IR and NLP Research in the Webis Group

—Overview and Background—

Benno Stein
Bauhaus-Universität Weimar
webis.de

Data Science Colloquium · Paderborn · December 16, 2022

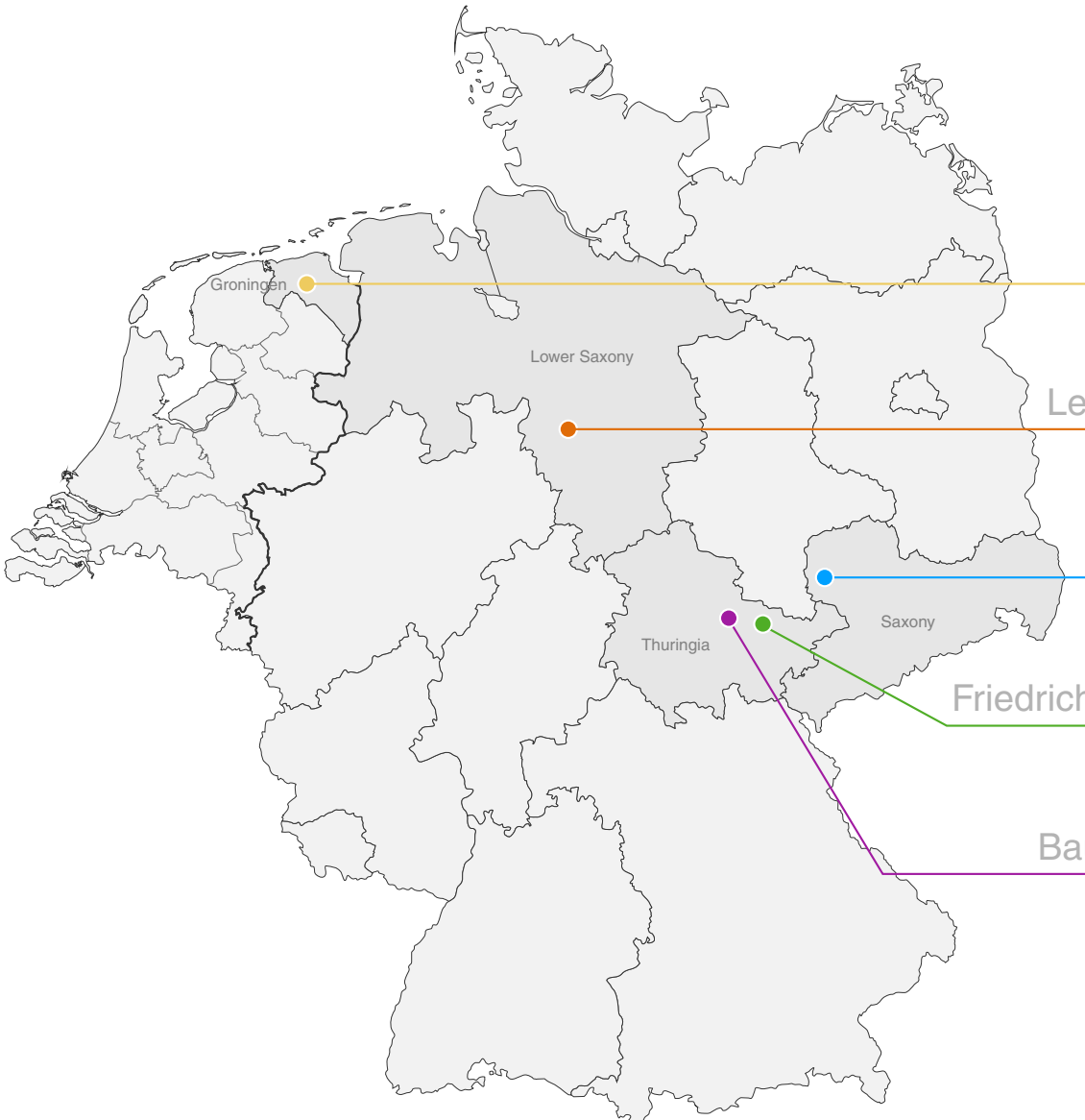
Outline

① About us (Webis)

② Archive Data

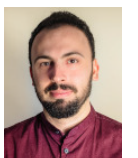
③ Data Analytics @ Webis

Webis Group



University of Groningen

Prof. Dr. Khalid Al-Khatib



Leibniz University Hannover

Prof. Dr. Henning Wachsmuth

Leipzig University

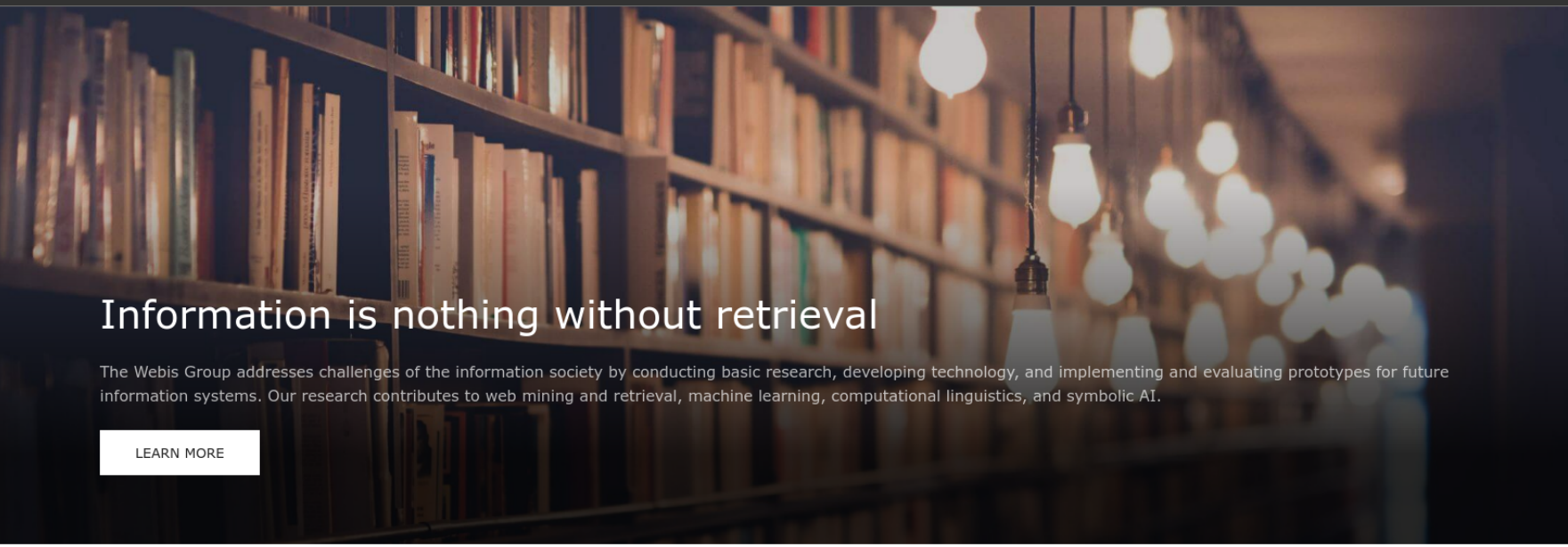
Prof. Dr. Martin Potthast

Friedrich-Schiller-Universität Jena

Prof. Dr. Matthias Hagen

Bauhaus-Universität Weimar

Prof. Dr. Benno Stein



Information is nothing without retrieval

The Webis Group addresses challenges of the information society by conducting basic research, developing technology, and implementing and evaluating prototypes for future information systems. Our research contributes to web mining and retrieval, machine learning, computational linguistics, and symbolic AI.

LEARN MORE

Search Engines

- Args
  Argument search
- ChatNoir
  Web search
- Netspeak
  Writing assistance
- Picapica
  Plagiarism detection

GRONINGEN	HANNOVER	JENA	LEIPZIG	WEIMAR
Home	Home	Home	Home	Home
People	People	People	People	People
Teaching	Teaching	Teaching	Teaching	Teaching
Research	Research	Research	Research	Research

Archive Data: The Global Datasphere

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 in 2020

ca. 59ZB Entire data generated in 2020

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

ca. 59ZB Transient data

ca. 200EB — Public access

- **Web pages (< 1EB)**
- 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

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

The Global Datasphere in 2020

ca. 59ZB Entire data generated in 2020

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

ca. 59ZB Transient data

ca. 200EB — Public access

- Web pages (< 1EB)
- Books and texts
- Audio recordings
- Videos
- Images
- Software programs

ca. 800EB — Restricted access

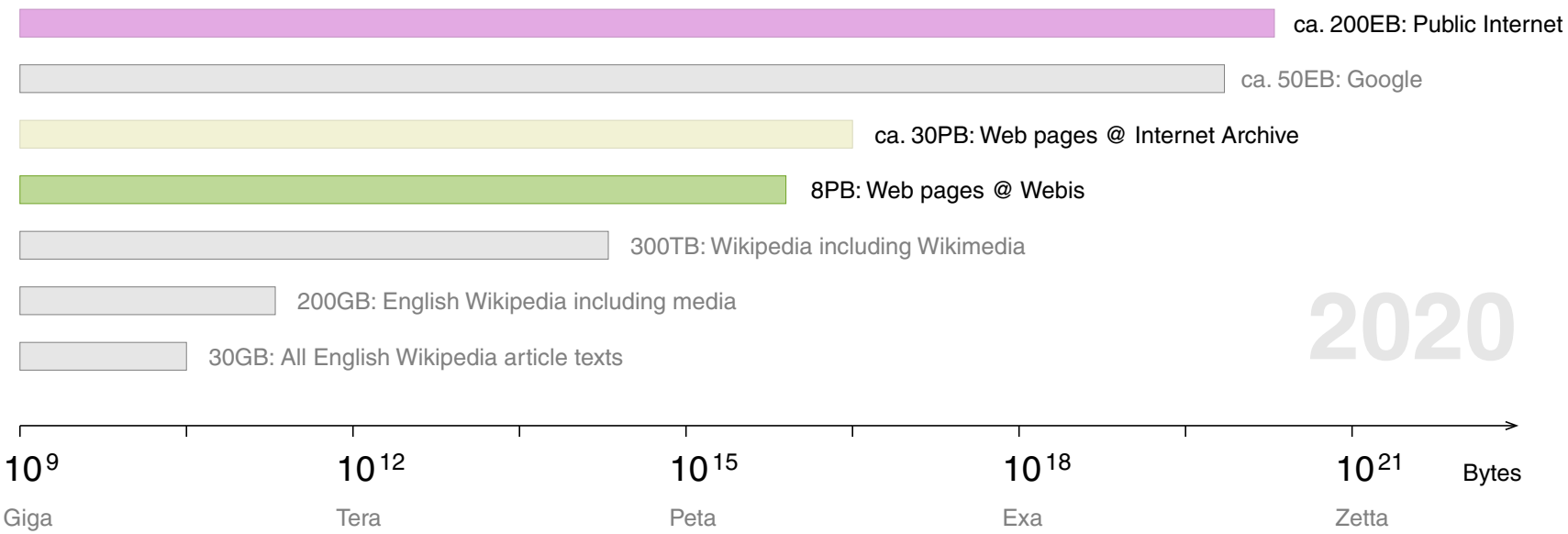
- Data of individuals
- Data in enterprises
- Data of public bodies

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

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

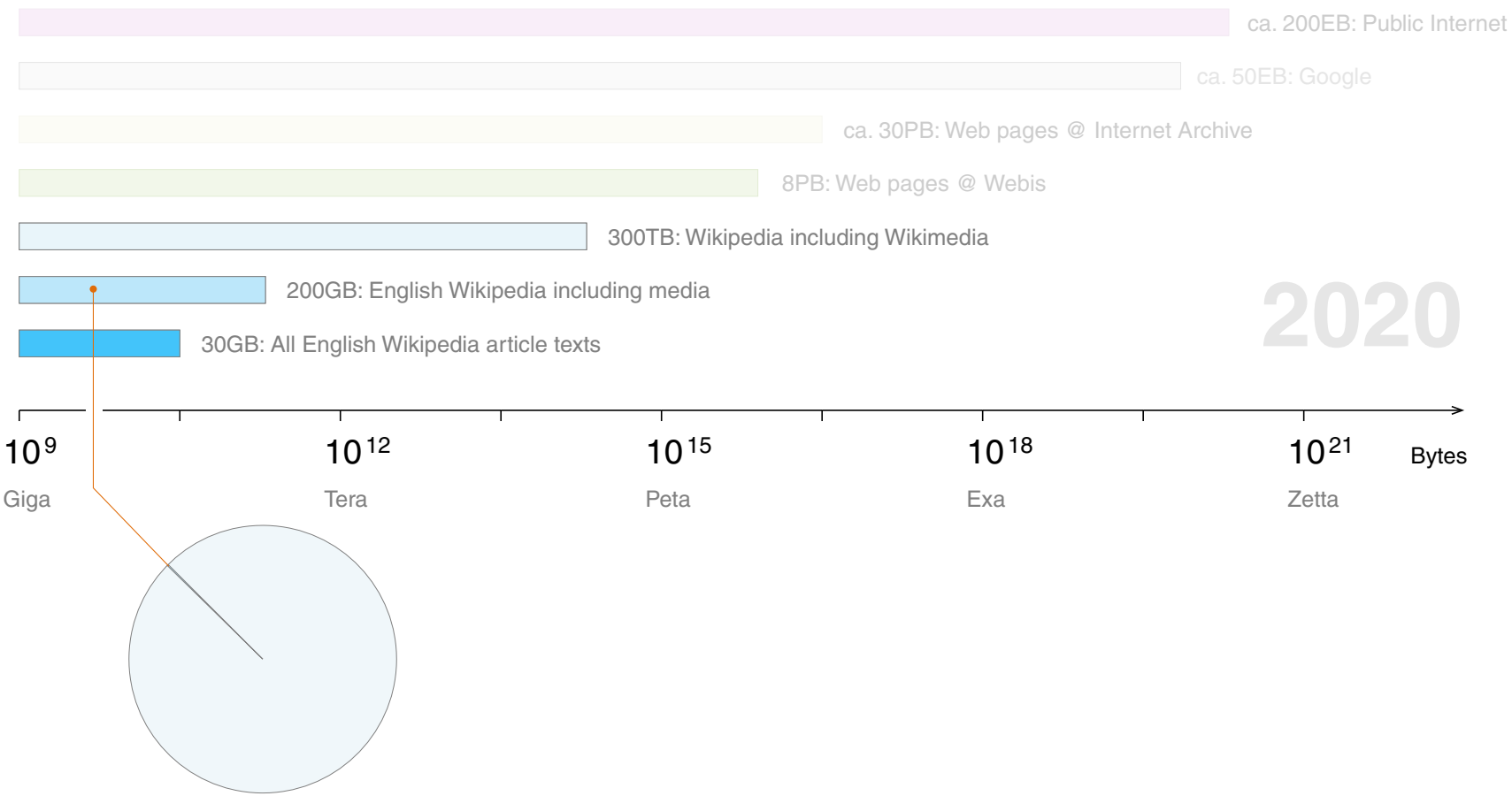
The Global Datasphere in 2020

Relating Data Source Sizes [Vöslke et al. 2021]



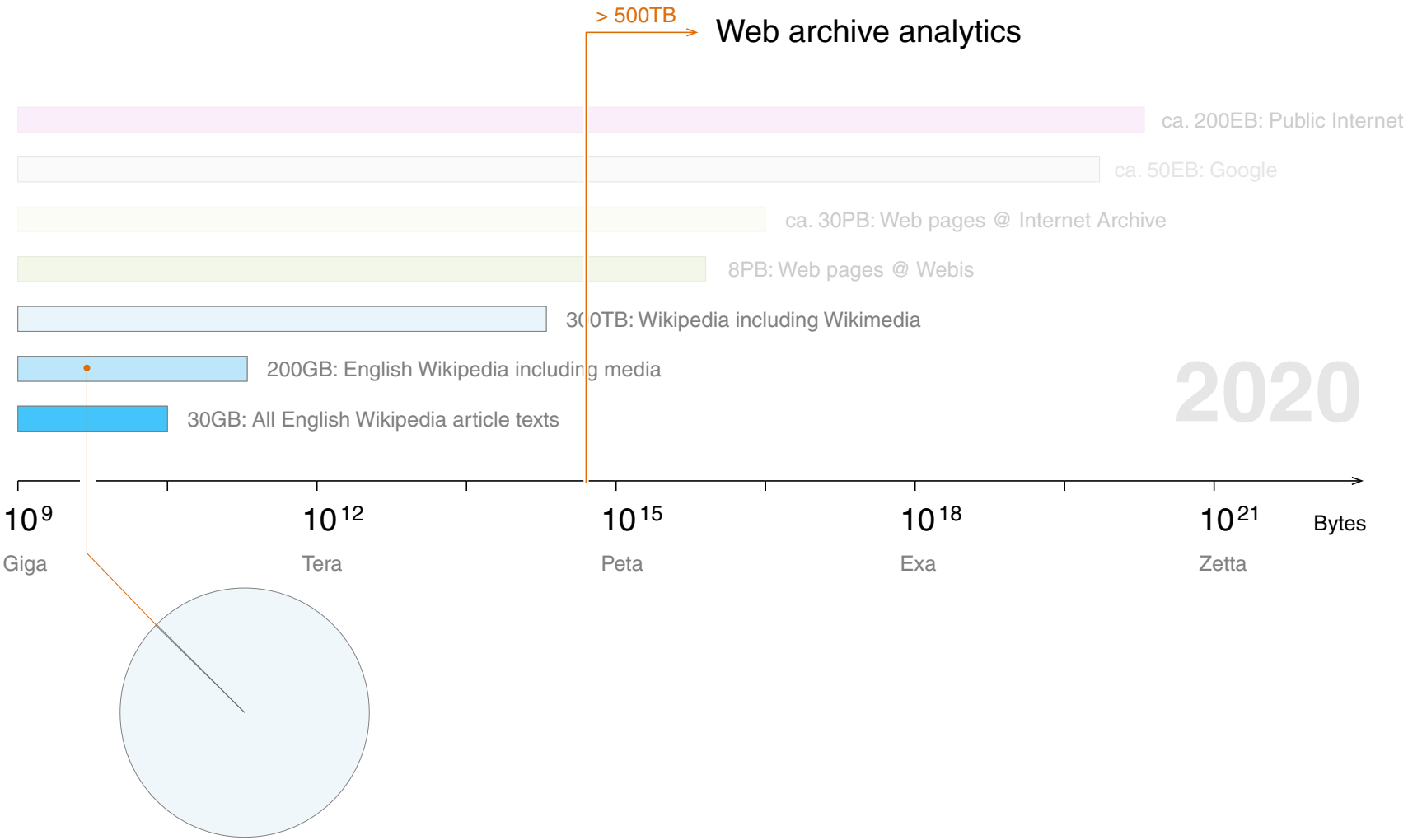
The Global Datasphere in 2020

Relating Data Source Sizes [Vöslke et al. 2021]



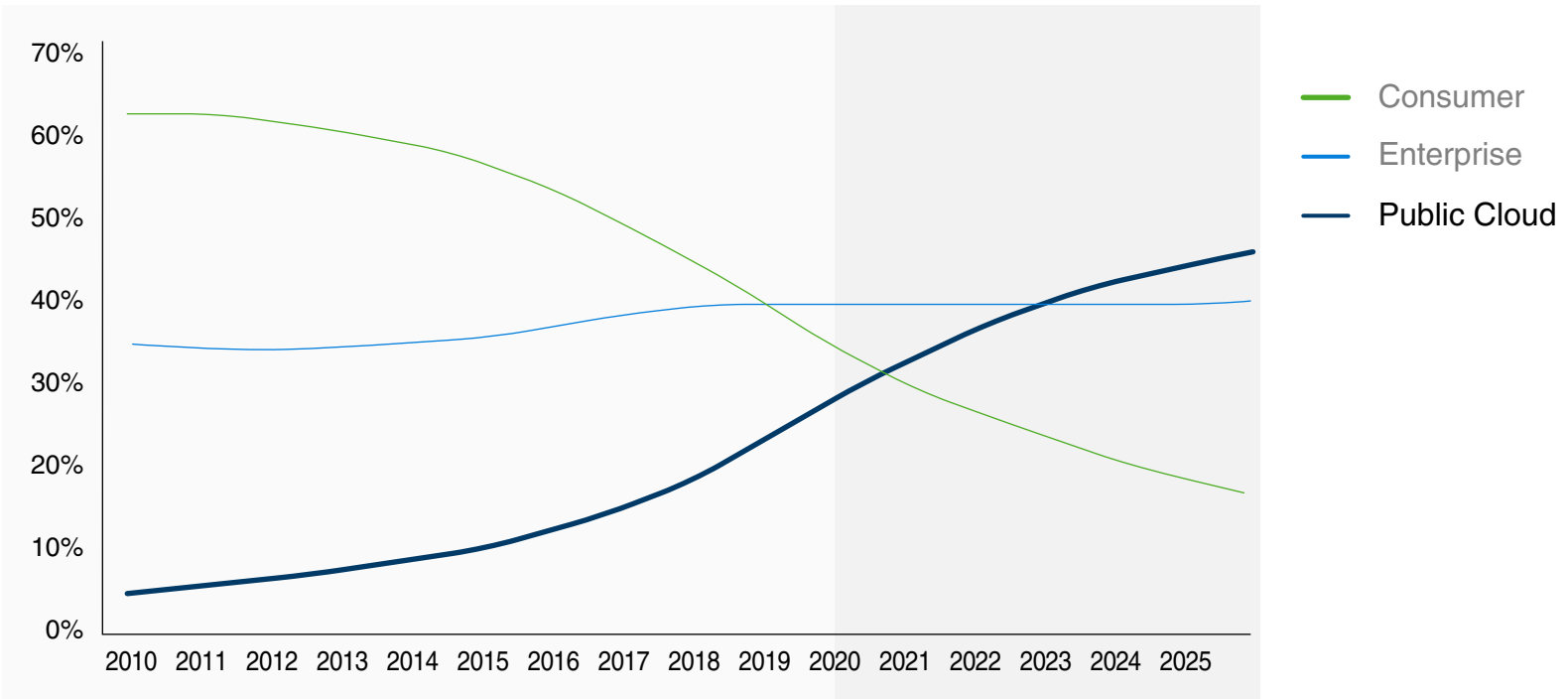
The Global Datasphere in 2020

Relating Data Source Sizes [Vöslke et al. 2021]



The Global Datasphere in 2020

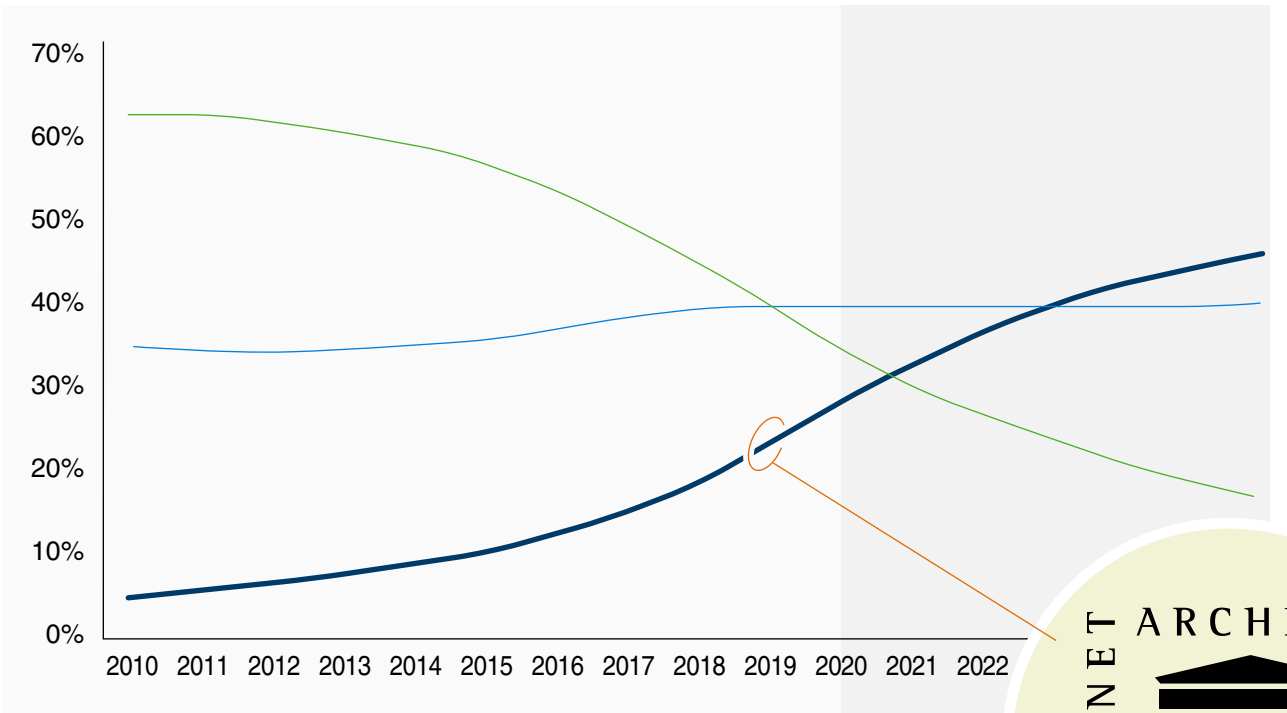
Where is the Data Stored?



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

The Global Datasphere in 2020

Where is the Data Stored?



Among others:

INTERNET ARCHIVE

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

Archive Data: The Internet Archive

INTERNET ARCHIVE

- Founded 1996 by Brewster Kahle
- For all things digital:
  - 477 billion web pages (ca. 30PB) – accessible via the INTERNET ARCHIVE WayBackMachine
  - 20 million books and texts
  - 4.5 million audio recordings (including 180,000 live concerts)
  - 4 million videos (including 1.6 million Television News programs)
  - 3 million images
  - 200,000 software programs

INTERNET ARCHIVE

- Founded 1996 by Brewster Kahle
- For all things digital:
  - 477 billion web pages (ca. 30PB) – accessible via the INTERNET ARCHIVE WayBackMachine
  - 20 million books and texts
  - 4.5 million audio recordings (including 180,000 live concerts)
  - 4 million videos (including 1.6 million Television News programs)
  - 3 million images
  - 200,000 software programs

INTERNET ARCHIVE

Mission: “Universal access to all knowlege.”

- One full copy in San Francisco
- Part at the new Library of Alexandria
- Part in Amsterdam
- Copy representative portion (8PB) to the Digital Bauhaus Lab / Webis group:

[archive.webis.de]

Data Analytics @ Webis: Platform and Stacks

Webis Data Center (Digital Bauhaus Lab)

Webis Data Center (Digital Bauhaus Lab)

	α-web [2009]	β-web [2015]	γ-web [2016 + 2021]	δ-web [2018]	ε-web [2020]
Nodes	44	135	9	78	55
Disk [PB]	0.2	4.1	0.08	12	0.1
Cores	176 ≅ 3.2 TFLOPs	1,740 ≅ 67.4 TFLOPs	672 + 227,328 ≅ 8 PFLOPs	1,248 ≅ 119.8 TFLOPs	1,100 ≅ 44 TFLOPs
RAM [TB]	0.8	28	7.5	10	7

Typical research:

- α-Web. Teaching, Staging environment
- β-Web. Web mining (map reduce), Authorship analytics, Virtualization (compute, web services)
- γ-Web. Machine learning (embedding, deep learning), Text synthesis, Language modeling
- δ-Web. Web archiving, Virtualization (storage)
- ϵ-Web. Search index construction, Argument search

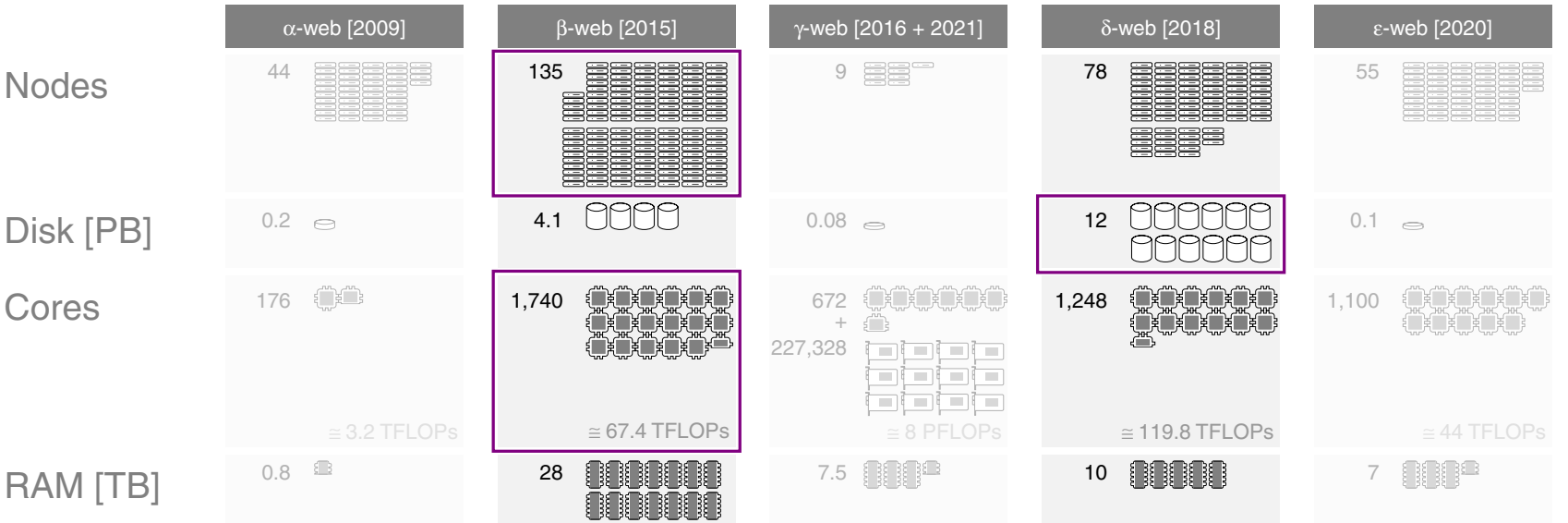
Webis Data Center (Digital Bauhaus Lab)

	α-web [2009]	β-web [2015]	γ-web [2016 + 2021]	δ-web [2018]	ε-web [2020]
Nodes	44	135	9	78	55
Disk [PB]	0.2	4.1	0.08	12	0.1
Cores	176	1,740	672 + 227,328	1,248	1,100
	≅ 3.2 TFLOPs	≅ 67.4 TFLOPs	≅ 8 PFLOPs	≅ 119.8 TFLOPs	≅ 44 TFLOPs
RAM [TB]	0.8	28	7.5	10	7

Typical research:

- α-Web. Teaching, Staging environment
- β-Web. Web mining (map reduce), Authorship analytics, Virtualization (compute, web services)
- γ-Web. Machine learning (embedding, deep learning), Text synthesis, Language modeling
- δ-Web. Web archiving, Virtualization (storage)
- ϵ-Web. Search index construction, Argument search

Webis Data Center (Digital Bauhaus Lab)



Typical research:

- α-Web. Teaching, Staging environment
- β-Web. Web mining (map reduce), Authorship analytics, Virtualization (compute, web services)
- γ-Web. Machine learning (embedding, deep learning), Text synthesis, Language modeling
- δ-Web. Web archiving, Virtualization (storage)
- ϵ-Web. Search index construction, Argument search

Webis Data Center (Digital Bauhaus Lab)

	α-web [2009]	β-web [2015]	γ-web [2016 + 2021]	δ-web [2018]	ε-web [2020]
Nodes	44	135	9	78	55
Disk [PB]	0.2	4.1	0.08	12	0.1
Cores	176 ≅ 3.2 TFLOPs	1,740 ≅ 67.4 TFLOPs	672 + 227,328 ≅ 8 PFLOPs	1,248 ≅ 119.8 TFLOPs	1,100 ≅ 44 TFLOPs
RAM [TB]	0.8	28	7.5	10	7

Typical research:

- α-Web. Teaching, Staging environment
- β-Web. Web mining (map reduce), Authorship analytics, Virtualization (compute, web services)
- γ-Web. Machine learning (embedding, deep learning), Text synthesis, Language modeling
- δ-Web. Web archiving, Virtualization (storage)
- ϵ-Web. Search index construction, Argument search

Webis Analytics Stack

Data Consumption Layer

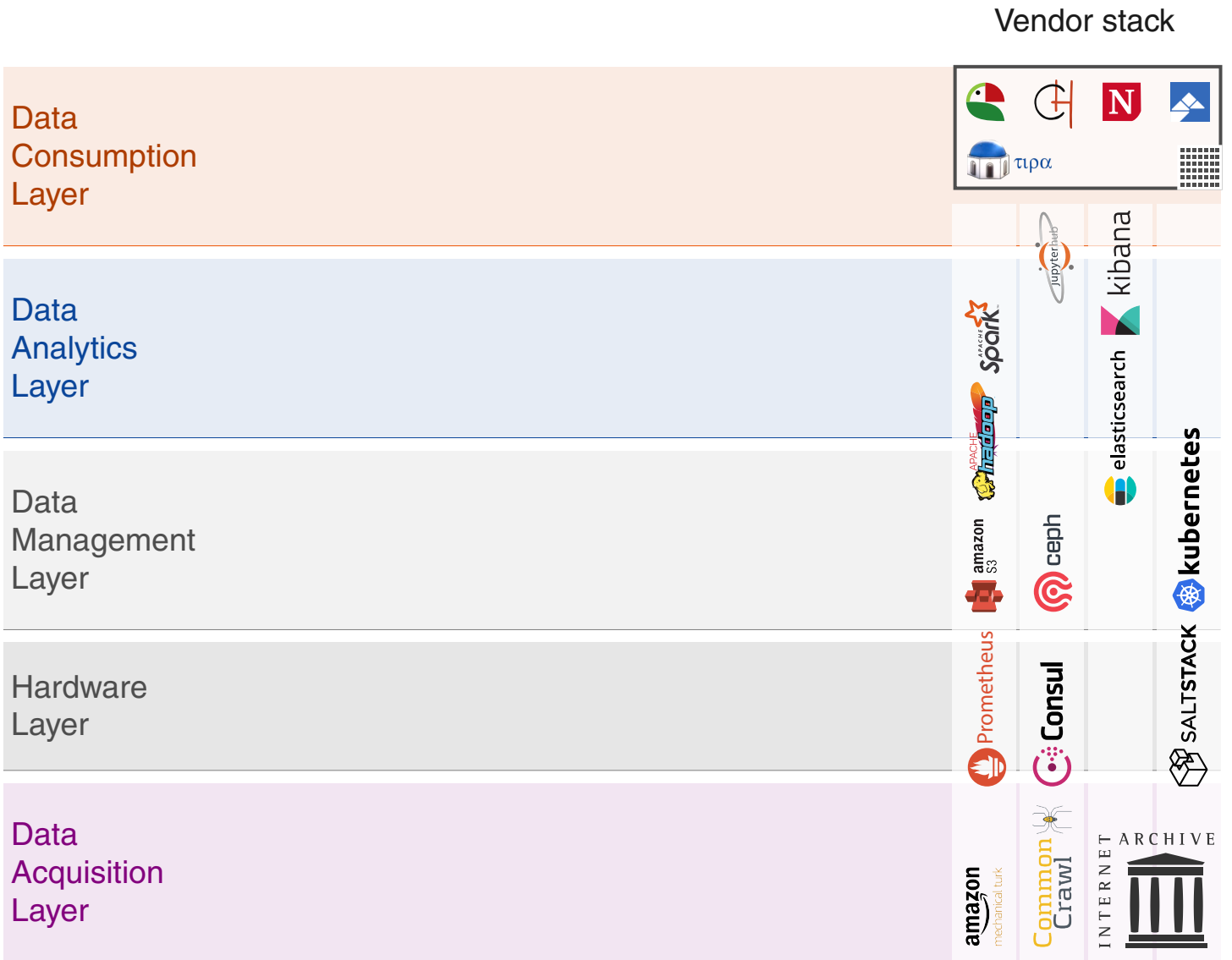
Data Analytics Layer

Data Management Layer

Hardware Layer

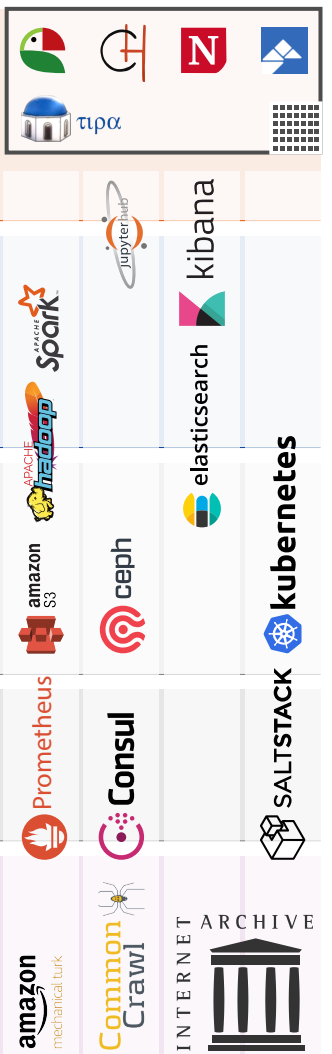
Data Acquisition Layer

Webis Analytics Stack



Webis Analytics Stack

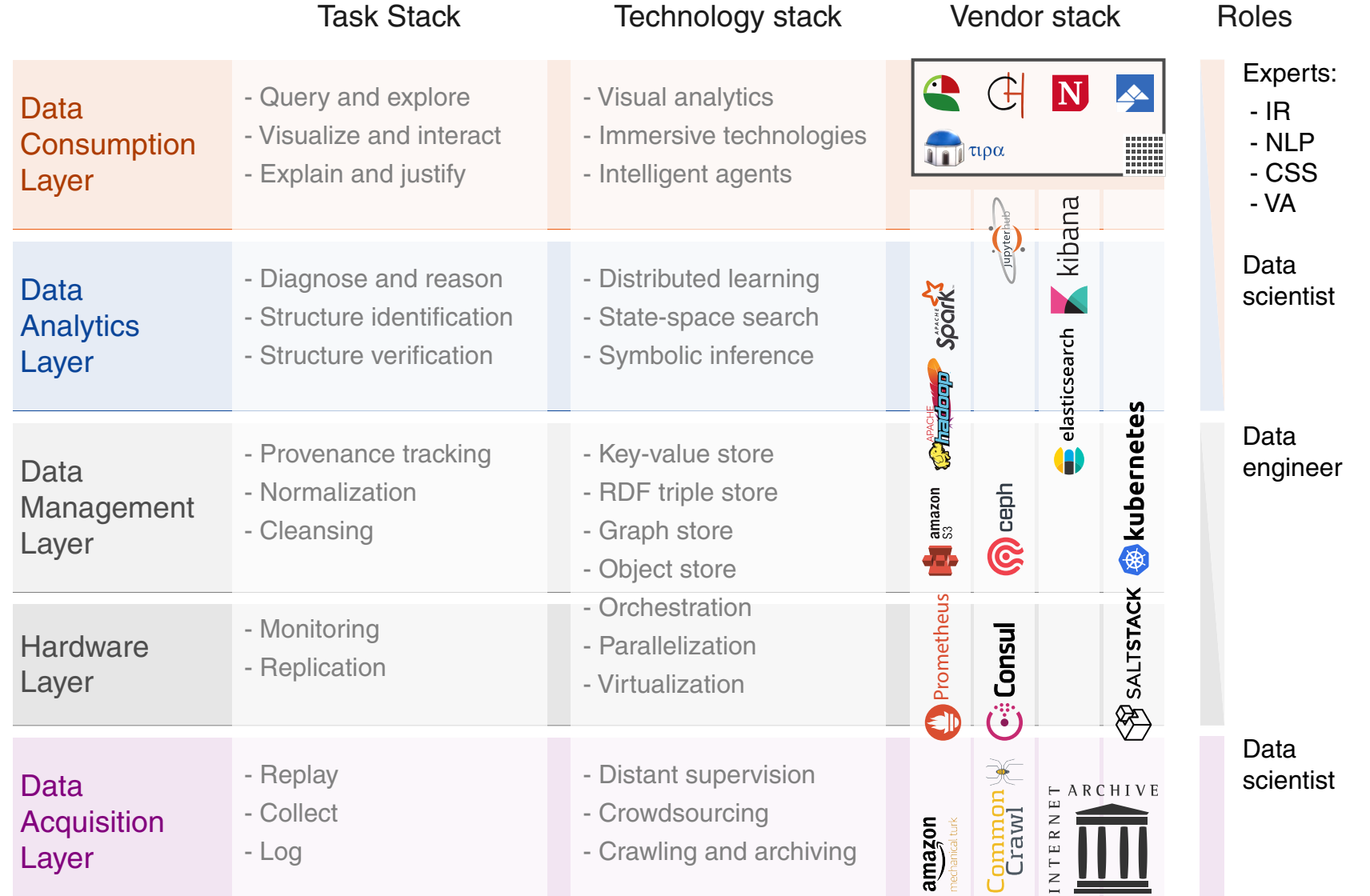
	Technology stack	Vendor stack
Data Consumption Layer	- Visual analytics - Immersive technologies - Intelligent agents	
Data Analytics Layer	- Distributed learning - State-space search - Symbolic inference	
Data Management Layer	- Key-value store - RDF triple store - Graph store - Object store	
Hardware Layer	- Orchestration - Parallelization - Virtualization	
Data Acquisition Layer	- Distant supervision - Crowdsourcing - Crawling and archiving	



Webis Analytics Stack

	Task Stack	Technology stack	Vendor stack
Data Consumption Layer	- Query and explore - Visualize and interact - Explain and justify	- Visual analytics - Immersive technologies - Intelligent agents	
Data Analytics Layer	- Diagnose and reason - Structure identification - Structure verification	- Distributed learning - State-space search - Symbolic inference	
Data Management Layer	- Provenance tracking - Normalization - Cleansing	- Key-value store - RDF triple store - Graph store - Object store	
Hardware Layer	- Monitoring - Replication	- Orchestration - Parallelization - Virtualization	
Data Acquisition Layer	- Replay - Collect - Log	- Distant supervision - Crowdsourcing - Crawling and archiving	

Webis Analytics Stack



Data Analytics @ Webis: Research

Webis Research [publications.webis.de]

Webis Research [publications.webis.de]

args.me

The first (2017) search engine for arguments on the web.

Recent extension: Search for argumentative images.

Background: Argument ranking and search.

ChatNoir

Search engine with rank explanation, indexing the ClueWeb and the CommonCrawl.

Netspeak

Phrase search engine for text correction and idiomatic writing.

Picapica

Search engine for text reuse detection.

- Truths and Myths of the Mnemonic Password Advice

  Approach: Construction of a position-dependent, higher-order language model, based on word initials of two billion sentences of verified casual language.

  Background: The BSI password creation advice.

Example:

"The quick brown fox jumps over the lazy dog."

⇝ Is "**`Tqbfjotld`**" a strong password?

Webis Research [publications.webis.de]

- Detect and Visualize Vandalism in Social Software

  Approach: Spatio-temporal analysis of reverted Wikipedia edits.

  Service: Data analytics and visualization.

  Background: Vandalism in Wikipedia.

- “Celebrity” Profiling

  Goal: Following personal traits on the Internet.

- Hyperpartisan News Detection

  Goal: Analyzing political bias and illustrating provenance on the Internet.

Webis Research [publications.webis.de]

- Learn Discussion Strategies

  Approach: Harvesting talk pages, email repositories, Reddit threads.

- Acquire Justification and Reasoning Knowledge

  Approach: Construction of a causality graph from causal statements.

- Compute Ranking Functions for Arguments

  Approach: Analysis of the hyperlink graph of web pages.

Webis Research [publications.webis.de]

- Who Wrote the Web?

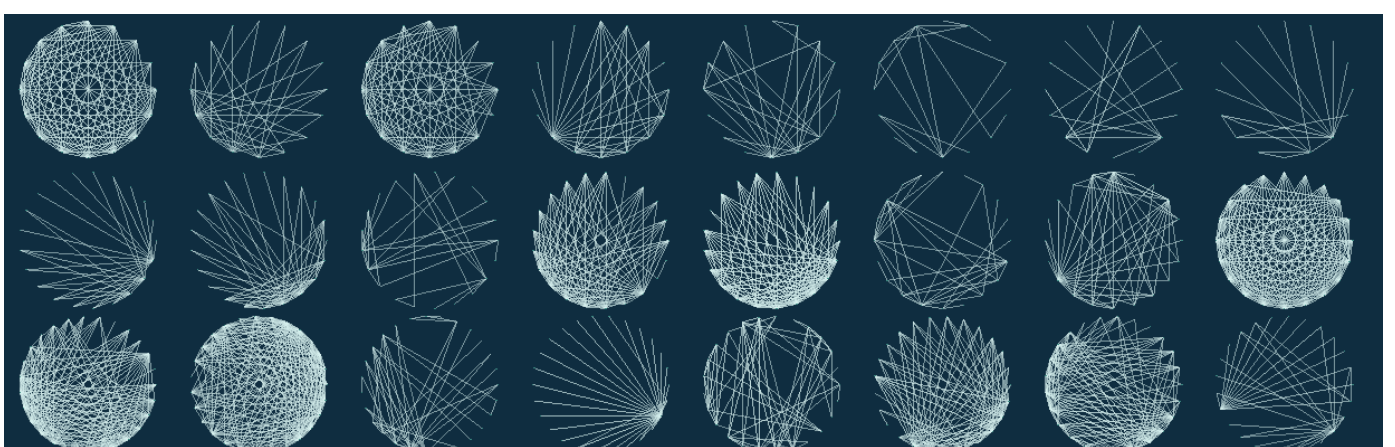
  Applying author identification technology at web-scale.

- Text Reuse Analytics

  Goals: (1) Finding Wikipedia text reuse (on the web).
  (2) Quantifying the prevalence of scientific text reuse.

- Text Reuse Illustration

  Example: Visualizing article similarities in Wikipedia.



Riemann et al.:
Visualizing Article Similarities in Wikipedia.
EuroVis 2016

Webis Research [publications.webis.de]

- Abstractive Snippet Generation

  Approach: Use of anchor contexts to generate abstractive snippets with a pointer-generator network, exploiting ClueWeb09, ClueWeb12, and the DMOZ Open Directory Project.

- Learn Automatic Summarization

  Approach: Exploit author-provided summaries, taking advantage of the common practice of appending a “TL;DR” to long posts.

Webis Research [publications.webis.de]

- Web Page Segmentation

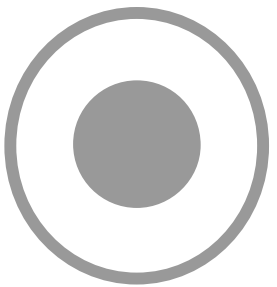
  Goal: Improve reliability of semantic web page segmentation.

- Web Crawling Quality Analysis

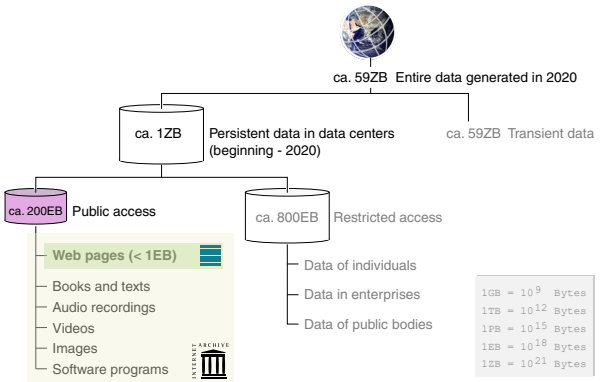
  Goals: (1) Detect incomplete crawls.
  (2) Improve the web page reconstructability from crawls.

- Personal Web Archival

  Goal: Technology for individual web archive creation and search.

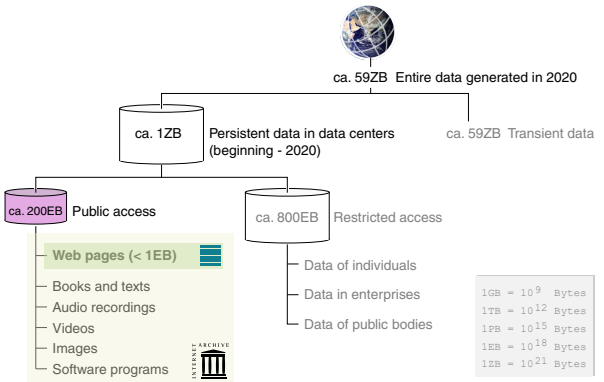


Summary

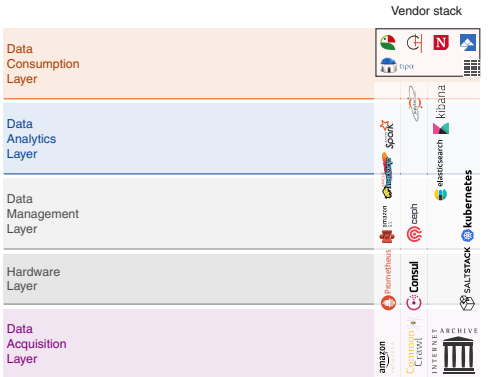


Global Datasphere and the Internet Archive

Summary

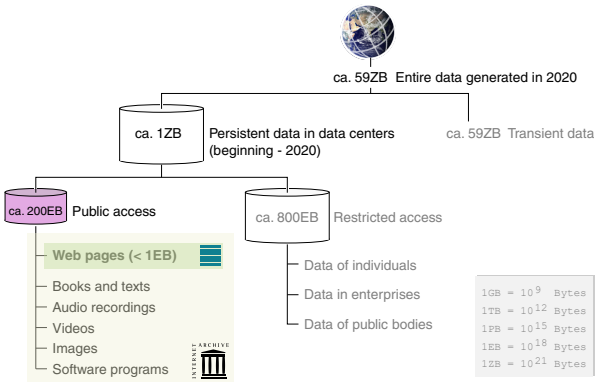


Global Datasphere and the Internet Archive

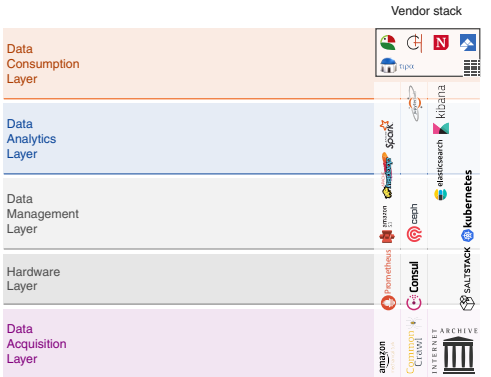


Webis Analytics Stack

Summary



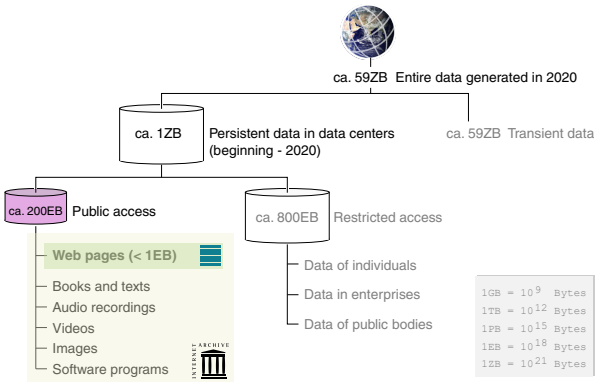
Global Datasphere and the Internet Archive



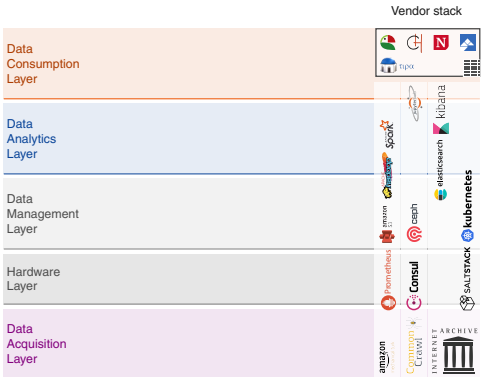
Webis Analytics Stack

Webis Archive Research

Summary



Global Datasphere and the Internet Archive



Webis Analytics Stack

Webis Archive Research

Webis Events

Thank You!