Open Web Search and Web Archive Analytics at Scale

Information Retrieval · Natural Language Processing · Machine Learning

Martin Potthast
Leipzig University and ScaDS.AI
leipzig.webis.de | webis.de

Karlsruhe, June 2023

• Webis Computing Facilities
• Web Archive
• Web Search Engines
• Big Data Analytics Stack
"Artificial Intelligence is the science of making machines do things that would require intelligence if done by men."

[Marvin Minsky 1966]
Artificial Intelligence

Computer Science

Theory
- Theory of computation
- Information and coding theory
- Programming language theory
- Data structures and algorithms

Systems
- Architectures
- Parallel and distributed systems
- Networks
- Security and cryptography
- Databases
- Software Engineering

Applications
- Graphics and visualization
- Human–computer interaction
- Scientific computing and simulation
- Artificial Intelligence

AI [Stuart Russell & Peter Norvig]

Problem Solving
- Search
- Adversarial strategies
- Constraint satisfaction

Knowledge / Reasoning / Planning
- Propositional logic
- First-order logic
- Planning and acting
- Knowledge representation

Dealing with Uncertainty
- Quantifying uncertainty
- Probabilistic reasoning
- Reasoning over time

Learning
- Learning from examples
- Knowledge in learning
- Reinforcement learning

Communicating / Perceiving / Acting
- Natural language processing
- Communication
- Perception
- Robotics

Logics

Linguistics

Mathematics

Statistics

AI [AAAI.org]

Games

Representation

Cognitive science

Machine learning

Natural language

Speech

Vision

Assistive technology

Robotics
Artificial Intelligence

Computer Science

Theory
- Theory of computation
- Information and coding theory
- Programming language theory
- Data structures and algorithms

Systems
- Architectures
- Parallel and distributed systems
- Networks
- Security and cryptography
- Databases
- Software Engineering

Applications
- Graphics and visualization
- Human–computer interaction
- Scientific computing and simulation
- Artificial Intelligence

Logics

Linguistics

Mathematics

Statistics

AI [Stuart Russell & Peter Norvig]

Problem Solving
- Search
- Adversarial strategies
- Constraint satisfaction

Knowledge / Reasoning / Planning
- Propositional logic
- First-order logic
- Planning and acting
- Knowledge representation

Dealing with Uncertainty
- Quantifying uncertainty
- Probabilistic reasoning
- Reasoning over time

Learning
- Learning from examples
- Knowledge in learning
- Reinforcement learning

Communicating / Perceiving / Acting
- Natural language processing
- Communication
- Perception
- Robotics

AI [AAAI.org]

Games

Representation

Reasoning

Cognitive science

Machine learning

Natural language

Speech

Vision

Assistive technology

Robotics

© Potthast 2023
Webis Research Areas

- **Information Retrieval and Search**
  (web search engines, query analytics, argument search, conversational search)

- **Natural Language Processing**
  (argumentation technologies, authorship analytics, computational humanities)

- **Data Mining and Machine Learning**
  (big data analytics, semi-supervised, one-class, domain transfer, LLMs)

- **Shared Tasks and Reproducibility**
  (corpus construction, crowdsourcing, acquisition of compiled expertise)
Webis Computing Facilities
# Webis Computing Facilities

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td></td>
<td>135</td>
<td>9</td>
<td>78</td>
<td>55</td>
</tr>
<tr>
<td>0.8</td>
<td></td>
<td>4.1</td>
<td>0.08</td>
<td>12</td>
<td>0.1</td>
</tr>
<tr>
<td>176</td>
<td></td>
<td>1,740</td>
<td>672 + 227,328</td>
<td>1,248</td>
<td>1,100</td>
</tr>
<tr>
<td></td>
<td>≅ 3.2 TFLOPs</td>
<td>≅ 67.4 TFLOPs</td>
<td>≅ 8 PFLOPs</td>
<td>≅ 119.8 TFLOPs</td>
<td>≅ 44 TFLOPs</td>
</tr>
</tbody>
</table>

© Potthast 2023
### Webis Computing Facilities

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nodes</strong></td>
<td>44</td>
<td>135</td>
<td>9</td>
<td>78</td>
<td>55</td>
</tr>
<tr>
<td><strong>Disk [PB]</strong></td>
<td>0.2</td>
<td>4.1</td>
<td>0.08</td>
<td>12</td>
<td>0.1</td>
</tr>
<tr>
<td><strong>Cores</strong></td>
<td>176</td>
<td>1,740</td>
<td>672 + 227,328</td>
<td>1,248</td>
<td>1,100</td>
</tr>
<tr>
<td><strong>RAM [TB]</strong></td>
<td>0.8</td>
<td>28</td>
<td>7.5</td>
<td>10</td>
<td>7</td>
</tr>
</tbody>
</table>

**Typical research tasks:**

**α-Web.** Teaching, Staging environment

**β-Web.** Virtualization (compute, web services), Web mining (map reduce), Authorship analytics

**γ-Web.** Machine learning (embedding, deep learning), Text synthesis, Language modeling

**δ-Web.** Virtualization (storage), Web archiving

**ε-Web.** Search index construction, Argument search
## Webis Computing Facilities

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nodes</strong></td>
<td>44</td>
<td>135</td>
<td>9</td>
<td>78</td>
<td>55</td>
</tr>
<tr>
<td><strong>RAM [TB]</strong></td>
<td>0.8</td>
<td>28</td>
<td>7.5</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td><strong>Cores</strong></td>
<td>176</td>
<td>1,740</td>
<td>672 + 227,328</td>
<td>1,248</td>
<td>1,100</td>
</tr>
<tr>
<td><strong>Disk [PB]</strong></td>
<td>0.2</td>
<td>4.1</td>
<td>0.08</td>
<td>12</td>
<td>0.1</td>
</tr>
</tbody>
</table>

*Typical research tasks:*

**α-Web.** Teaching, Staging environment

**β-Web.** Virtualization (compute, web services), Web mining (map reduce), Authorship analytics

**γ-Web.** Machine learning (embedding, deep learning), Text synthesis, Language modeling

**δ-Web.** Virtualization (storage), Web archiving

**ε-Web.** Search index construction, Argument search

© Potthast 2023
## Webis Computing Facilities

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nodes</strong></td>
<td>44</td>
<td>135</td>
<td>9</td>
<td>78</td>
<td>55</td>
</tr>
<tr>
<td><strong>Disk [PB]</strong></td>
<td>0.2</td>
<td>4.1</td>
<td>0.08</td>
<td>12</td>
<td>0.1</td>
</tr>
<tr>
<td><strong>Cores</strong></td>
<td>176</td>
<td>1,740</td>
<td>672 + 227,328</td>
<td>1,248</td>
<td>1,100</td>
</tr>
<tr>
<td><strong>RAM [TB]</strong></td>
<td>0.8</td>
<td>28</td>
<td>7.5</td>
<td>10</td>
<td>7</td>
</tr>
</tbody>
</table>

### Typical research tasks:

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td>44</td>
<td>135</td>
<td>9</td>
<td>78</td>
<td>55</td>
</tr>
<tr>
<td>Disk [PB]</td>
<td>0.2</td>
<td>4.1</td>
<td>0.08</td>
<td>12</td>
<td>0.1</td>
</tr>
<tr>
<td>Cores</td>
<td>176</td>
<td>1,740</td>
<td>672 + 227,328</td>
<td>1,248</td>
<td>1,100</td>
</tr>
<tr>
<td>RAM [TB]</td>
<td>0.8</td>
<td>28</td>
<td>7.5</td>
<td>10</td>
<td>7</td>
</tr>
</tbody>
</table>

≡ 3.2 TFLOPs
≡ 67.4 TFLOPs
≡ 8 PFLOPs
≡ 119.8 TFLOPs
≡ 44 TFLOPs

### Typical research tasks:

α-Web. Teaching, Staging environment

β-Web. Virtualization (compute, web services), Web mining (map reduce), Authorship analytics

γ-Web. Machine learning (embedding, deep learning), Text synthesis, Language modeling

δ-Web. Virtualization (storage), Web archiving

ε-Web. Search index construction, Argument search

© Potthast 2023
Web Archive Analytics
Founded 1996 by Brewster Kahle

For all things digital:

- 808 billion web pages (ca. 60PB+) – accessible via the WayBack Machine
- 38 million books and texts
- 15 million audio recordings (including 180,000 live concerts)
- 10 million videos (including 2.5 million Television News programs)
- 5 million images
- 1 million pieces of software (with emulators)
Founded 1996 by Brewster Kahle

For all things digital:

- 808 billion web pages (ca. 60PB+) – accessible via the WayBack Machine
- 38 million books and texts
- 15 million audio recordings (including 180,000 live concerts)
- 10 million videos (including 2.5 million Television News programs)
- 5 million images
- 1 million pieces of software (with emulators)
Mission: “Universal access to all knowledge.”

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

[archive.webis.de]
Web Search Engines

EU project OpenWebSearch
Web Search in a Nutshell

- A vague request.
  Expression of a complex information need: a question

- Billions of documents.
  Text, images, audio files, videos, …

How can a computer's intelligence be tested?
Web Search in a Nutshell

- A vague request.
  Expression of a complex information need: a question, or just a few keywords.

- Billions of documents.
  Text, images, audio files, videos, ... 

- High class imbalance.
  Only a tiny fraction of all documents are relevant to the request.

→ Retrieve relevant documents in milliseconds.

How can a computer's intelligence be tested?

Computer Intelligence Test

Turing test - Wikipedia
Jump to Human intelligence vs. intelligence" (1950) was the first published Turing test (disambiguation) · Graph
Web Search Architecture

Indexing

Storage

Retrieval

© Potthast 2023
Web Search Architecture

Acquisition
- crawling
- plain text conversion
- encoding unification

Text analysis
- index terms
- meta data
- classification

Indexing
- term weighting
- doc. embedding
- collection statistics

Storage
- Document store
- Index

Retrieval
In general, a query is a form of questioning, in a line of inquiry.

Dictionary.com > browse > query
Query | Dictionary.com
to ask or inquire about: No one queried his presence.
Web Search Architecture

**Acquisition**
- crawling
- plain text conversion
- encoding unification

**Text analysis**
- index terms
- meta data
- classification

**Indexing**
- term weighting
- doc. embedding
- collection statistics

**Storage**
- Log
  - queries, clicks, dwell times
- Document store
- Index

**Indexing**

**Query analysis**  
- query
- query meaning
- query definition
- query string

**Query synthesis**

**Snippet generation**

**Query processing**

**Retrieval model**

---

In general, a query is a form of questioning, in a line of inquiry.

dictionary.com › browse › query
Query I Dictionary.com
to ask or inquire about: No one queried his presence.

en.wikipedia.org › wiki › Query
Query - Wikipedia
In general, a query is a form of questioning, in a line of inquiry.
In general, a query is a form of questioning, in a line of inquiry. 

No one queried his presence.
The Archive Query Log (AQL):

- We have uncovered and acquired an extensive query log at Internet Archive.
- It features queries and archived search engine result pages (SERPs).
  SERPs archived actively, or accidentally, e.g., crawled when linked on another page.
- User privacy implications only inasmuch as web data has privacy implications.
The Archive Query Log (AQL):

- We have uncovered and acquired an extensive query log at Internet Archive.
- It features queries and archived search engine result pages (SERPs).
  - SERPs archived actively, or accidentally, e.g., crawled when linked on another page.
- User privacy implications only inasmuch as web data has privacy implications.

<table>
<thead>
<tr>
<th>Search provider</th>
<th>URLs (total)</th>
<th>Queries (total)</th>
<th>Queries (unique)</th>
<th>SERPs (estimate)</th>
<th>Results (estimate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>89.4 M</td>
<td>72.7 M</td>
<td>20.0 M</td>
<td>34.0 M</td>
<td>270.9 M</td>
</tr>
<tr>
<td>YouTube</td>
<td>41.8 M</td>
<td>41.4 M</td>
<td>11.3 M</td>
<td>19.3 M</td>
<td>411.8 M</td>
</tr>
<tr>
<td>Baidu</td>
<td>78.5 M</td>
<td>69.6 M</td>
<td>2.9 M</td>
<td>32.5 M</td>
<td>130.7 M</td>
</tr>
<tr>
<td>Ø 547 others</td>
<td>646.6 M</td>
<td>161.8 M</td>
<td>27.7 M</td>
<td>75.4 M</td>
<td>839.5 M</td>
</tr>
<tr>
<td><strong>∑ 550</strong></td>
<td><strong>1,010.0 M</strong></td>
<td><strong>356.5 M</strong></td>
<td><strong>64.5 M</strong></td>
<td><strong>166.4 M</strong></td>
<td><strong>1,709.0 M</strong></td>
</tr>
</tbody>
</table>

- Use case: Train “Large Ranking Model” for re-ranking.
- Use case: Promote search industry transparency at scale (e.g., fairness).
To test computer intelligence, you can give it problems to solve and evaluate how it performs. You can also give it tasks that require it to think outside the box, such as creating its own algorithms or coming up with innovative solutions to problems. Additionally, you can give it tests that measure its knowledge and understanding of a given subject.

This is outlined in [1], which also covers other definitions and tests of machine intelligence proposed, such as [2] which provides an overview of these definitions and tests. Finally, [3] is a standard test for machine intelligence.

[1] How to Pass the Turing Artificial Intelligence Test | WIRED
How to Pass the Turing Artificial Intelligence Test | WIRED WIRED Staff Science Jun 19, 2012 2:00 PM How to Pass the Turing Artificial Intelligence Test Are you human or a machine? Prove it, by passing the Turing Test. It’s the test of a machine’s ability to exhibit intelligent behavior.
www.shivonzilis.com/

Machine Intelligence - Shivon Zilis ClueWeb22 • Crawled Aug 2022
Teams need to figure out how to test newfound capabilities, and applications need to change so they can create a system of record; they also need to coach employees and learn from the data they enter.
www.geeksforgeeks.org/turing-test-artificial-intelligence/

Turing Test in Artificial Intelligence - GeeksforGeeks ClueWeb22 • Crawled Aug 2022
If a machine engages in a conversation with a human how to process the data it has been demonstrated to be a machine, He has proposed the following skills of the test as follows: The turning judges the conversation is that of two humans.

Show more...
Big Data Analytics Stack
<table>
<thead>
<tr>
<th>Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Consumption Layer</td>
</tr>
<tr>
<td>Data Analytics Layer</td>
</tr>
<tr>
<td>Data Management Layer</td>
</tr>
<tr>
<td>Hardware Layer</td>
</tr>
<tr>
<td>Data Acquisition Layer</td>
</tr>
<tr>
<td>Layer</td>
</tr>
<tr>
<td>-------------------------------</td>
</tr>
<tr>
<td>Data Consumption Layer</td>
</tr>
<tr>
<td>Data Analytics Layer</td>
</tr>
<tr>
<td>Data Management Layer</td>
</tr>
<tr>
<td>Hardware Layer</td>
</tr>
<tr>
<td>Data Acquisition Layer</td>
</tr>
</tbody>
</table>

Vendor stack:
- Prometheus
- kibana
- ceph
- elasticsearch
- SALTSTACK
- Saltstack
- Amazon
- Consul
- Archival internet crawl
<table>
<thead>
<tr>
<th>Layer</th>
<th>Technology stack</th>
<th>Vendor stack</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Consumption Layer</td>
<td>- Visual analytics</td>
<td>Prometheus</td>
</tr>
<tr>
<td></td>
<td>- Immersive technologies</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Intelligent agents</td>
<td></td>
</tr>
<tr>
<td>Data Analytics Layer</td>
<td>- Distributed learning</td>
<td>elastic</td>
</tr>
<tr>
<td></td>
<td>- State-space search</td>
<td>search</td>
</tr>
<tr>
<td></td>
<td>- Symbolic inference</td>
<td>kibana</td>
</tr>
<tr>
<td>Data Management Layer</td>
<td>- Key-value store</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- RDF triple store</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Graph store</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Object store</td>
<td></td>
</tr>
<tr>
<td>Hardware Layer</td>
<td>- Orchestration</td>
<td>Prometheus</td>
</tr>
<tr>
<td></td>
<td>- Parallelization</td>
<td>Consul</td>
</tr>
<tr>
<td></td>
<td>- Virtualization</td>
<td>Saltstack</td>
</tr>
<tr>
<td>Data Acquisition Layer</td>
<td>- Distant supervision</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Crowdsourcing</td>
<td>Amazon</td>
</tr>
<tr>
<td></td>
<td>- Crawling and archiving</td>
<td>Common</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Crawl</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Archive</td>
</tr>
<tr>
<td>Task Stack</td>
<td>Technology stack</td>
<td>Vendor stack</td>
</tr>
<tr>
<td>------------</td>
<td>------------------</td>
<td>-------------</td>
</tr>
<tr>
<td><strong>Data Consumption Layer</strong></td>
<td>- Query and explore - Visualize and interact - Explain and justify</td>
<td>- Visual analytics - Immersive technologies - Intelligent agents</td>
</tr>
<tr>
<td><strong>Data Analytics Layer</strong></td>
<td>- Diagnose and reason - Structure identification - Structure verification</td>
<td>- Distributed learning - State-space search - Symbolic inference</td>
</tr>
<tr>
<td><strong>Data Management Layer</strong></td>
<td>- Provenance tracking - Normalization - Cleansing</td>
<td>- Key-value store - RDF triple store - Graph store - Object store</td>
</tr>
<tr>
<td><strong>Hardware Layer</strong></td>
<td>- Monitoring - Replication</td>
<td>- Orchestration - Parallelization - Virtualization</td>
</tr>
<tr>
<td><strong>Data Acquisition Layer</strong></td>
<td>- Replay - Collect - Log</td>
<td>- Distant supervision - Crowdsourcing - Crawling and archiving</td>
</tr>
</tbody>
</table>

**Technology stack**:
- Visual analytics
- Immersive technologies
- Intelligent agents
- Distributed learning
- State-space search
- Symbolic inference
- Key-value store
- RDF triple store
- Graph store
- Object store
- Orchestration
- Parallelization
- Virtualization
- Distant supervision
- Crowdsourcing
- Crawling and archiving

**Vendor stack**:
- Prometheus
- Kibana
- Elasticsearch
- SaltStack
- Kubernetes
- Consul
- Common Crawl
- Internet Archive
- Amazon Web Services
<table>
<thead>
<tr>
<th>Data Consumption Layer</th>
<th>Data Analytics Layer</th>
<th>Data Management Layer</th>
<th>Hardware Layer</th>
<th>Data Acquisition Layer</th>
<th>Technology stack</th>
<th>Vendor stack</th>
<th>Roles</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Query and explore</td>
<td>- Diagnose and reason</td>
<td>- Provenance tracking</td>
<td>- Monitoring</td>
<td>- Replay</td>
<td>- Visual analytics</td>
<td>Prometheus</td>
<td>Experts:</td>
</tr>
<tr>
<td>- Visualize and interact</td>
<td>- Structure identification</td>
<td>- Normalization</td>
<td>- Replication</td>
<td>- Collect</td>
<td>- Immersive technologies</td>
<td>Consul</td>
<td>- IR</td>
</tr>
<tr>
<td>- Explain and justify</td>
<td>- Structure verification</td>
<td>- Cleansing</td>
<td>-</td>
<td>- Log</td>
<td>- Intelligent agents</td>
<td>elasticsearch</td>
<td>- NLP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>kibana</td>
<td>- CSS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>salthack</td>
<td>- VA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>kubernetes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Technology stack**
- Visual analytics
- Immersive technologies
- Intelligent agents
- Distributed learning
- State-space search
- Symbolic inference
- Key-value store
- RDF triple store
- Graph store
- Object store
- Orchestration
- Parallelization
- Virtualization
- Distant supervision
- Crowdsourcing
- Crawling and archiving

**Vendor stack**
- Prometheus
- Consul
- SALTSTACK
- Kubernetes
- Elasticsearch
- Amazon
- Internet Archive

**Roles**
- Data engineer
- Data scientist
- Experts:
  - IR
  - NLP
  - CSS
  - VA
### Summary

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nodes</strong></td>
<td>44</td>
<td>135</td>
<td>9</td>
<td>78</td>
<td>55</td>
</tr>
<tr>
<td><strong>Disk [PB]</strong></td>
<td>0.2</td>
<td>4.1</td>
<td>0.09</td>
<td>12</td>
<td>0.1</td>
</tr>
<tr>
<td><strong>Cores</strong></td>
<td>176</td>
<td>1,740</td>
<td>672</td>
<td>1,248</td>
<td>52</td>
</tr>
<tr>
<td><strong>RAM [TB]</strong></td>
<td>0.8</td>
<td>28</td>
<td>7.5</td>
<td>10</td>
<td>7</td>
</tr>
</tbody>
</table>

The Webis clusters are our computing foundation.
## Summary

<table>
<thead>
<tr>
<th></th>
<th>Nodes</th>
<th>RAM [TB]</th>
<th>Cores</th>
<th>Disk [PB]</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>α-web [2009]</td>
<td>44</td>
<td>0.8</td>
<td>176</td>
<td>0.2</td>
<td>≅ 3.2 TFLOPs</td>
</tr>
<tr>
<td>β-web [2015]</td>
<td>135</td>
<td>28</td>
<td>1,740</td>
<td>0.09</td>
<td>≅ 67.4 TFLOPs</td>
</tr>
<tr>
<td>γ-web [2016 + 2021]</td>
<td>9</td>
<td>0.09</td>
<td>1,248</td>
<td>0.1</td>
<td>≅ 8 PFLOPs</td>
</tr>
<tr>
<td>δ-web [2018]</td>
<td>78</td>
<td>10</td>
<td>1,100</td>
<td>0.1</td>
<td>≅ 119.8 TFLOPs</td>
</tr>
<tr>
<td>ε-web [2020]</td>
<td>55</td>
<td>7</td>
<td>1,100</td>
<td>0.1</td>
<td>≅ 44 TFLOPs</td>
</tr>
</tbody>
</table>

The Webis clusters are our computing foundation.

The web archive data is our data foundation.
## Summary

<table>
<thead>
<tr>
<th>Nodes</th>
<th>2009</th>
<th>2015</th>
<th>2016-2021</th>
<th>2018</th>
<th>2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>α-web</td>
<td>44</td>
<td>135</td>
<td>9</td>
<td>76</td>
<td>55</td>
</tr>
<tr>
<td>β-web</td>
<td>0.8</td>
<td>28</td>
<td>0.09</td>
<td>12</td>
<td>0.1</td>
</tr>
<tr>
<td>γ-web</td>
<td>1,740</td>
<td>1,248</td>
<td>1,100</td>
<td>1,100</td>
<td>7</td>
</tr>
</tbody>
</table>

### The Webis clusters are our computing foundation.

The web archive data is our data foundation.

### We develop open search engines that scale to the web.
The Webis clusters are our computing foundation.

The web archive data is our data foundation.

We develop open search engines that scale to the web.

An in-house Big Data stack frees research and teaching.
The Webis clusters are our computing foundation.

The web archive data is our data foundation.

We develop open search engines that scale to the web.

An in-house Big Data stack frees research and teaching.