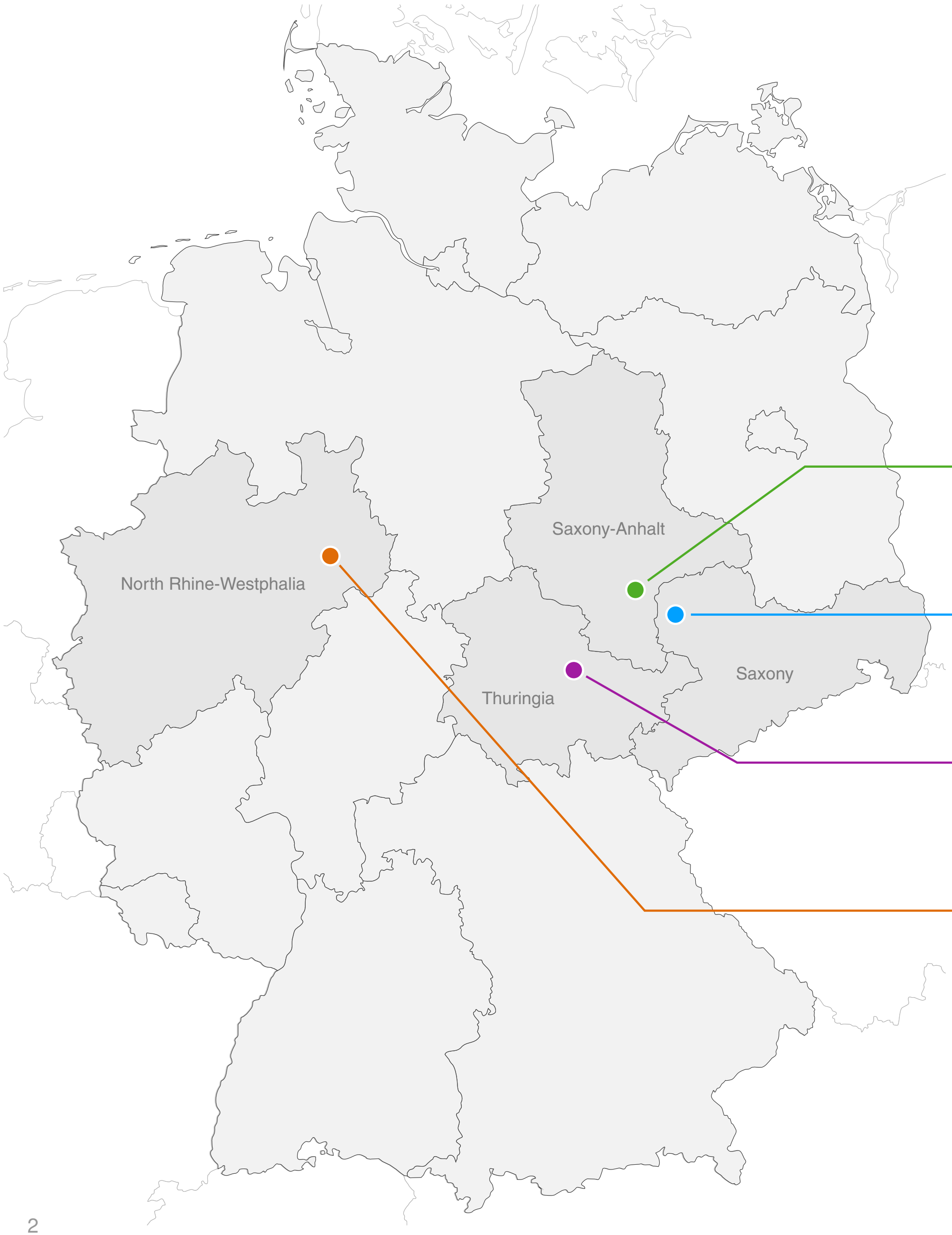
Towards an Open Web Index

Lessons from the Past

Michael Völske
Bauhaus-Universität Weimar
webis.de

Second International Open Search Symposium · Geneva · October 13, 2020



MLU Halle-Wittenberg

Prof. Dr. Matthias Hagen

Leipzig University

Prof. Dr. Martin Potthast

Bauhaus-Universität Weimar

Prof. Dr. Benno Stein

Paderborn University

Prof. Dr. Henning Wachsmuth

Open Web Search

Open Web Search

Challenges

Independence

Scale

User data

Market penetration

Funding

Transparency

Open Web Search

Challenges

Independence

Scale

User data

Market penetration

Funding

Transparency

Independence

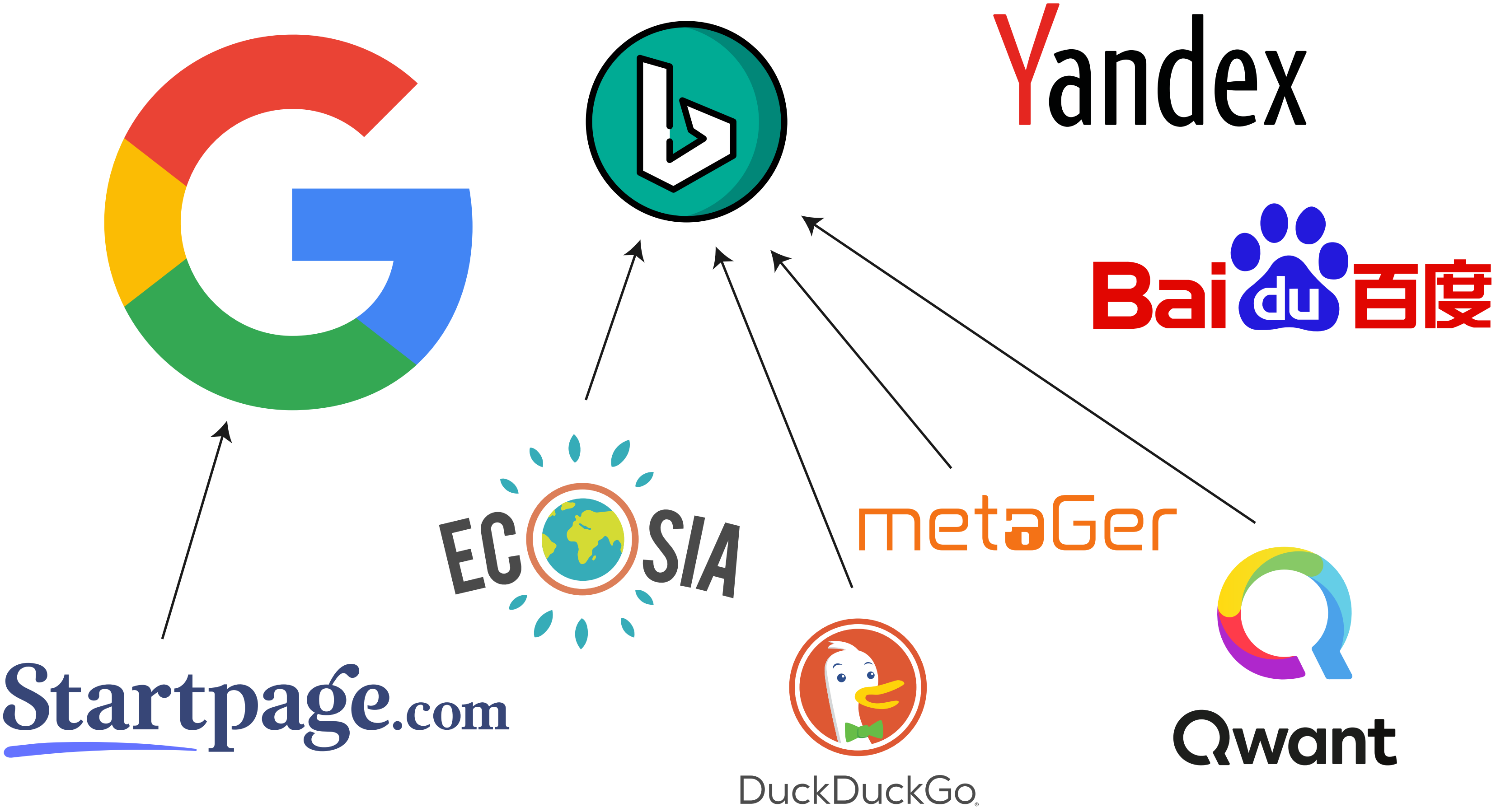
The Search Market in 2020

Independence

The Search Market in 2020

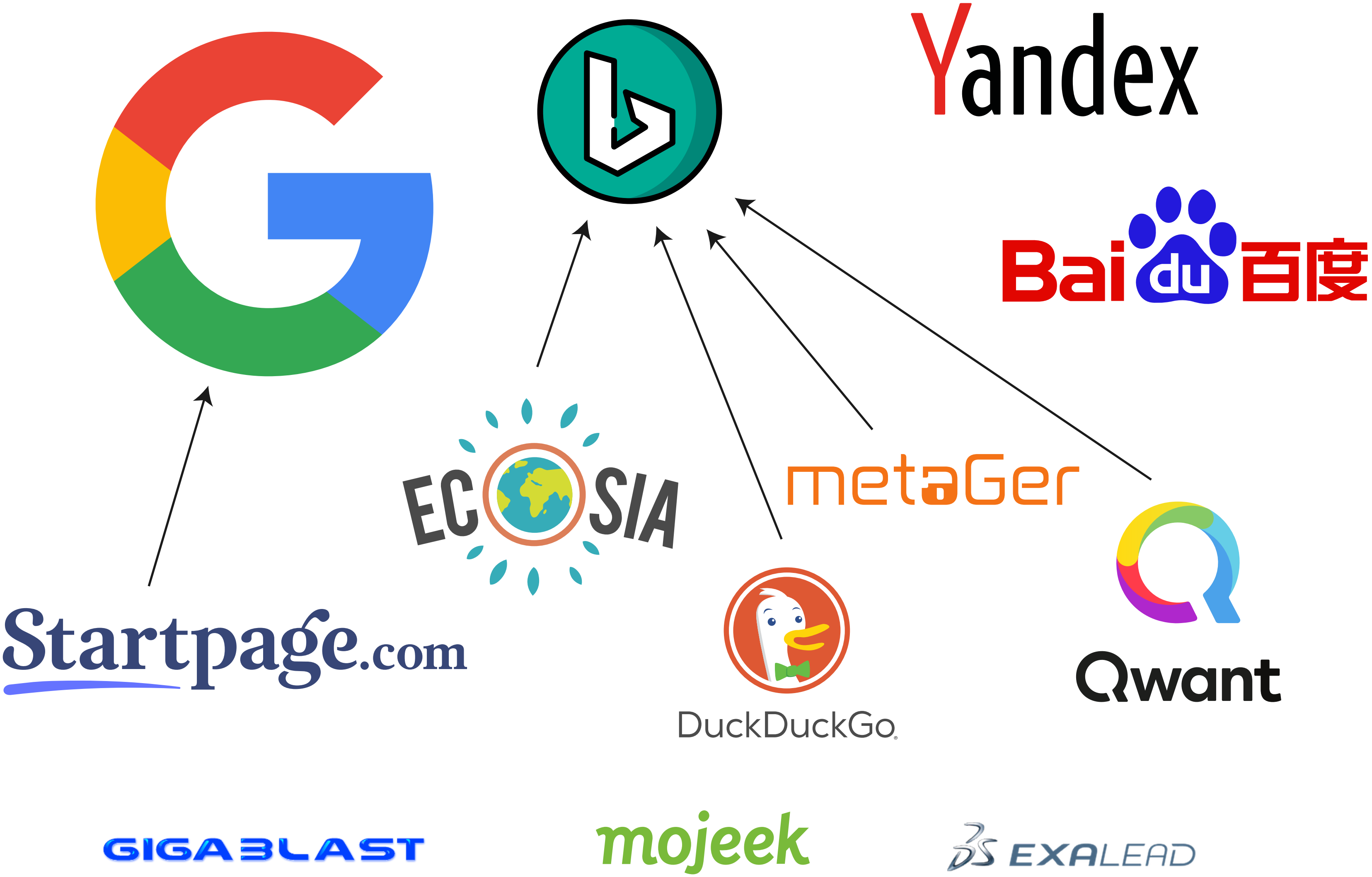
Independence

The Search Market in 2020



Independence

The Search Market in 2020



Scale

The Global Datasphere

The Global Datasphere

“A measure of all new data captured, created, and replicated in a single year.”

[IDC, 2018]

“. . . images and videos on mobile phones uploaded to YouTube, digital movies populating the pixels of our high-definition TVs, security footage at airports and major events such as the Olympic Games, subatomic collisions recorded by the Large Hadron Collider at CERN, banking data swiped in an ATM, transponders recording highway tolls, voice calls zipping through digital phone lines, texting as a widespread means of communications, . . . ”

[IDC, 2012]

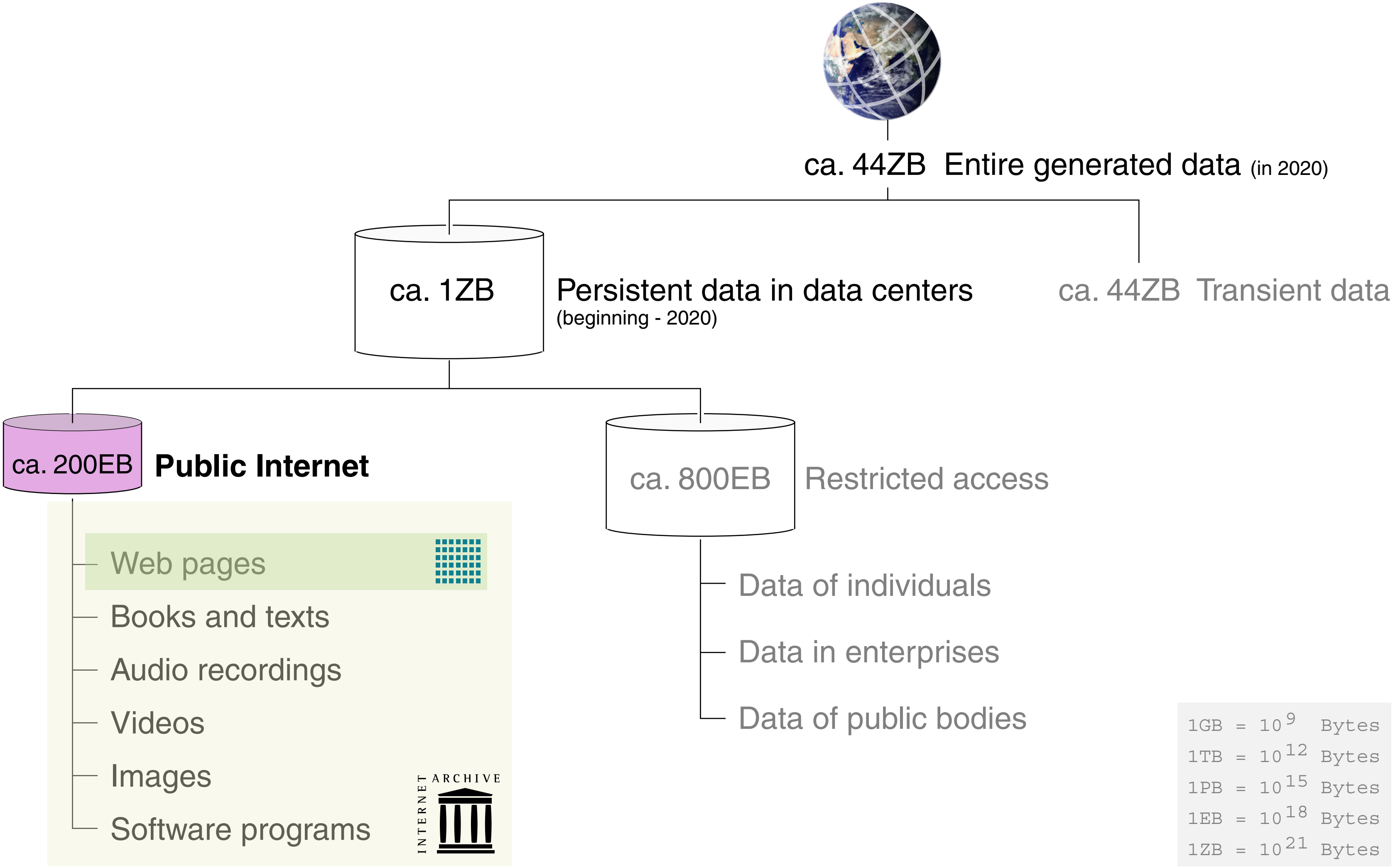
The Global Datasphere

- ca. 44ZB Entire generated data (in 2020)
  - ca. 1ZB Persistent data in data centers (beginning - 2020)
    - ca. 200EB **Public Internet**
      - Web pages
      - Books and texts
      - Audio recordings
      - Videos
      - Images
      - Software programs
    - ca. 800EB Restricted access
      - Data of individuals
      - Data in enterprises
      - Data of public bodies
  - ca. 44ZB Transient data

1GB	=	10^9	Bytes
1TB	=	10^{12}	Bytes
1PB	=	10^{15}	Bytes
1EB	=	10^{18}	Bytes
1ZB	=	10^{21}	Bytes

Basis: IDC (2014-20) • Seagate (2018-20) • Cisco Systems (2018) • Statista (2020) • Domo Inc. (2018-20)

The Global Datasphere



ca. 44ZB Entire generated data (in 2020)

ca. 1ZB Persistent data in data centers (beginning - 2020)

ca. 44ZB Transient data

ca. 200EB **Public Internet**

- Web pages
- Books and texts
- Audio recordings
- Videos
- Images
- Software programs

INTERNET ARCHIVE

ca. 800EB Restricted access

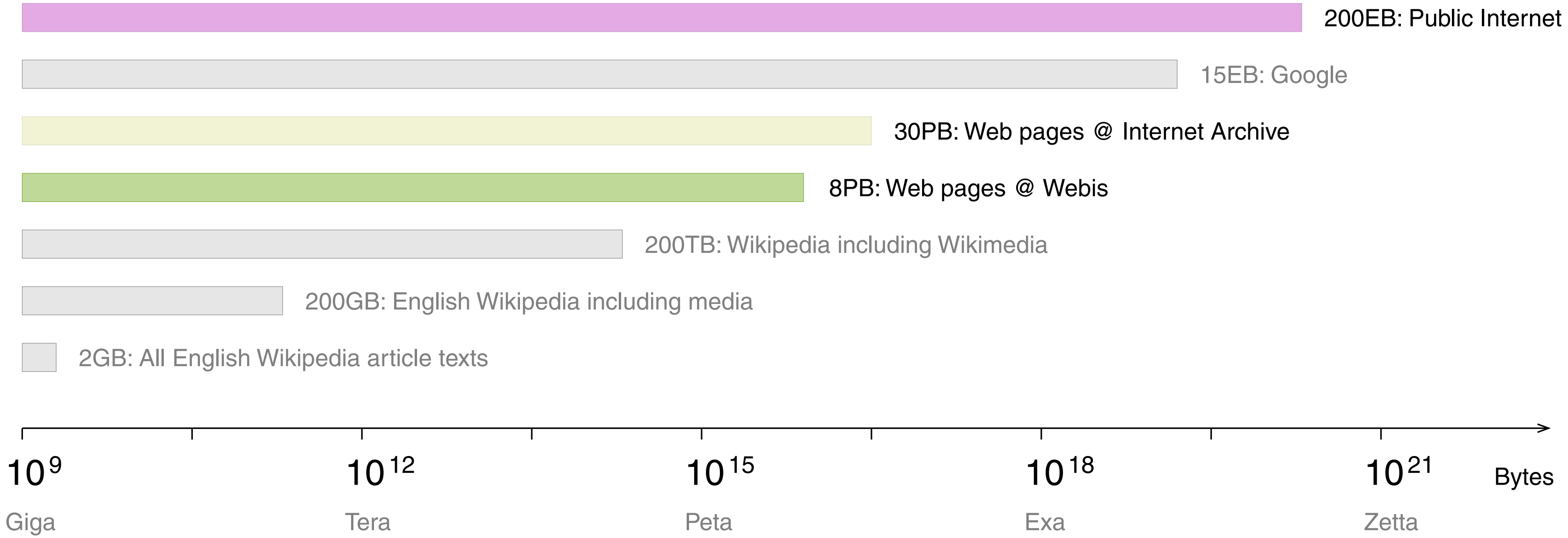
- Data of individuals
- Data in enterprises
- Data of public bodies

1GB	=	10^9	Bytes
1TB	=	10^{12}	Bytes
1PB	=	10^{15}	Bytes
1EB	=	10^{18}	Bytes
1ZB	=	10^{21}	Bytes

Basis: IDC (2014-20) • Seagate (2018-20) • Cisco Systems (2018) • Statista (2020) • Domo Inc. (2018-20)

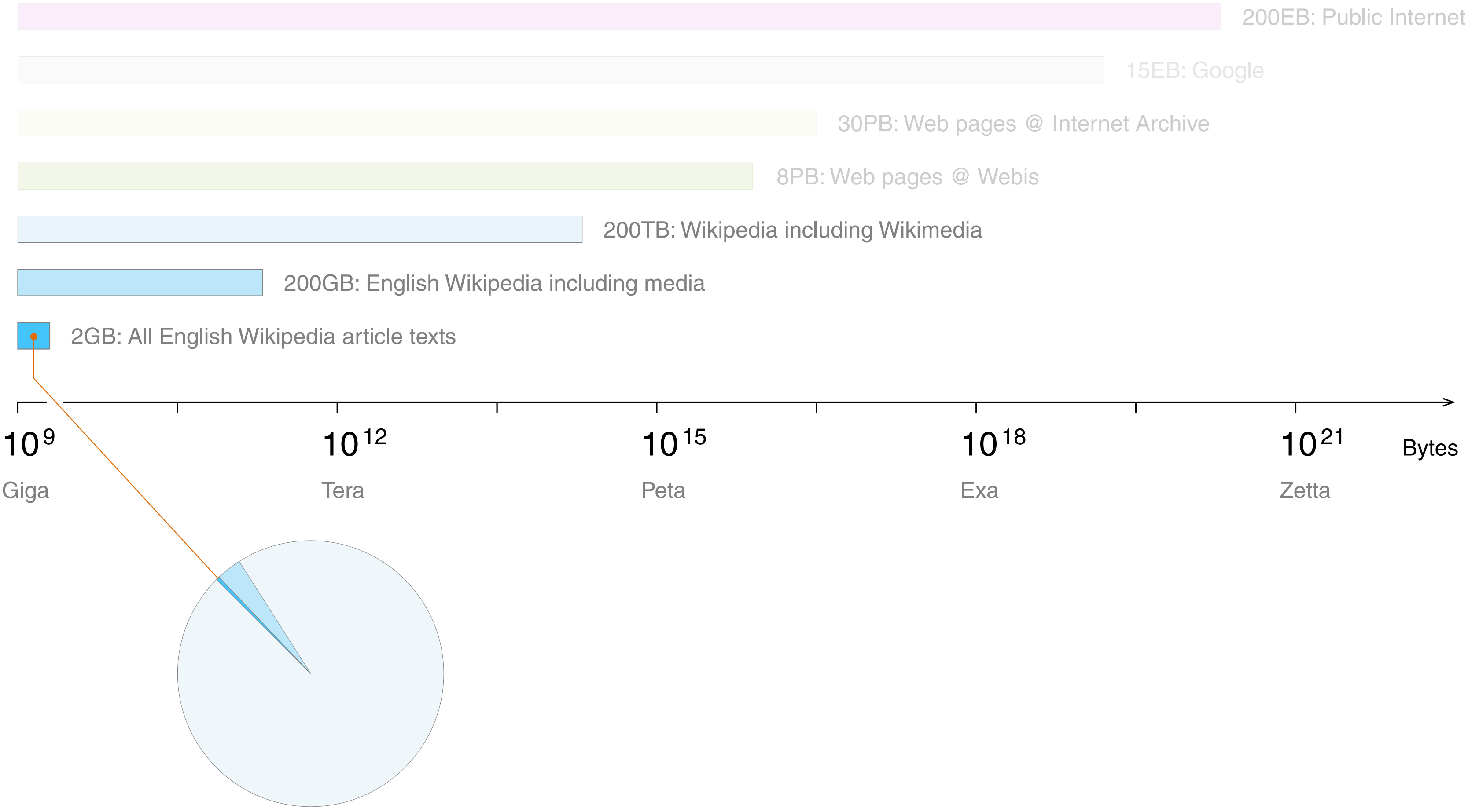
The Global Datasphere

Relating Data Source Sizes



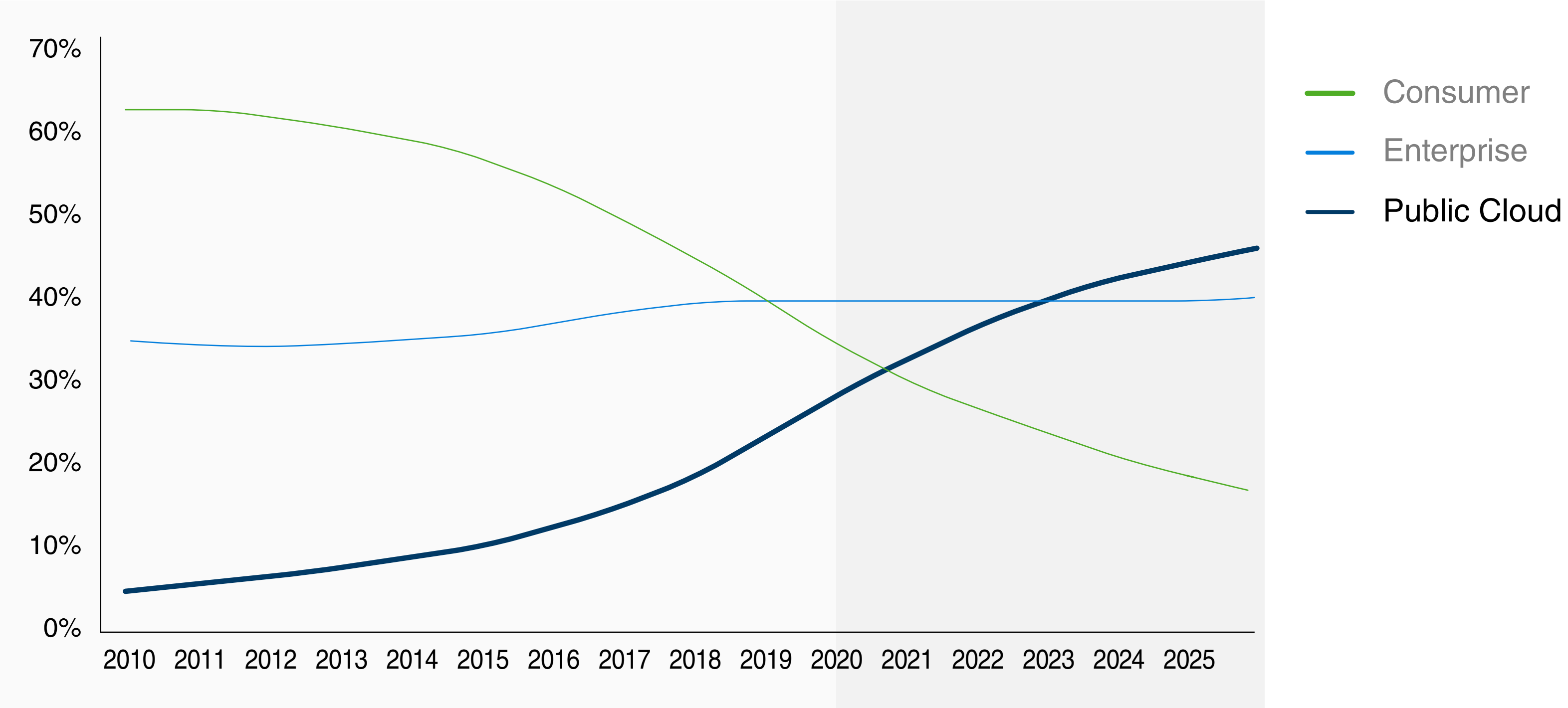
The Global Datasphere

Relating Data Source Sizes



The Global Datasphere

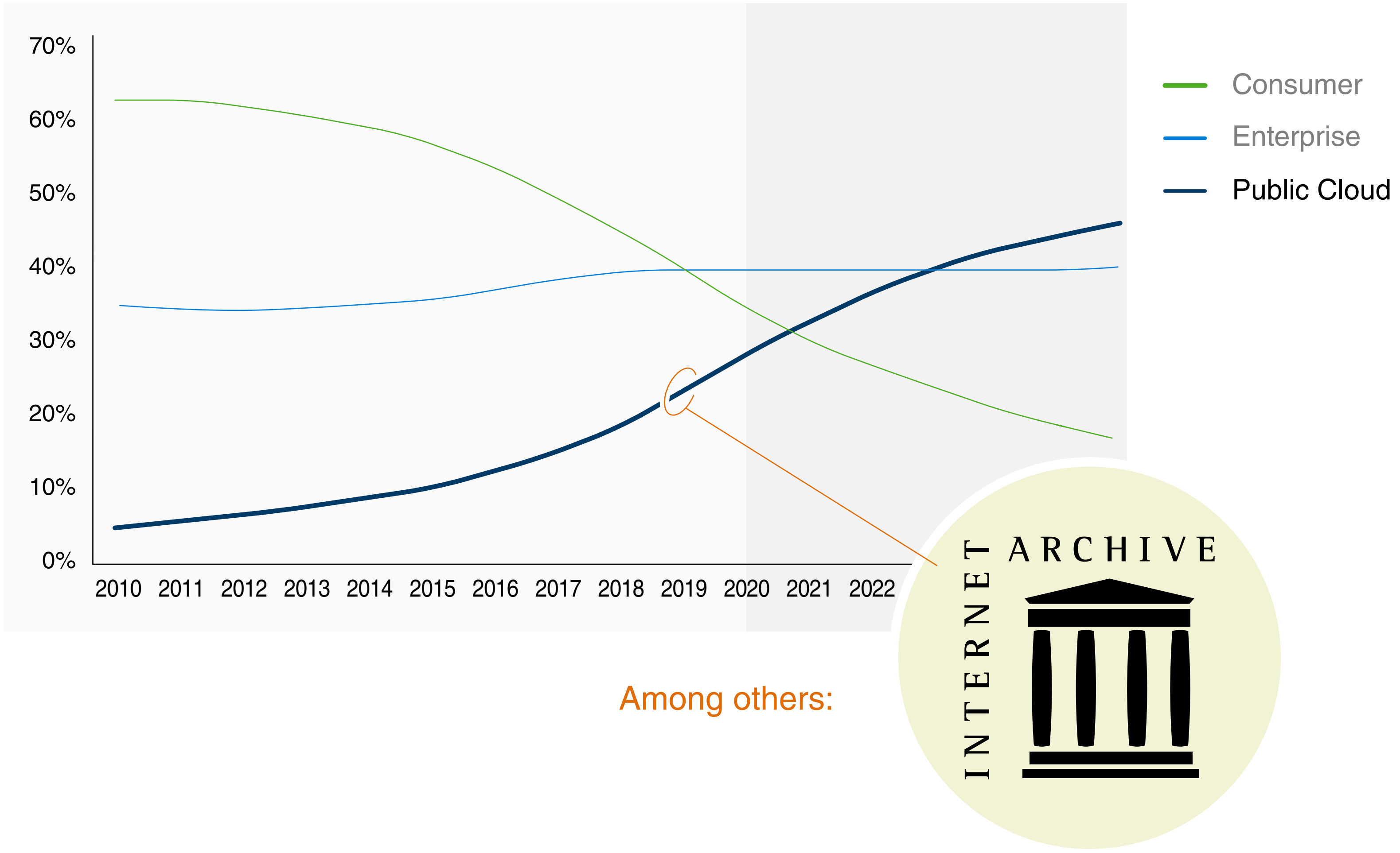
Where is the Data Stored?



Basis: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, May 2020.

The Global Datasphere

Where is the Data Stored?



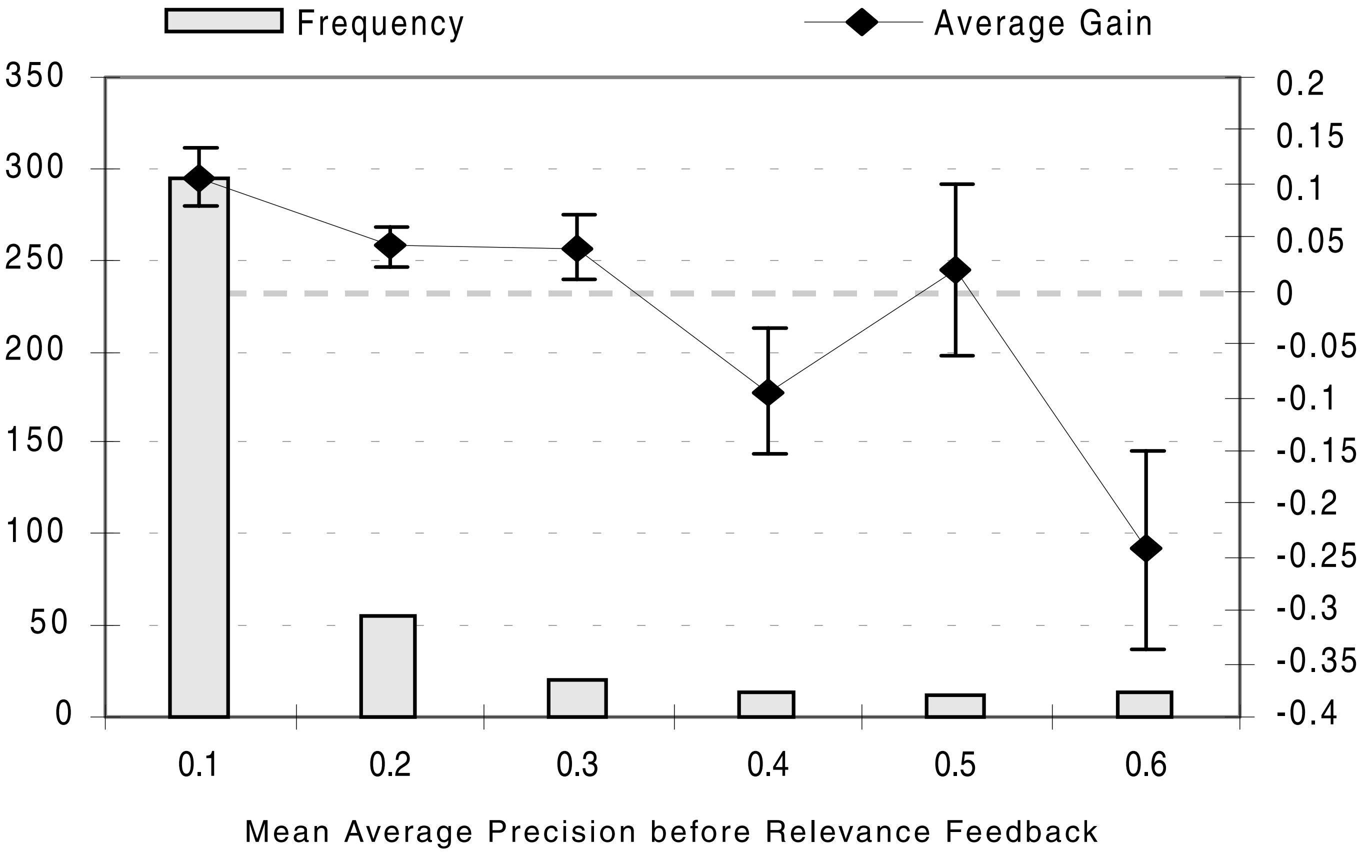
Among others:

Basis: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, May 2020.

User Data

User Data

Benefits of Implicit Relevance Feedback



Basis: Agichtein et al., Improving Web Search Ranking by Incorporating User Behavior Information, SIGIR (2006)

User Data

Risks: Record Linkage

Timestamp	UID	URL
2020-10-13 14:06	A1234	sportsnews.com/sports/
2020-10-13 14:07	A1234	news.com/myaccount/solso
2020-10-13 14:07	A1234	ecommerce.com/receipt_id=1234054
2020-10-13 14:08	A1234	sportsnews.com/sports
2020-10-13 14:06	B5678	sportsnews.com/sports
2020-10-13 14:07	B5678	sportsnews.com/sports

Basis: `https://0x65.dev/blog` • Human Web—Collecting Data in a Socially Responsible Manner (2019)

User Data

Risks: Record Linkage

Domain	Unique Visitors
sportsnews.com	2
news.com	1
ecommerce.com	1

Timestamp	UID	URL
2020-10-13 14:06	A1234	sportsnews.com/sp
2020-10-13 14:07	A1234	news.com/myaccount/soiso
2020-10-13 14:07	A1234	ecommerce.com/receipt_id=1234054
2020-10-13 14:08	A1234	sportsnews.com/sports
2020-10-13 14:06	B5678	sportsnews.com/sports
2020-10-13 14:07	B5678	sportsnews.com/sports

Basis: `https://0x65.dev/blog` • Human Web—Collecting Data in a Socially Responsible Manner (2019)

User Data

Risks: Record Linkage

Domain	Unique Visitors
sportsnews.com	2
news.com	1
ecommerce.com	1

Timestamp	UID	URL
2020-10-13 14:06	A1234	sportsnews.com/sp
2020-10-13 14:07	A1234	news.com/myaccount/soiso
2020-10-13 14:07	A1234	ecommerce.com/receipt_id=1234054
2020-10-13 14:08	A1234	sportsnews.com/sports
2020-10-13 14:06	B5678	sportsnews.com/sports
2020-10-13 14:07	B5678	sportsn

URL	UID
sportsnews.com/sports	A1234
news.com/myaccount/alice	A1234
ecommerce.com/receipt_id=1234054	A1234

Basis: `https://0x65.dev/blog` • Human Web—Collecting Data in a Socially Responsible Manner (2019)

User Data

Risks: Record Linkage

Timestamp	UID	URL
2020-10-13	A1234	sportsnews.com/sp
2020-10-13 14:07	A1234	news.com/myaccount/alice
2020-10-13 14:07	A1234	ecommerce.com/receipt_id=1234054
2020-10-13 14:08	A1234	sportsnews.com/sports
2020-10-13 14:06	B5678	sportsnews.com/sports
2020-10-13 14:07	B5678	sportsn

Domain	Unique Visitors
sportsnews.com	2
news.com	1
ecommerce.com	1

URL	UID
sportsnews.com/sports	A1234
news.com/myaccount/alice	A1234
ecommerce.com/receipt_id=1234054	A1234

Basis: `https://0x65.dev/blog` • Human Web—Collecting Data in a Socially Responsible Manner (2019)

User Data

Risks: Record Linkage

Tim st m	UI	URL
20 0-1		port s.com/sp
2020-10-1		ount/alice
2020-10-		receipt_id=1234054
2020-10-1 14:0	1234	sportsnews.com/sports
2020-10-13 14:06	B5678	sportsnews.com/sports
2020-10-13 14:07	B5678	sportsn

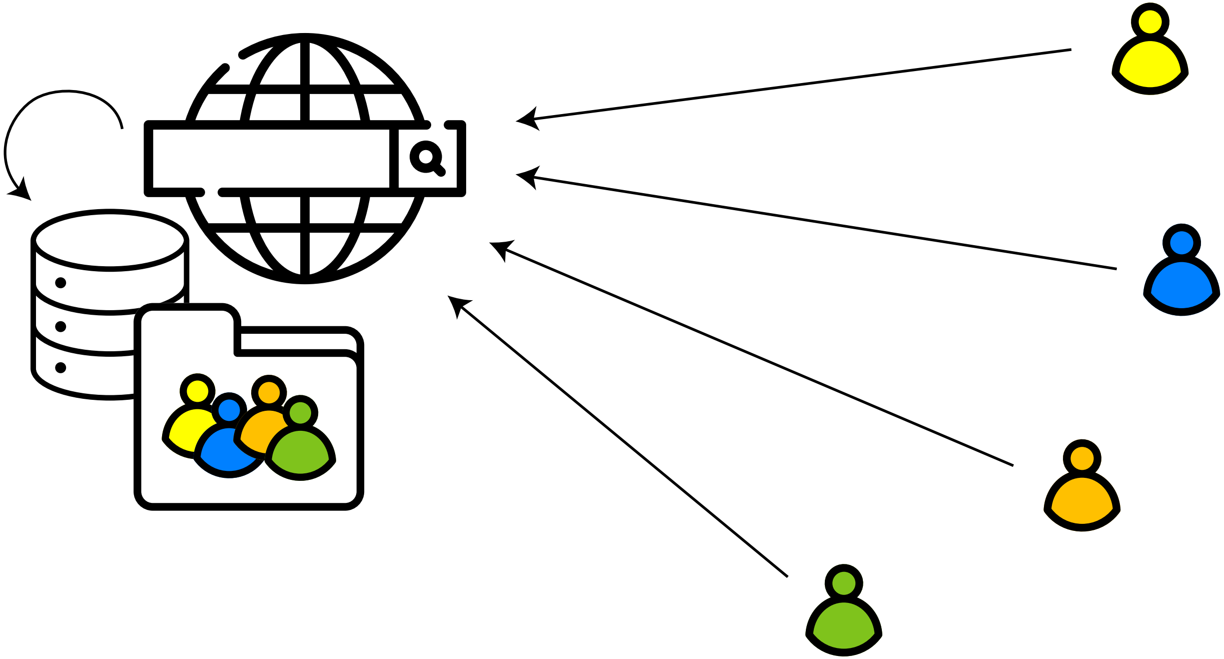
Domain	Unique Visitors
sportsnews.com	2
news.com	1
ecommerce.com	1

URL	UID
sportsnews.com/sports	A1234
news.com/myaccount/alice	A1234
ecommerce.com/receipt_id=1234054	A1234

Basis: `https://0x65.dev/blog` • Human Web—Collecting Data in a Socially Responsible Manner (2019)

User Data

Traditional Data Collection



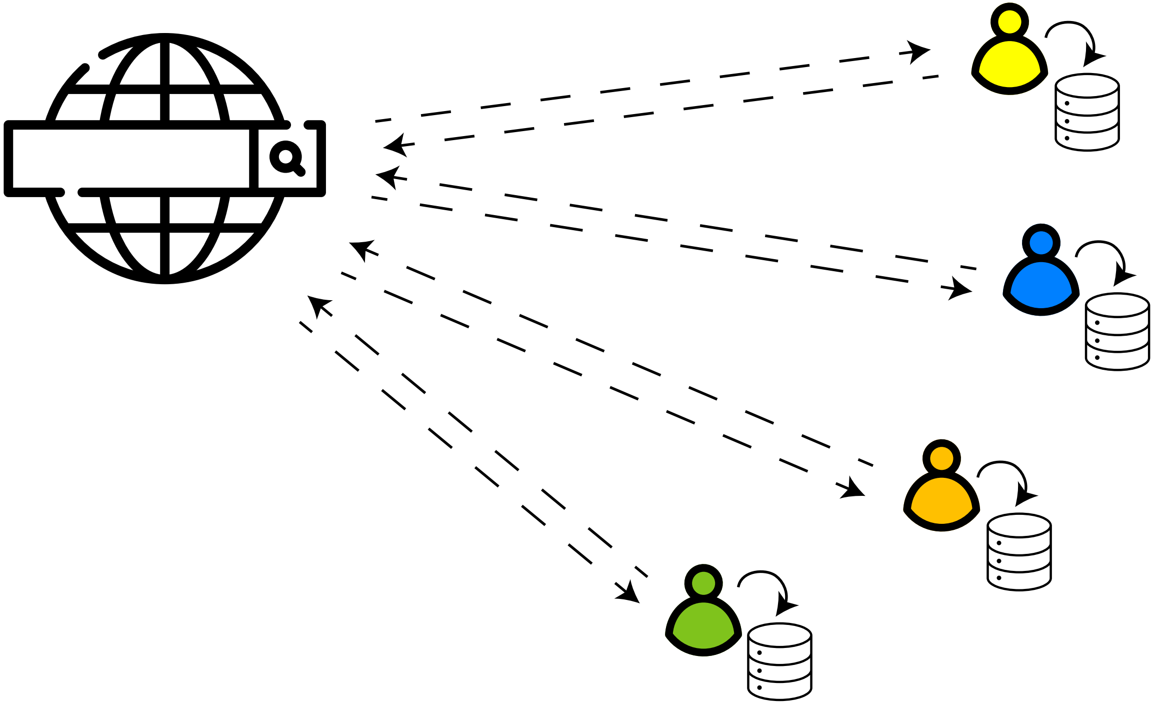
Basis: `https://0x65.dev/blog` • Human Web—Collecting Data in a Socially Responsible Manner (2019)

User Data

Privacy-preserving Data Collection

CLIQZ [2015 – 2020]

Human Web & Human Proxy Network



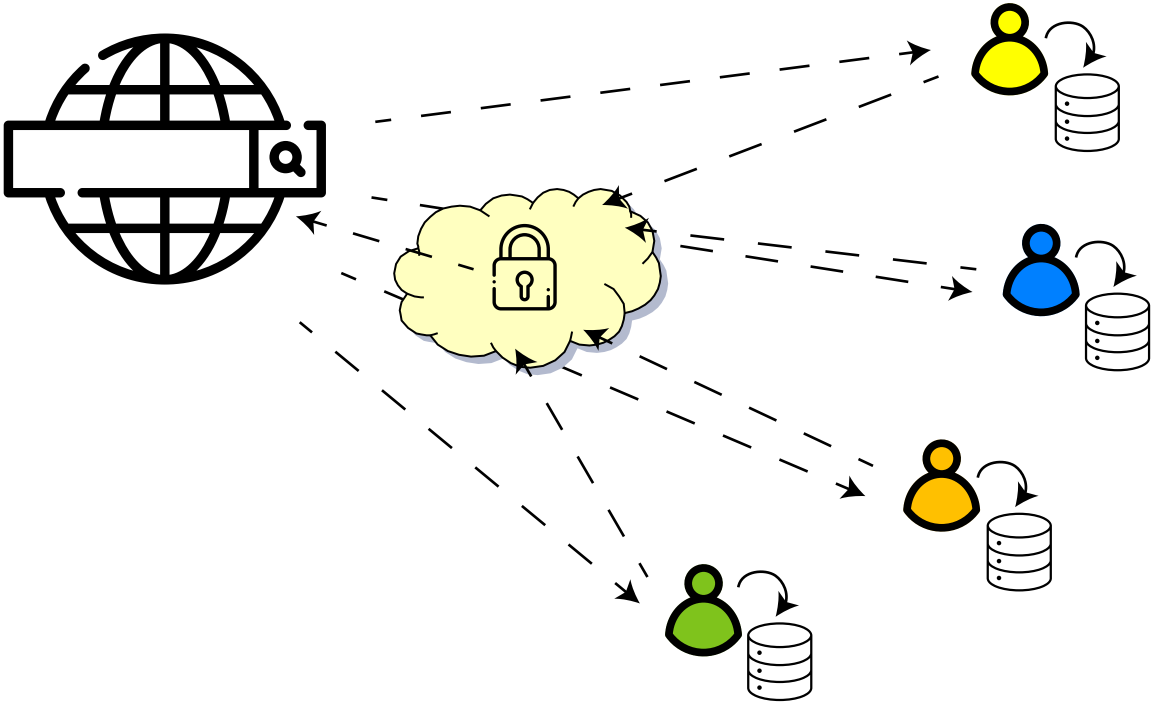
Basis: `https://0x65.dev/blog` • Human Web—Collecting Data in a Socially Responsible Manner (2019)

User Data

Privacy-preserving Data Collection

CLIQZ [2015 – 2020]

Human Web & Human Proxy Network



Basis: `https://0x65.dev/blog` • Human Web—Collecting Data in a Socially Responsible Manner (2019)

◉

Summary

Independence

Scale

User data

Market penetration

Funding

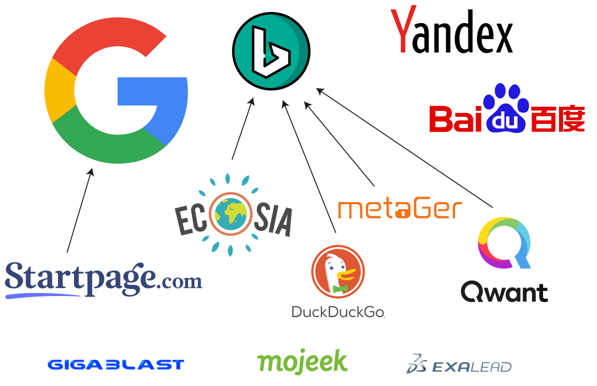
Transparency

Challenges

Summary

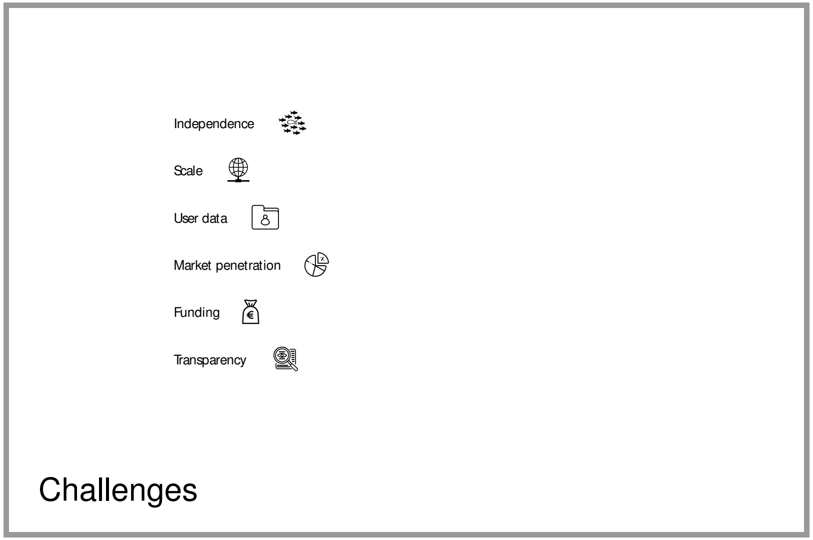
- Independence
- Scale
- User data
- Market penetration
- Funding
- Transparency

Challenges



Independence

Summary



Independence

Scale

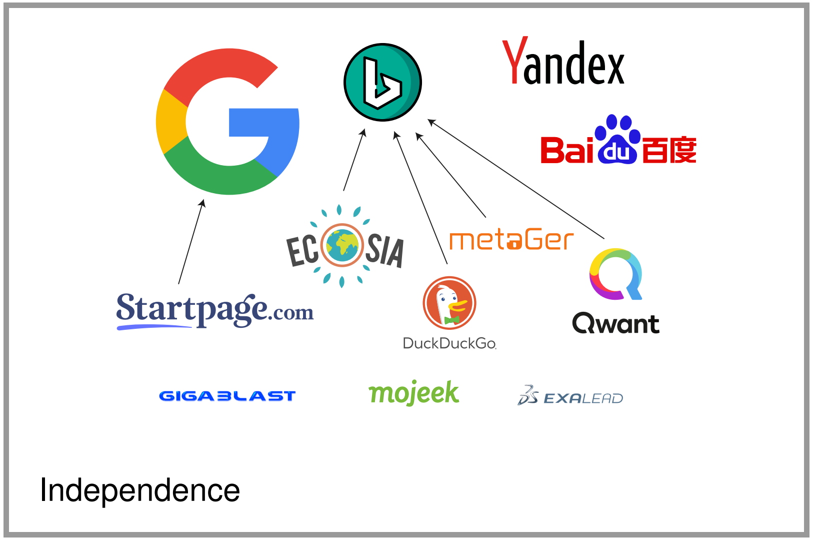
User data

Market penetration

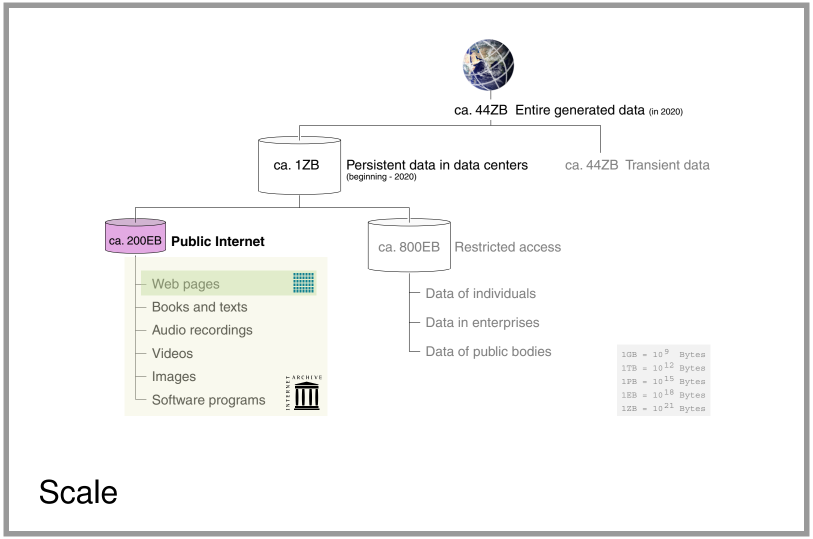
Funding

Transparency

Challenges



Independence



ca. 44ZB Entire generated data (in 2020)

ca. 1ZB Persistent data in data centers (beginning - 2020)

ca. 44ZB Transient data

ca. 200EB **Public Internet**

- Web pages
- Books and texts
- Audio recordings
- Videos
- Images
- Software programs

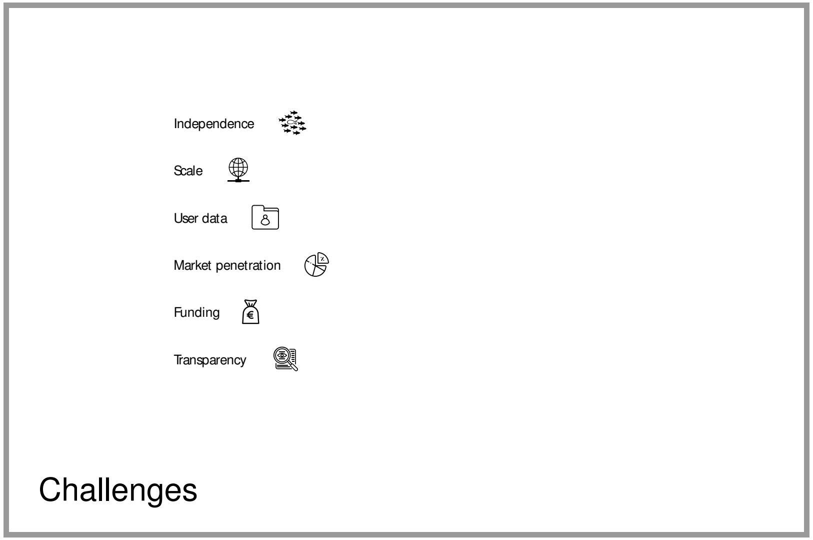
ca. 800EB Restricted access

- Data of individuals
- Data in enterprises
- Data of public bodies

1GB = 10^{9} Bytes
1TB = 10^{12} Bytes
1PB = 10^{15} Bytes
1EB = 10^{18} Bytes
1ZB = 10^{21} Bytes

Scale

Summary



Independence

Scale

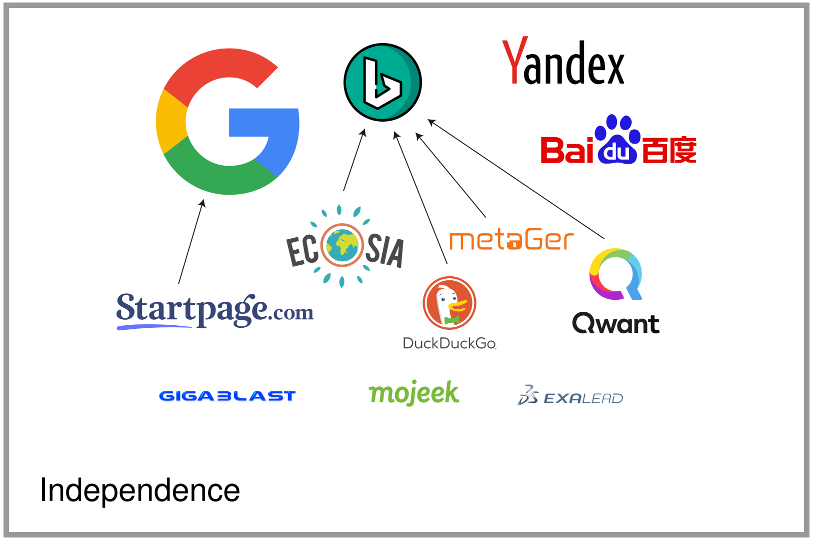
User data

Market penetration

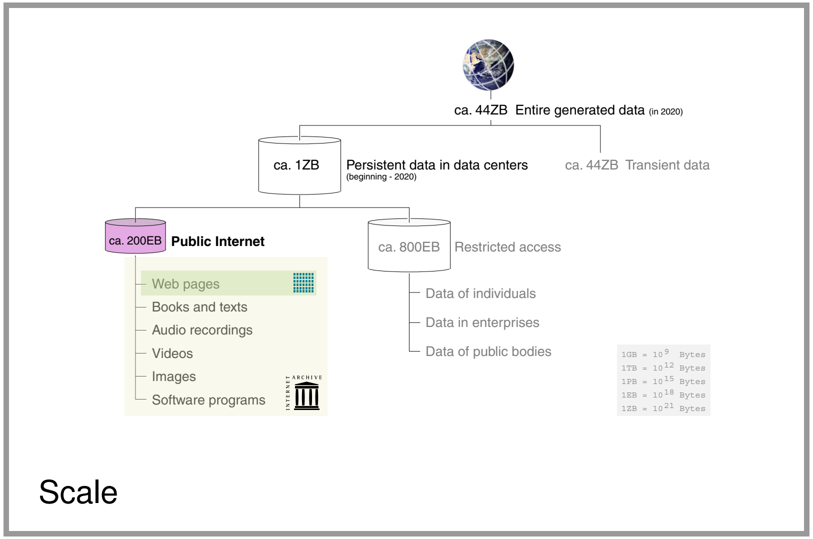
Funding

Transparency

Challenges



Independence



ca. 44ZB Entire generated data (in 2020)

ca. 1ZB Persistent data in data centers (beginning - 2020)

ca. 44ZB Transient data

ca. 200EB **Public Internet**

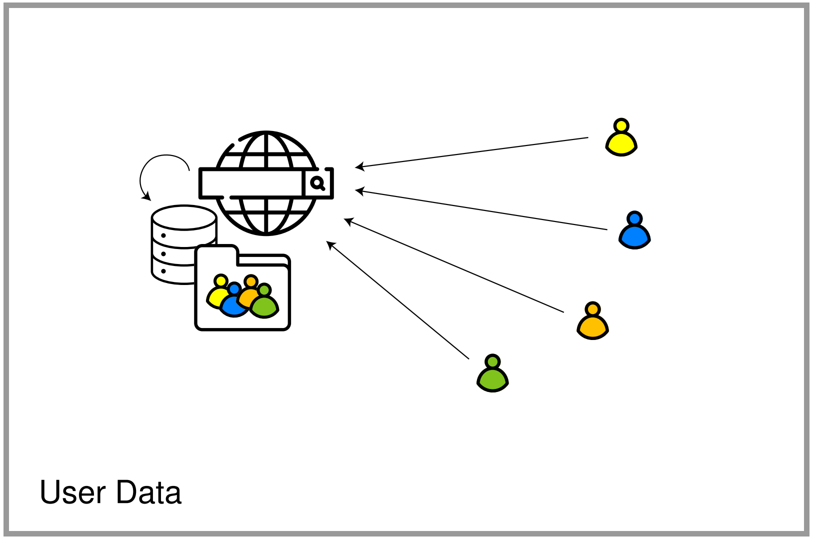
- Web pages
- Books and texts
- Audio recordings
- Videos
- Images
- Software programs

ca. 800EB Restricted access

- Data of individuals
- Data in enterprises
- Data of public bodies

```
1GB = 10^9  Bytes
1TB = 10^12 Bytes
1PB = 10^15 Bytes
1EB = 10^18 Bytes
1ZB = 10^21 Bytes
```

Scale



User Data

Thank You!

Graphics re-used in this presentation are © their original authors • Icons by Freepik, iconixar, Wichai.wi, DinosoftLabs, Kiranshastry @ Flaticon.com and Nick Dominguez, amy morgan @ theNounProject.com