

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

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

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

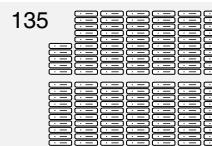
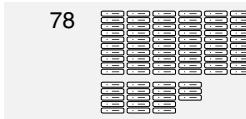
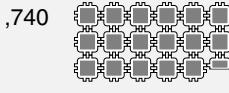
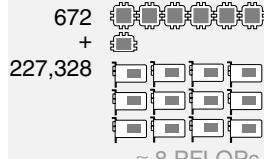
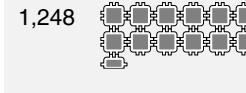
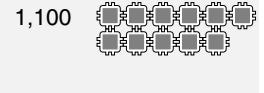
Robotics

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

	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 
	$\approx 3.2 \text{ TFLOPs}$	$\approx 67.4 \text{ TFLOPs}$	$\approx 8 \text{ PFLOPs}$	$\approx 119.8 \text{ TFLOPs}$	$\approx 44 \text{ TFLOPs}$
RAM [TB]	0.8 	28 	7.5 	10 	7 

Webis Computing Facilities

	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
	$\cong 3.2 \text{ TFLOPs}$	$\cong 67.4 \text{ TFLOPs}$	$\cong 8 \text{ PFLOPs}$	$\cong 119.8 \text{ TFLOPs}$	$\cong 44 \text{ TFLOPs}$
RAM [TB]	0.8	28	7.5	10	7

Typical research tasks:

α -Web. Teaching, Staging environment

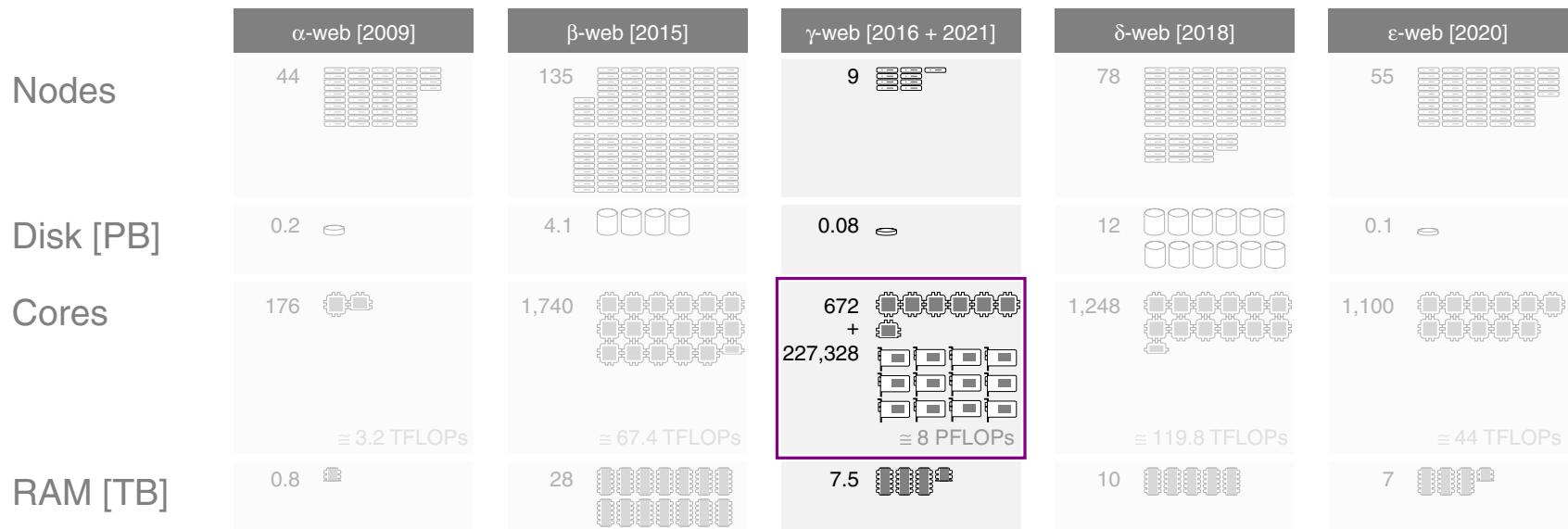
β -Web. Virtualization (compute, web services), Web mining (map reduce), Authorship analytics

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



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

	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
RAM [TB]	0.8	28	7.5	10	7

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

	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 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

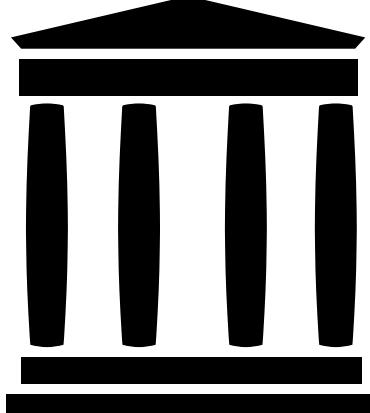
Web Archive Analytics



- Founded 1996 by Brewster Kahle
- For all things digital:
 - 808 billion web pages (ca. 60PB+) – accessible via the [WayBackMachine](#)
 - 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)

INTERNET ARCHIVE
[WayBackMachine](#)

INTERNET ARCHIVE



- 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)

INTERNET ARCHIVE
WayBack Machine



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

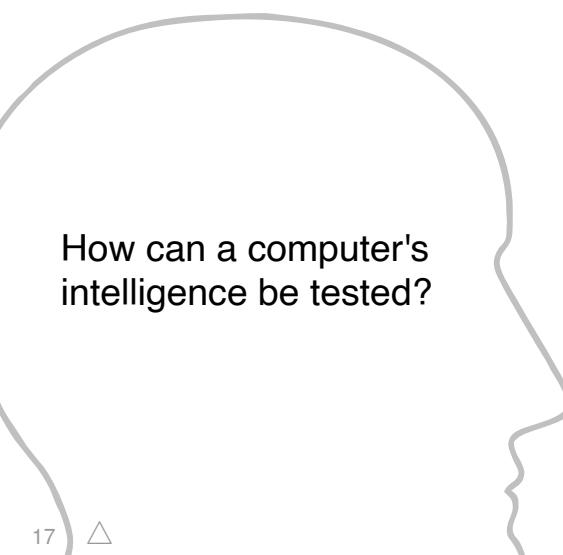
Web Search in a Nutshell

- ❑ 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.

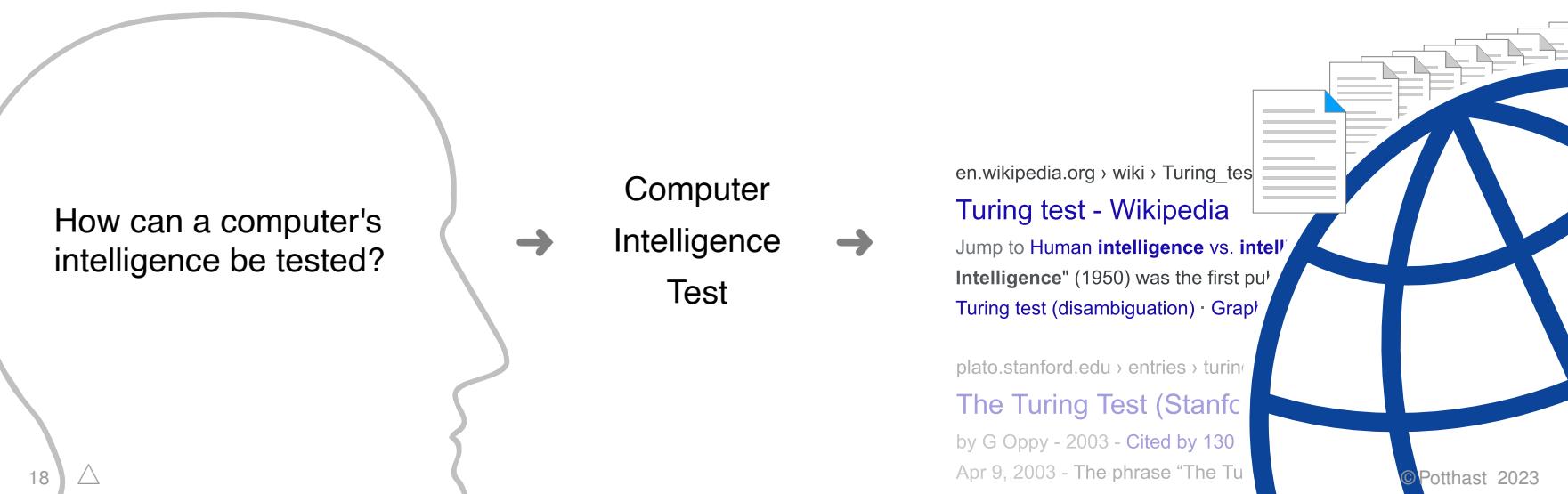
- ❑ Billions of documents.

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.

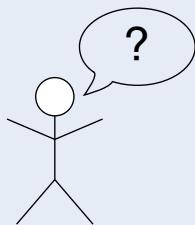


Web Search Architecture



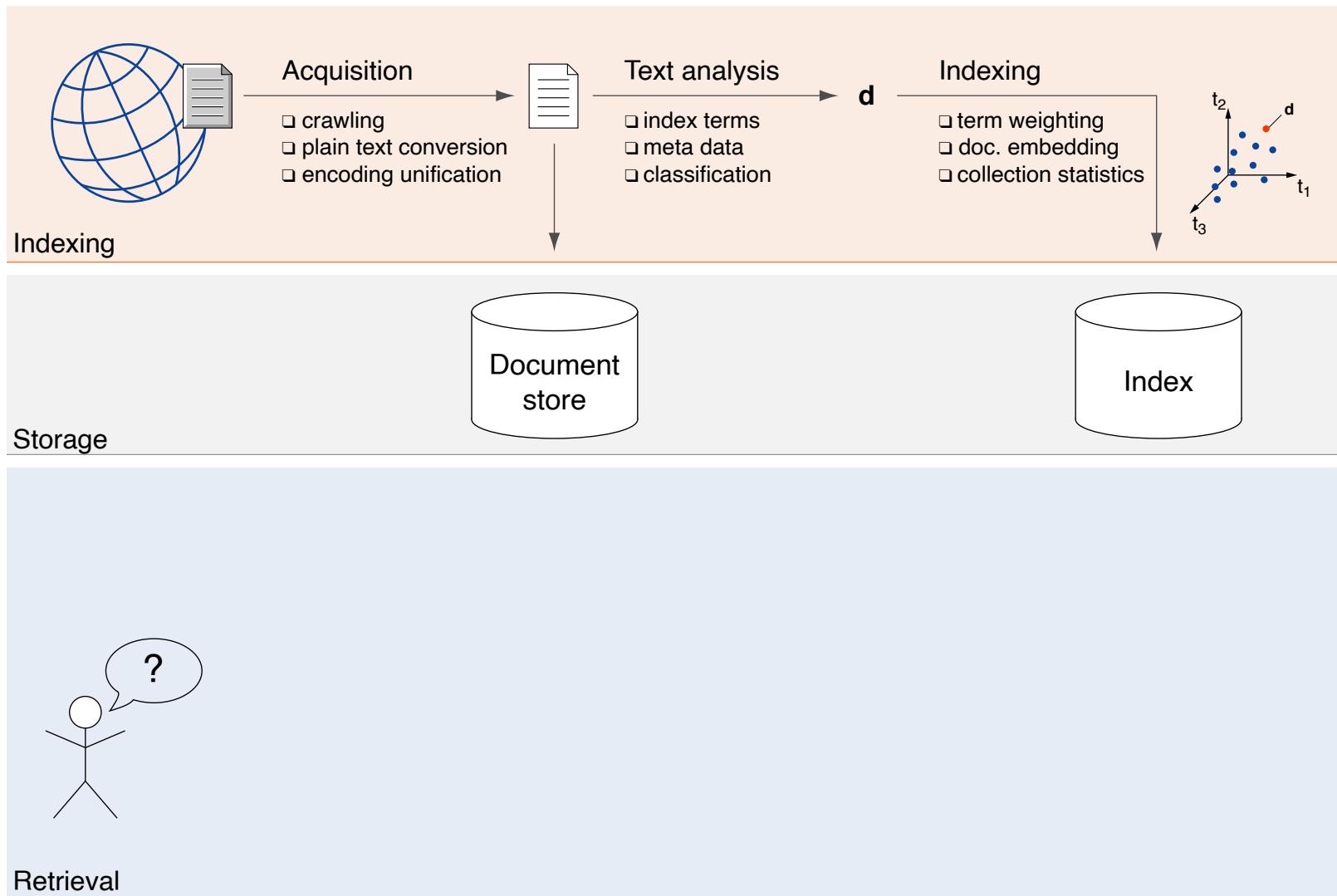
Indexing

Storage

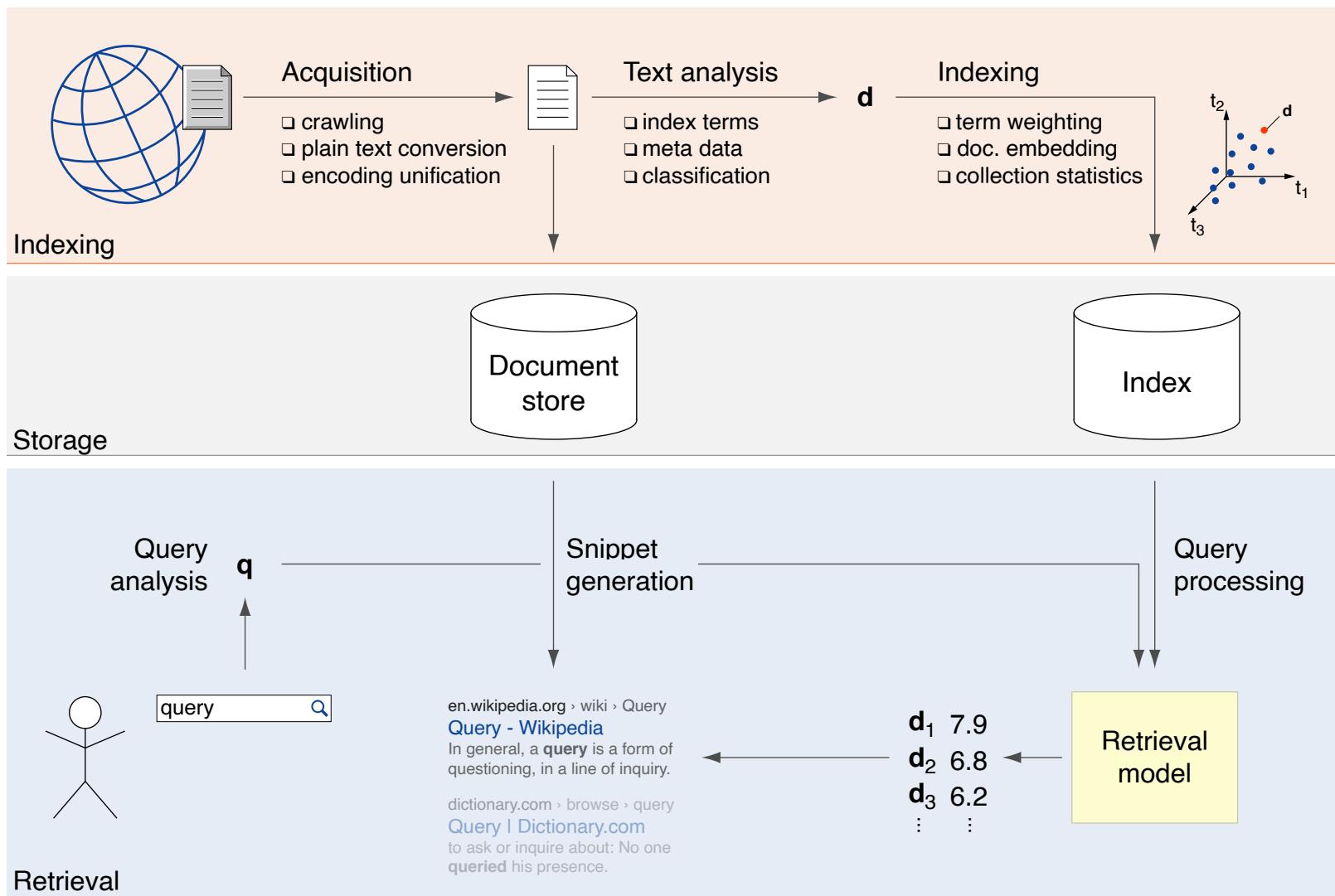


Retrieval

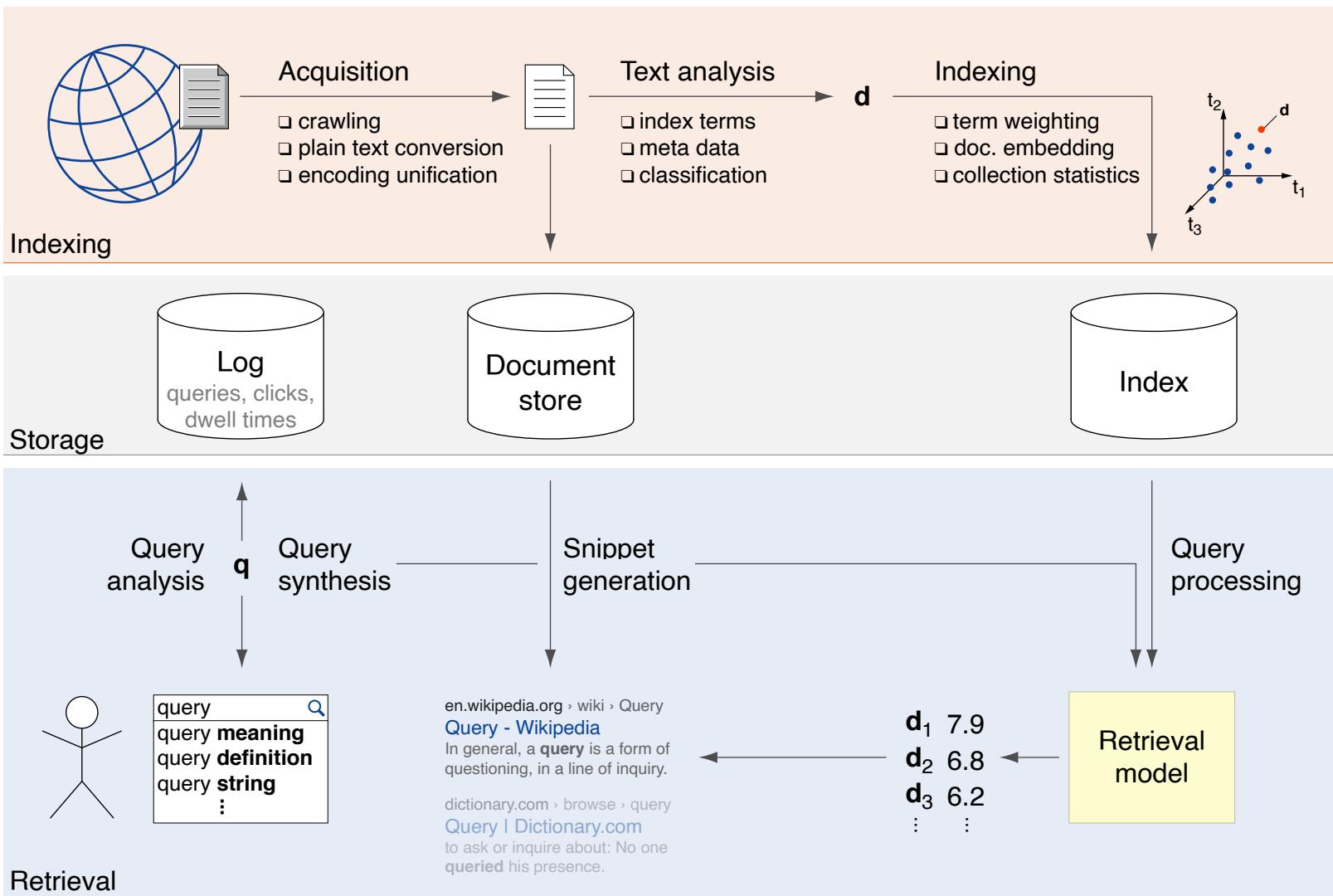
Web Search Architecture



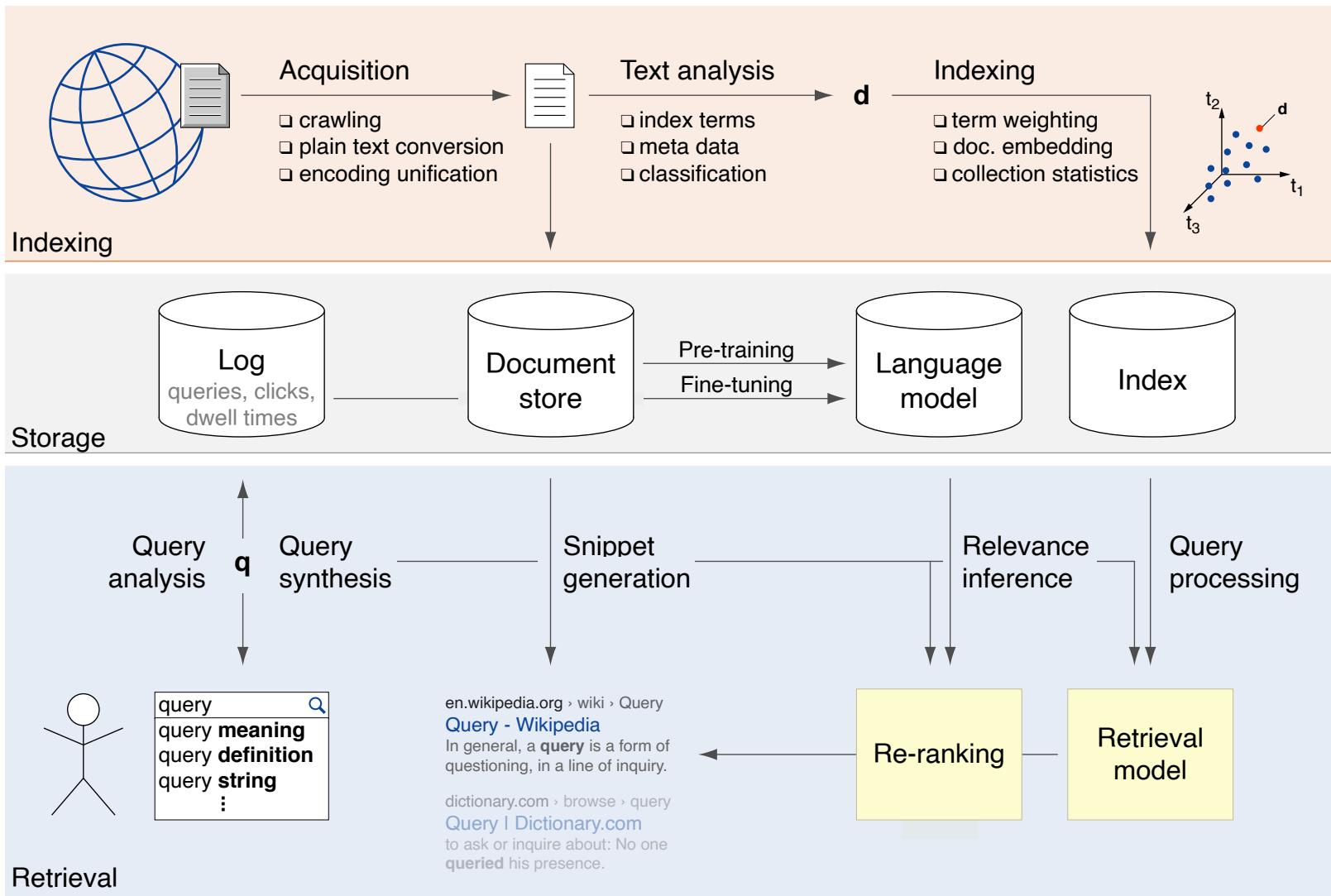
Web Search Architecture



Web Search Architecture



Web Search Architecture



The Archive Query Log (AQL)

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)

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.

Search provider (known domains)	URLs (total)	Queries (total)	Queries (unique)	SERPs (estimate)	Results (estimate)
 Google	89.4 M	72.7 M	20.0 M	34.0 M	270.9 M
 YouTube	41.8 M	41.4 M	11.3 M	19.3 M	411.8 M
 Baidu	78.5 M	69.6 M	2.9 M	32.5 M	130.7 M
: 547 others	646.6 M	161.8 M	27.7 M	75.4 M	839.5 M
Σ	550	1,010.0 M	356.5 M	166.4 M	1,709.0 M

- Use case: Train “Large Ranking Model” for re-ranking.
- Use case: Promote search industry transparency at scale (e.g., fairness).

The ChatNoir Web Search Engine



ChatNoir

Bonjour! I'm ChatNoir, your friendly cat
assistant. How can I help you today?

How can a computer's intelligence be tested?



ChatNoir

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.

www.wired.com/.../pass-turing-ai-test/

- [1] **How to Pass the Turing Artificial Intelligence Test | WIRED**

ClueWeb22 · Crawled Aug 2022 ·

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 – a test of a machine's ability to exhibit intelligent behavior.

www.shivonzilis.com/

- [2] **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 offer more than a system of record; they also need to coach employees and learn from the data they enter.

www.geeksforgeeks.org/turing-test-artificial-intelligence/

- [3] **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 by a machine, He has proposed the following skills of the test as follows: The turning judges the conversational AI to be humans.

Show more...

Big Data Analytics Stack

Data
Consumption
Layer

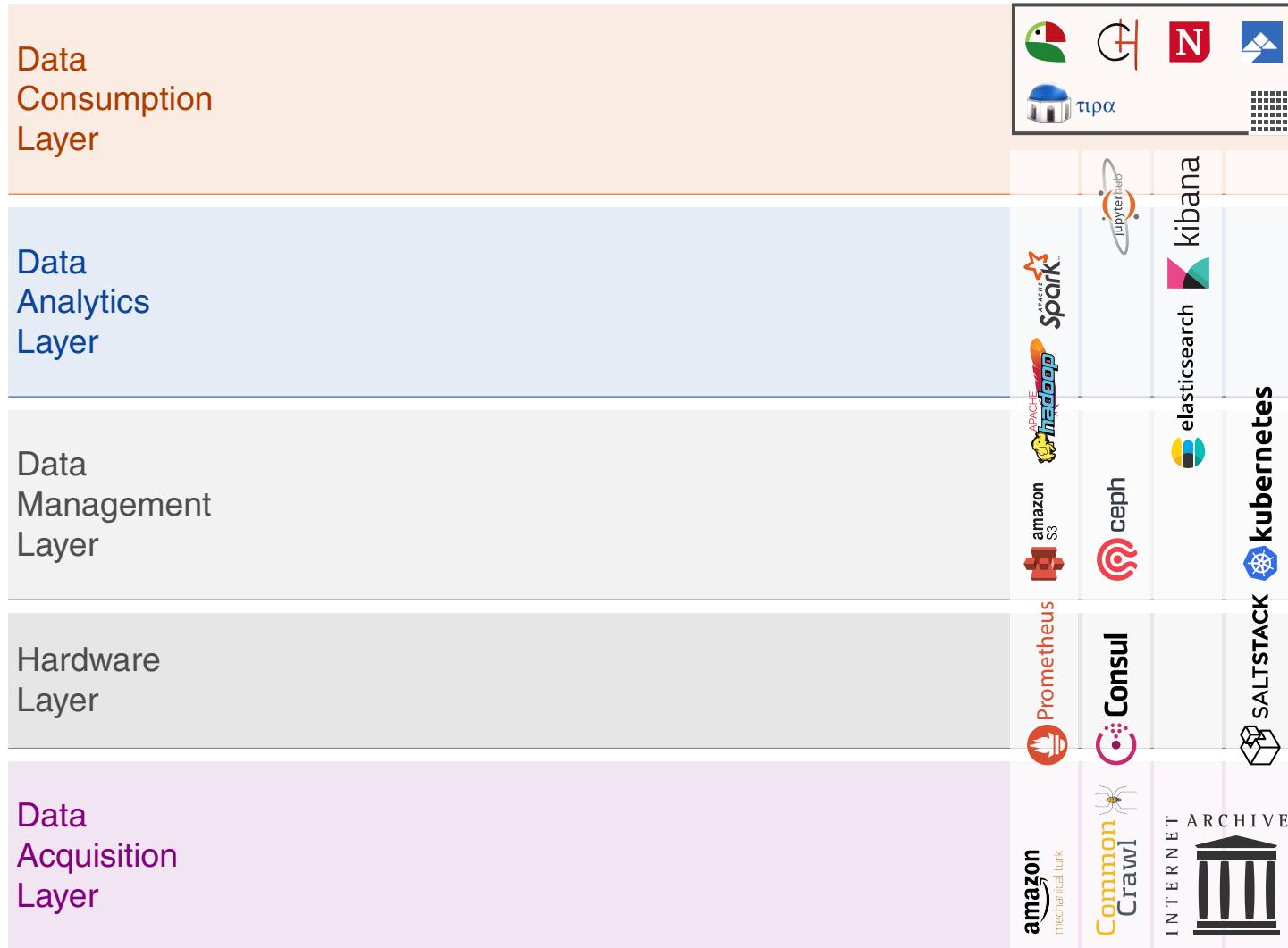
Data
Analytics
Layer

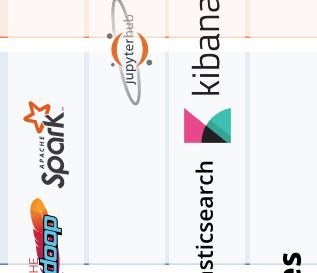
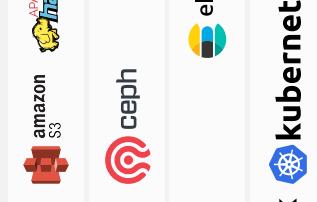
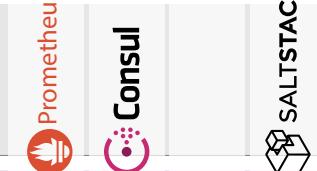
Data
Management
Layer

Hardware
Layer

Data
Acquisition
Layer

Vendor stack

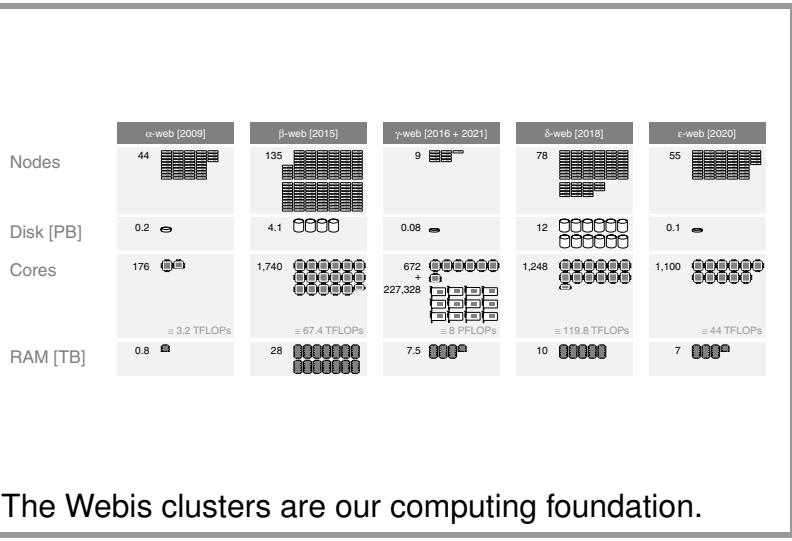


	Technology stack	Vendor stack
Data Consumption Layer	<ul style="list-style-type: none"> - Visual analytics - Immersive technologies - Intelligent agents 	
Data Analytics Layer	<ul style="list-style-type: none"> - Distributed learning - State-space search - Symbolic inference 	
Data Management Layer	<ul style="list-style-type: none"> - Key-value store - RDF triple store - Graph store - Object store 	
Hardware Layer	<ul style="list-style-type: none"> - Orchestration - Parallelization - Virtualization 	
Data Acquisition Layer	<ul style="list-style-type: none"> - Distant supervision - Crowdsourcing - Crawling and archiving 	

	Task Stack	Technology stack	Vendor stack
Data Consumption Layer	<ul style="list-style-type: none"> - Query and explore - Visualize and interact - Explain and justify 	<ul style="list-style-type: none"> - Visual analytics - Immersive technologies - Intelligent agents 	
Data Analytics Layer	<ul style="list-style-type: none"> - Diagnose and reason - Structure identification - Structure verification 	<ul style="list-style-type: none"> - Distributed learning - State-space search - Symbolic inference 	
Data Management Layer	<ul style="list-style-type: none"> - Provenance tracking - Normalization - Cleansing 	<ul style="list-style-type: none"> - Key-value store - RDF triple store - Graph store - Object store 	
Hardware Layer	<ul style="list-style-type: none"> - Monitoring - Replication 	<ul style="list-style-type: none"> - Orchestration - Parallelization - Virtualization 	
Data Acquisition Layer	<ul style="list-style-type: none"> - Replay - Collect - Log 	<ul style="list-style-type: none"> - Distant supervision - Crowdsourcing - Crawling and archiving 	

	Task Stack	Technology stack	Vendor stack	Roles
Data Consumption Layer	<ul style="list-style-type: none"> - Query and explore - Visualize and interact - Explain and justify 	<ul style="list-style-type: none"> - Visual analytics - Immersive technologies - Intelligent agents 		Experts: <ul style="list-style-type: none"> - IR - NLP - CSS - VA
Data Analytics Layer	<ul style="list-style-type: none"> - Diagnose and reason - Structure identification - Structure verification 	<ul style="list-style-type: none"> - Distributed learning - State-space search - Symbolic inference 		Data scientist
Data Management Layer	<ul style="list-style-type: none"> - Provenance tracking - Normalization - Cleansing 	<ul style="list-style-type: none"> - Key-value store - RDF triple store - Graph store - Object store 		Data engineer
Hardware Layer	<ul style="list-style-type: none"> - Monitoring - Replication 	<ul style="list-style-type: none"> - Orchestration - Parallelization - Virtualization 		
Data Acquisition Layer	<ul style="list-style-type: none"> - Replay - Collect - Log 	<ul style="list-style-type: none"> - Distant supervision - Crowdsourcing - Crawling and archiving 		Data scientist

Summary



The Webis clusters are our computing foundation.

Summary

	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
RAM [TB]	0.8	28	7.5	10	7
	$\equiv 3.2 \text{ TFLOPs}$	$\equiv 67.4 \text{ TFLOPs}$	$\equiv 8 \text{ PFLOPs}$	$\equiv 119.8 \text{ TFLOPs}$	$\equiv 44 \text{ TFLOPs}$

The Webis clusters are our computing foundation.



The web archive data is our data foundation.

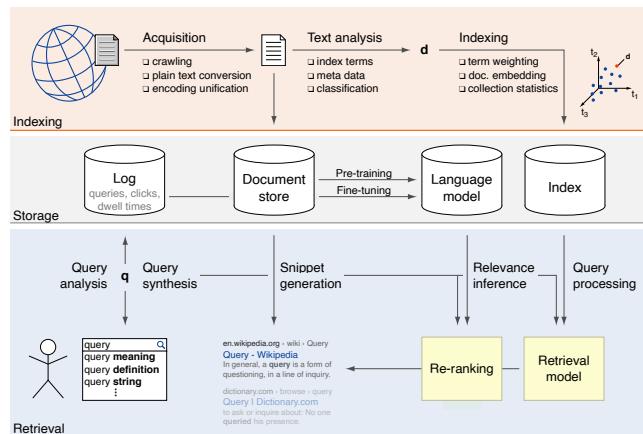
Summary

	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
RAM [TB]	0.8	28	7.5	10	7
	$\equiv 3.2 \text{ TFLOPs}$	$\equiv 67.4 \text{ TFLOPs}$	$\equiv 8 \text{ PFLOPs}$	$\equiv 119.8 \text{ TFLOPs}$	$\equiv 44 \text{ TFLOPs}$

The Webis clusters are our computing foundation.



The web archive data is our data foundation.



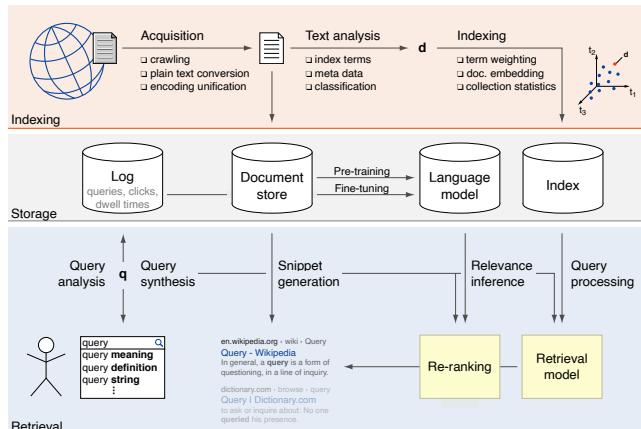
We develop open search engines that scale to the web.



Summary

	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	227,328	1,248	1,100
RAM [TB]	0.8	28	7.5	10	7
	$\equiv 3.2 \text{ TFLOPs}$	$\equiv 67.4 \text{ TFLOPs}$	$\equiv 8 \text{ PFLOPs}$	$\equiv 119.8 \text{ TFLOPs}$	$\equiv 44 \text{ TFLOPs}$

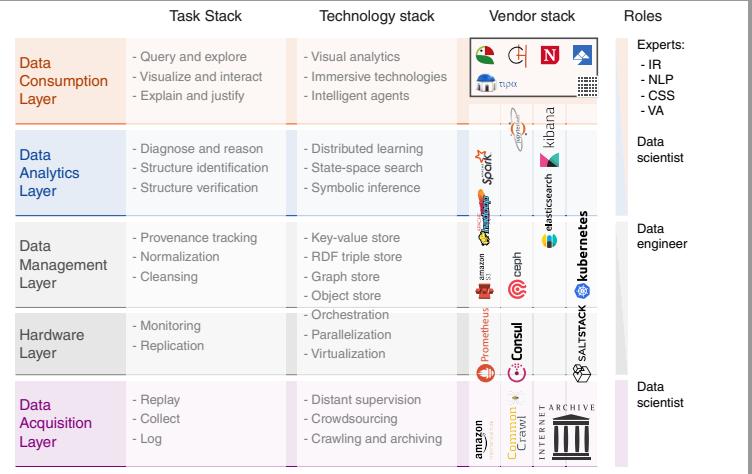
The Webis clusters are our computing foundation.



We develop open search engines that scale to the web.



The web archive data is our data foundation.



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

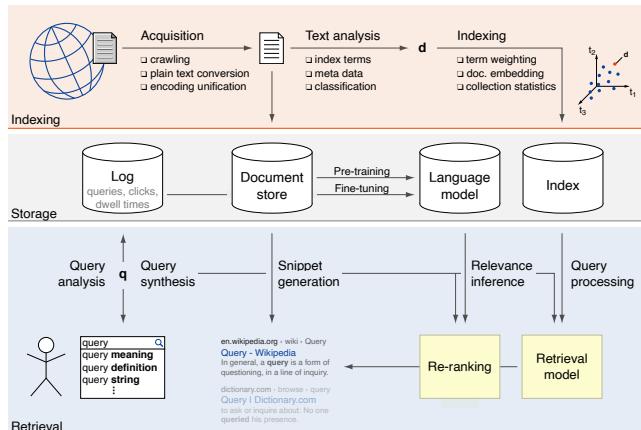
Summary

Thank you!

leipzig.webis.de

	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
RAM [TB]	0.8	28	7.5	10	7
	≈ 3.2 TFLOPs	≈ 67.4 TFLOPs	≈ 8 PFLOPs	≈ 119.8 TFLOPs	≈ 44 TFLOPs

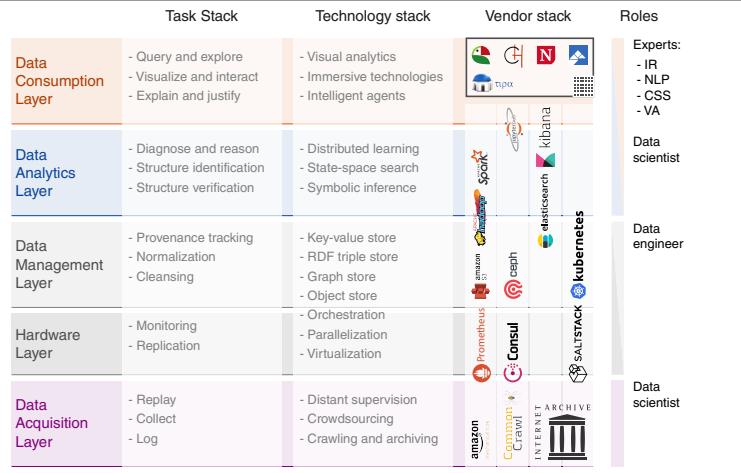
The Webis clusters are our computing foundation.



We develop open search engines that scale to the web.



The web archive data is our data foundation.



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