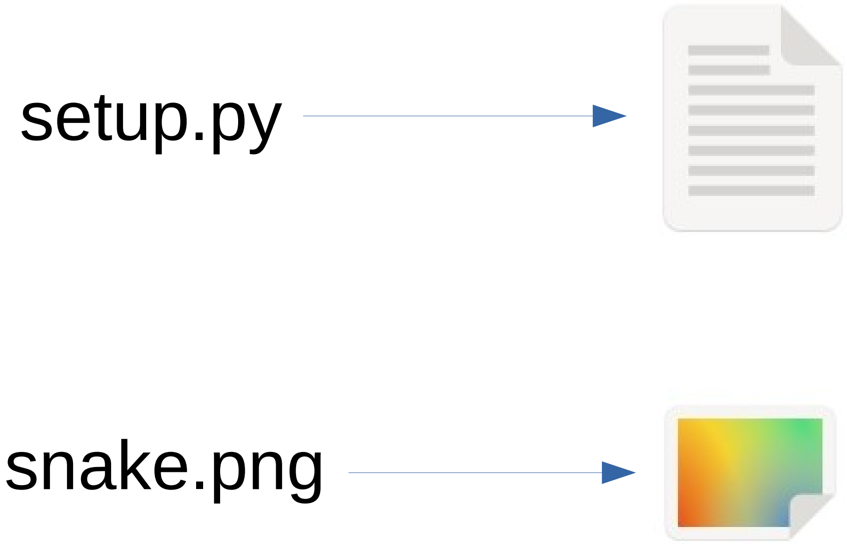
Hash Function

Hash Digest

291c9b90563d97a98e65d862f5
830148d56f639aa04aa3d7648a
48c1820d3515

MD5 ✗
SHA-1 ✗
SHA-2 ✓
SHA-3 ✓
...

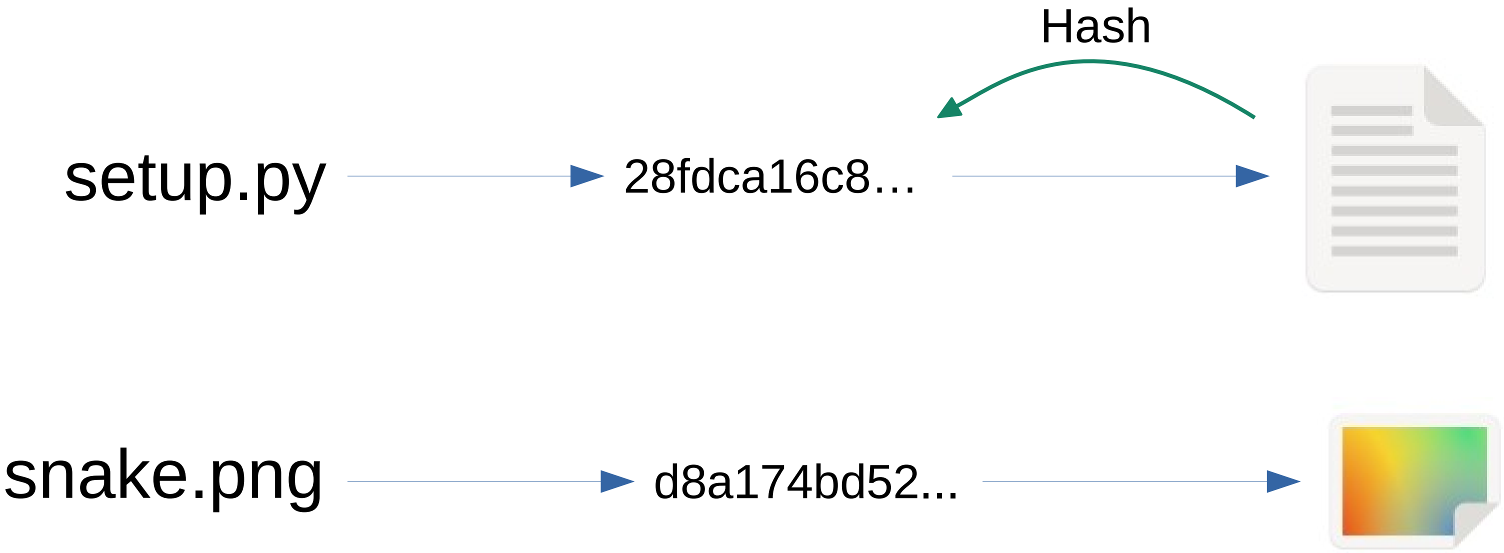
Normal filesystem



setup.py

snake.png

Content addressable store



Hash

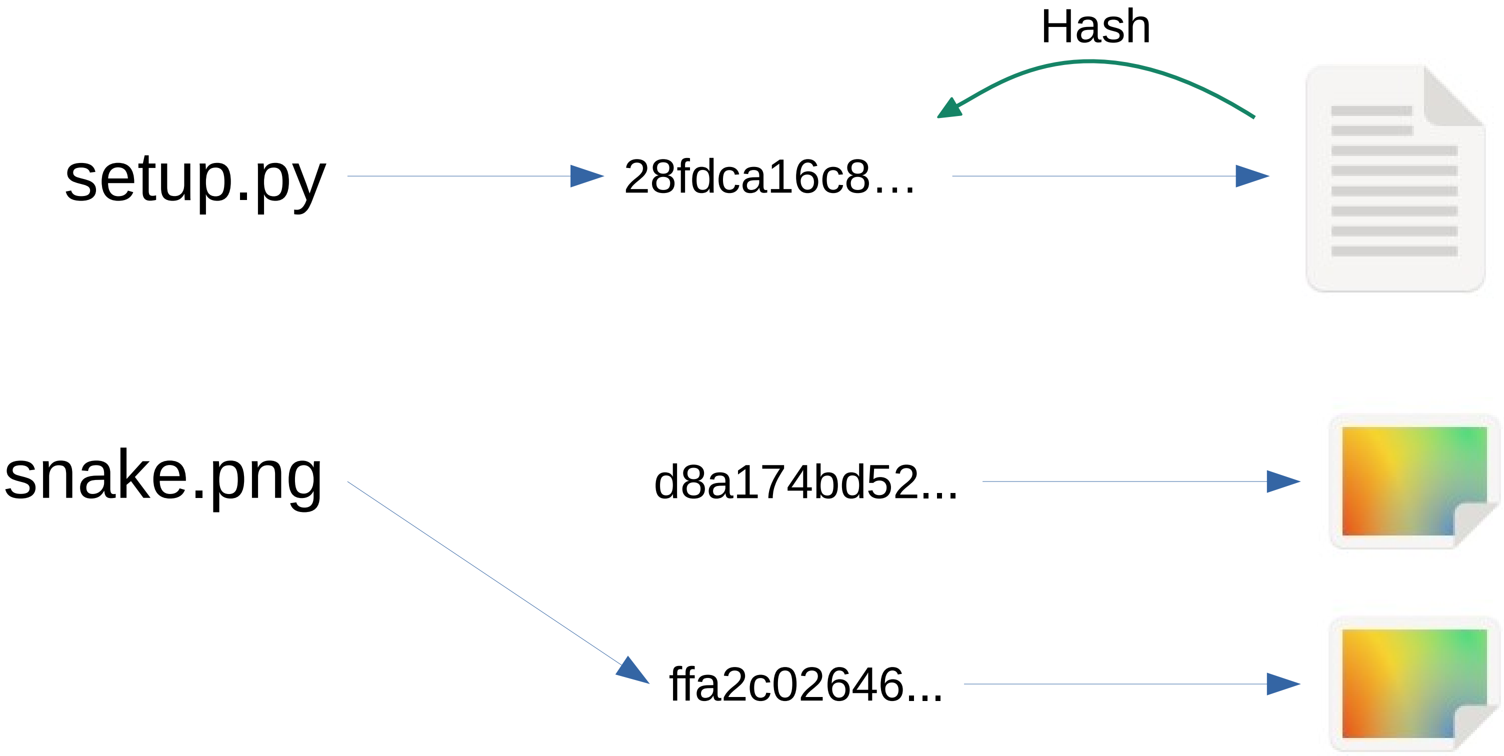
setup.py

28fdca16c8…

snake.png

d8a174bd52...

Content addressable store



Hash

setup.py

28fdca16c8…

snake.png

d8a174bd52...

ffa2c02646...

Knowing our data

- Don’t send data twice
- Don’t store data twice
- Check data is correct

Applications

- Efficient backup (e.g. Borg)
- Git
- Hangar (See Richard Izzo’s talk)
- Caching (e.g. pip)
- Peer-to-peer sharing (e.g. IPFS)