

# 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](http://leipzig.webis.de) | [webis.de](http://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**

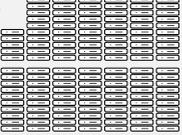
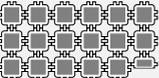
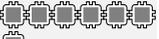
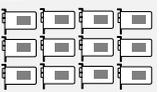
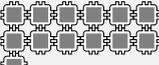
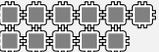
(big data analytics, semi-supervised, one-class, domain transfer, LLMs)

- **Shared Tasks and Reproducibility**

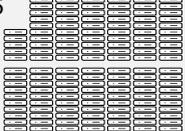
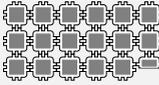
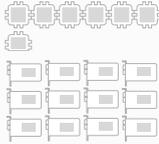
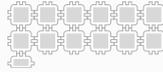
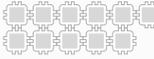
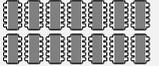
(corpus construction, crowdsourcing, acquisition of compiled expertise)

# Webis Computing Facilities

# Webis Computing Facilities

	$\alpha$ -web [2009]	$\beta$ -web [2015]	$\gamma$ -web [2016 + 2021]	$\delta$ -web [2018]	$\varepsilon$ -web [2020]
Nodes	44 	135 	9 	78 	55 
Disk [PB]	0.2 	4.1 	0.08 	12 	0.1 
Cores	176   $\cong 3.2$ TFLOPs	1,740   $\cong 67.4$ TFLOPs	672  + 227,328   $\cong 8$ PFLOPs	1,248   $\cong 119.8$ TFLOPs	1,100   $\cong 44$ TFLOPs
RAM [TB]	0.8 	28 	7.5 	10 	7 

# Webis Computing Facilities

	$\alpha$ -web [2009]	$\beta$ -web [2015]	$\gamma$ -web [2016 + 2021]	$\delta$ -web [2018]	$\epsilon$ -web [2020]
Nodes	44 	135 	9 	78 	55 
Disk [PB]	0.2 	4.1 	0.08 	12 	0.1 
Cores	176   $\approx 3.2$ TFLOPs	1,740   $\approx 67.4$ TFLOPs	672 + 227,328   $\approx 8$ PFLOPs	1,248   $\approx 119.8$ TFLOPs	1,100   $\approx 44$ TFLOPs
RAM [TB]	0.8 	28 	7.5 	10 	7 

## Typical research tasks:

$\alpha$ -Web. Teaching, Staging environment

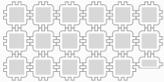
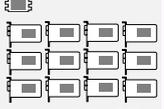
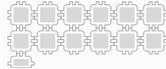
$\beta$ -Web. Virtualization (compute, web services), Web mining (map reduce), Authorship analytics

$\gamma$ -Web. Machine learning (embedding, deep learning), Text synthesis, Language modeling

$\delta$ -Web. Virtualization (storage), Web archiving

$\epsilon$ -Web. Search index construction, Argument search

# Webis Computing Facilities

	$\alpha$ -web [2009]	$\beta$ -web [2015]	$\gamma$ -web [2016 + 2021]	$\delta$ -web [2018]	$\epsilon$ -web [2020]
Nodes	44 	135 	9 	78 	55 
Disk [PB]	0.2 	4.1 	0.08 	12 	0.1 
Cores	176   $\approx 3.2$ TFLOPs	1,740   $\approx 67.4$ TFLOPs	672  + 227,328   $\approx 8$ PFLOPs	1,248   $\approx 119.8$ TFLOPs	1,100   $\approx 44$ TFLOPs
RAM [TB]	0.8 	28 	7.5 	10 	7 

## Typical research tasks:

$\alpha$ -Web. Teaching, Staging environment

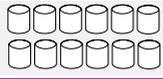
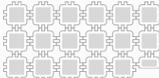
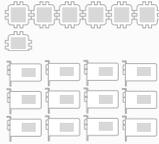
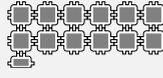
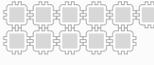
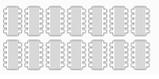
$\beta$ -Web. Virtualization (compute, web services), Web mining (map reduce), Authorship analytics

$\gamma$ -Web. Machine learning (embedding, deep learning), Text synthesis, Language modeling

$\delta$ -Web. Virtualization (storage), Web archiving

$\epsilon$ -Web. Search index construction, Argument search

# Webis Computing Facilities

	$\alpha$ -web [2009]	$\beta$ -web [2015]	$\gamma$ -web [2016 + 2021]	$\delta$ -web [2018]	$\epsilon$ -web [2020]
Nodes	44 	135 	9 	78 	55 
Disk [PB]	0.2 	4.1 	0.08 	12 	0.1 
Cores	176   $\approx 3.2$ TFLOPs	1,740   $\approx 67.4$ TFLOPs	672 + 227,328   $\approx 8$ PFLOPs	1,248   $\approx 119.8$ TFLOPs	1,100   $\approx 44$ TFLOPs
RAM [TB]	0.8 	28 	7.5 	10 	7 

## Typical research tasks:

$\alpha$ -Web. Teaching, Staging environment

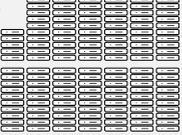
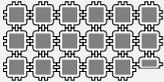
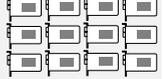
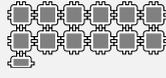
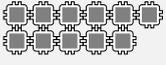
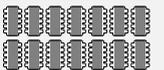
$\beta$ -Web. Virtualization (compute, web services), Web mining (map reduce), Authorship analytics

$\gamma$ -Web. Machine learning (embedding, deep learning), Text synthesis, Language modeling

$\delta$ -Web. Virtualization (storage), Web archiving

$\epsilon$ -Web. Search index construction, Argument search

# Webis Computing Facilities

	$\alpha$ -web [2009]	$\beta$ -web [2015]	$\gamma$ -web [2016 + 2021]	$\delta$ -web [2018]	$\epsilon$ -web [2020]
Nodes	44 	135 	9 	78 	55 
Disk [PB]	0.2 	4.1 	0.08 	12 	0.1 
Cores	176   $\cong 3.2$ TFLOPs	1,740   $\cong 67.4$ TFLOPs	672  + 227,328   $\cong 8$ PFLOPs	1,248   $\cong 119.8$ TFLOPs	1,100   $\cong 44$ TFLOPs
RAM [TB]	0.8 	28 	7.5 	10 	7 

## Typical research tasks:

$\alpha$ -Web. Teaching, Staging environment

$\beta$ -Web. Virtualization (compute, web services), Web mining (map reduce), Authorship analytics

$\gamma$ -Web. Machine learning (embedding, deep learning), Text synthesis, Language modeling

$\delta$ -Web. Virtualization (storage), Web archiving

$\epsilon$ -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 INTERNET ARCHIVE **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)



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

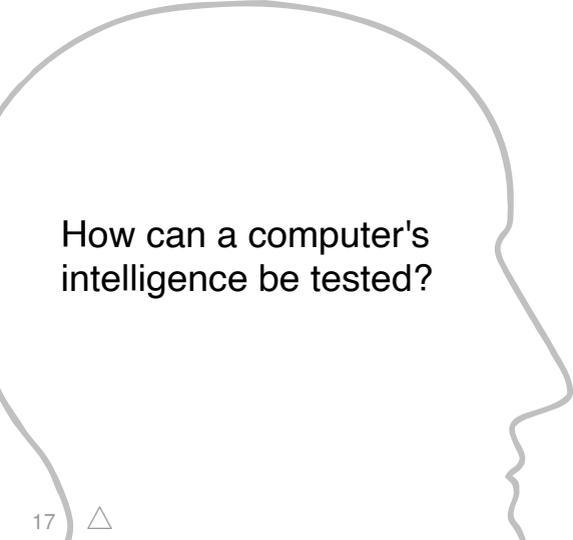
[[archive.webis.de](https://archive.webis.de)]

# 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, ...

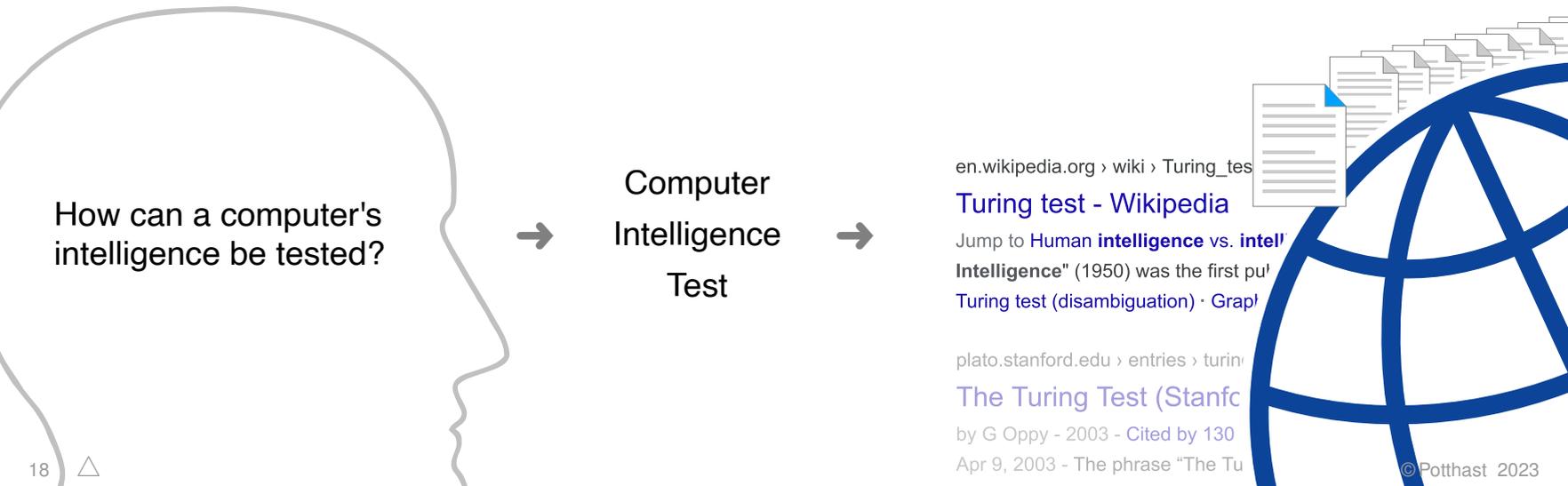


How can a computer's intelligence be tested?



# 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.

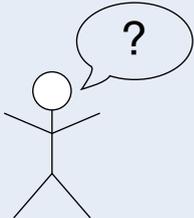


# Web Search Architecture



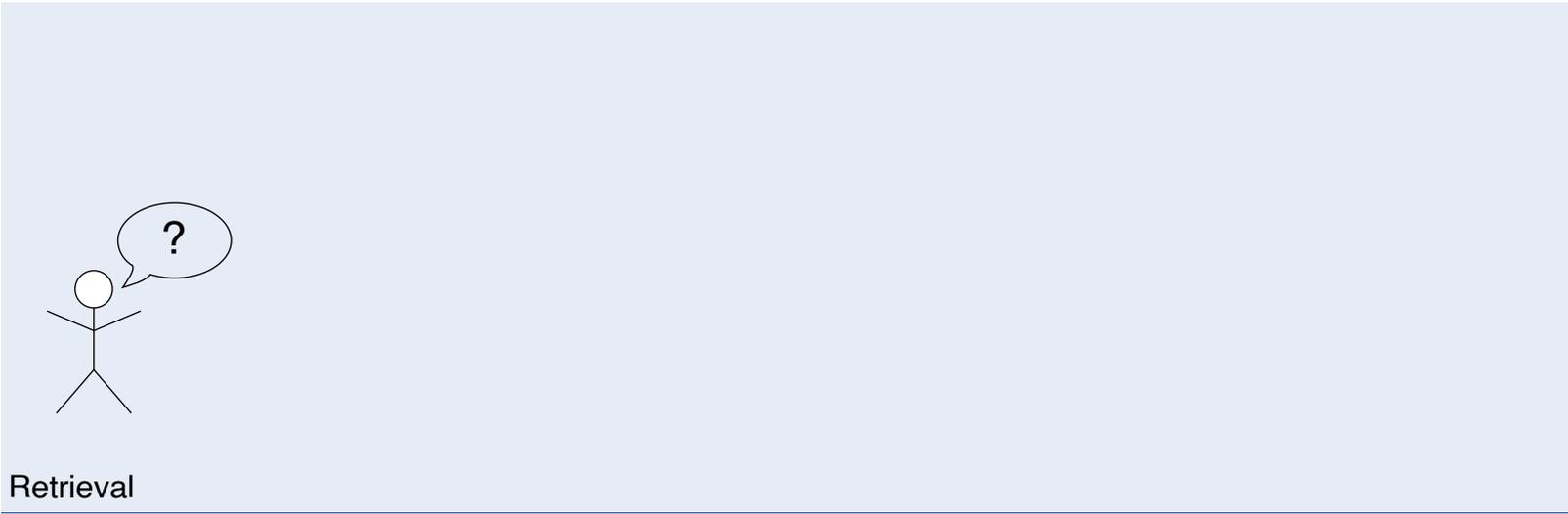
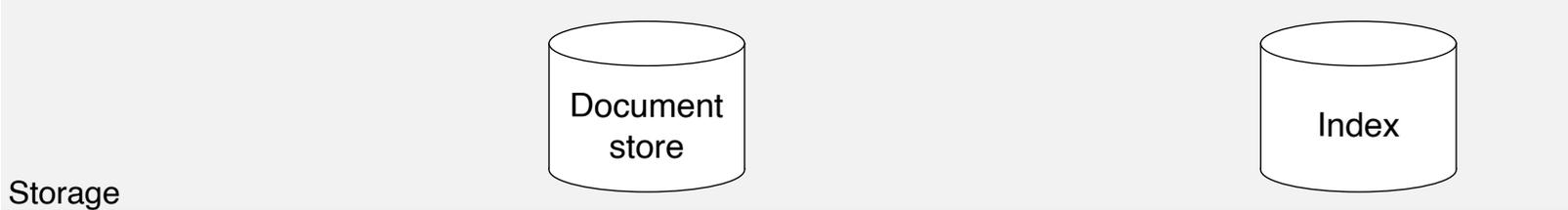
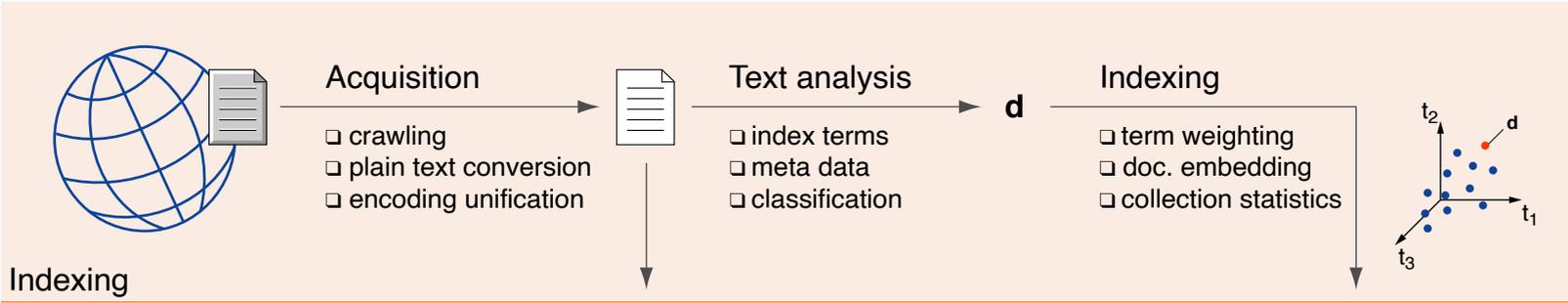
Indexing

Storage

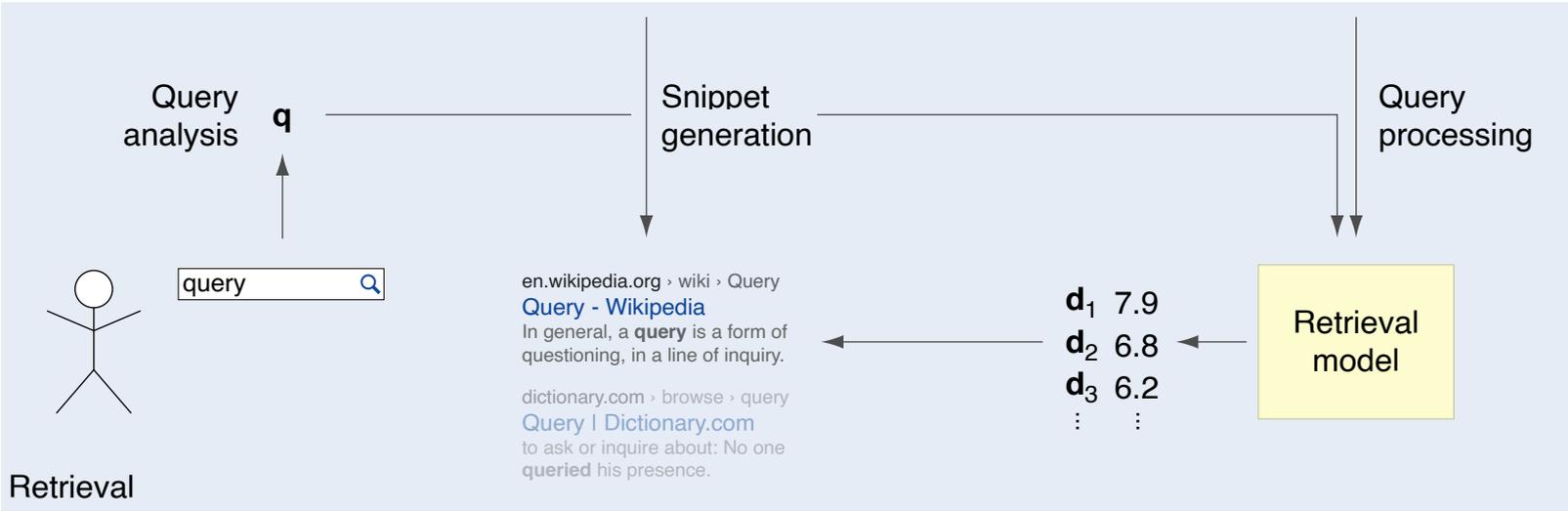
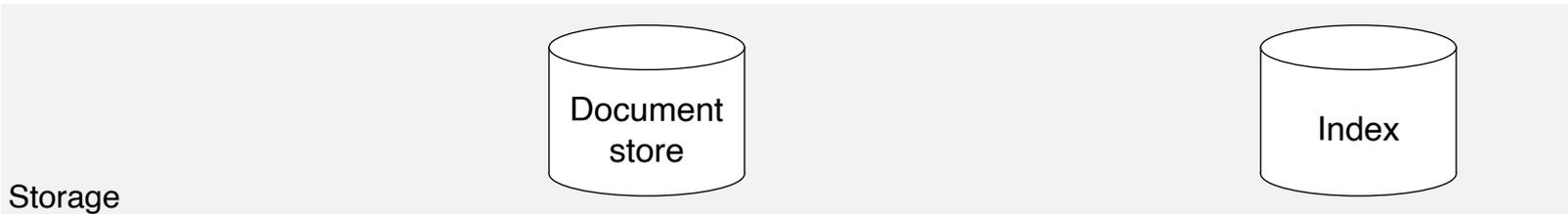
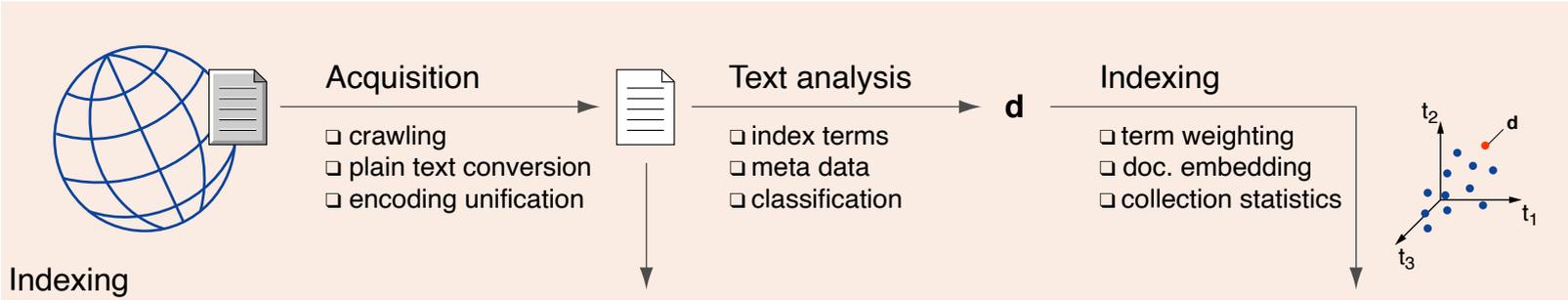


Retrieval

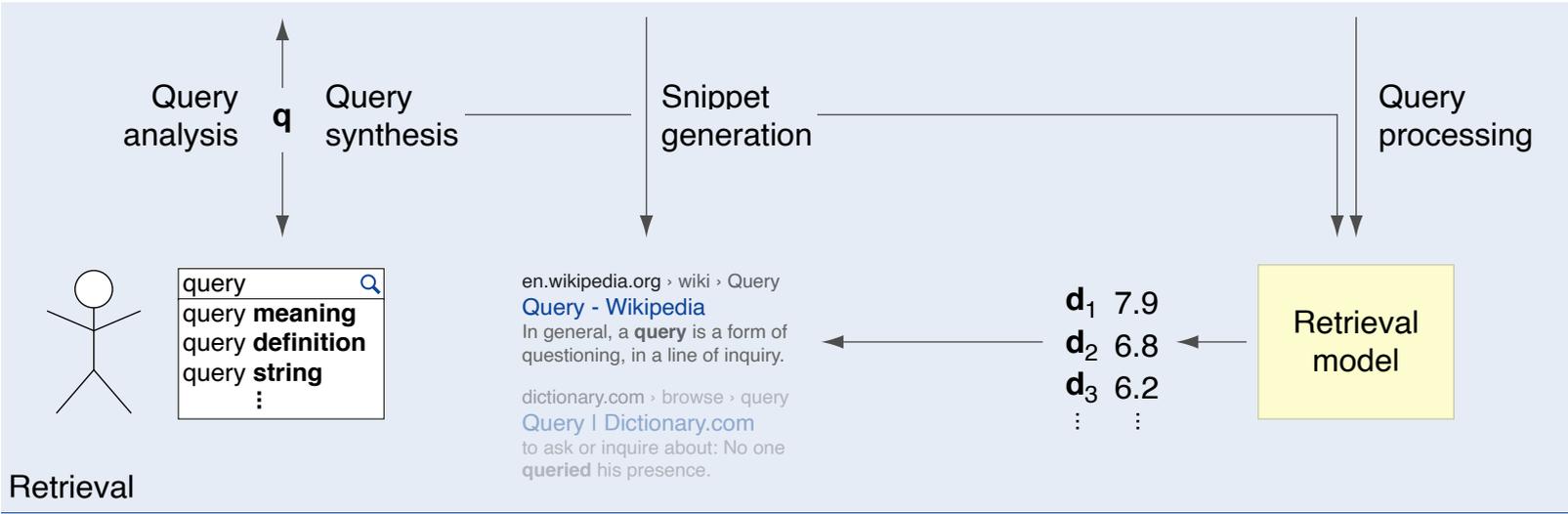
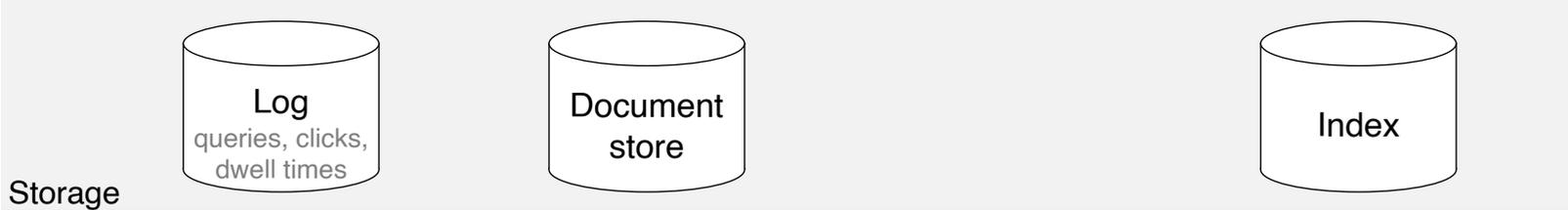
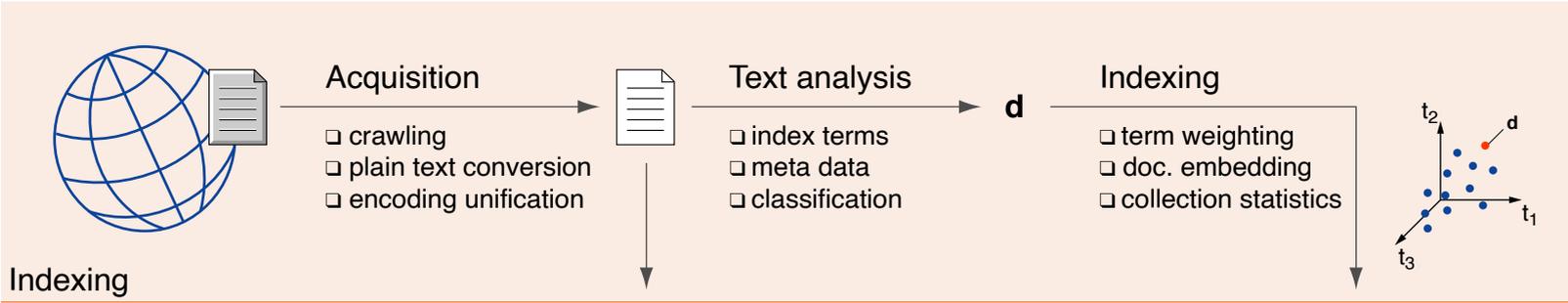
# Web Search Architecture



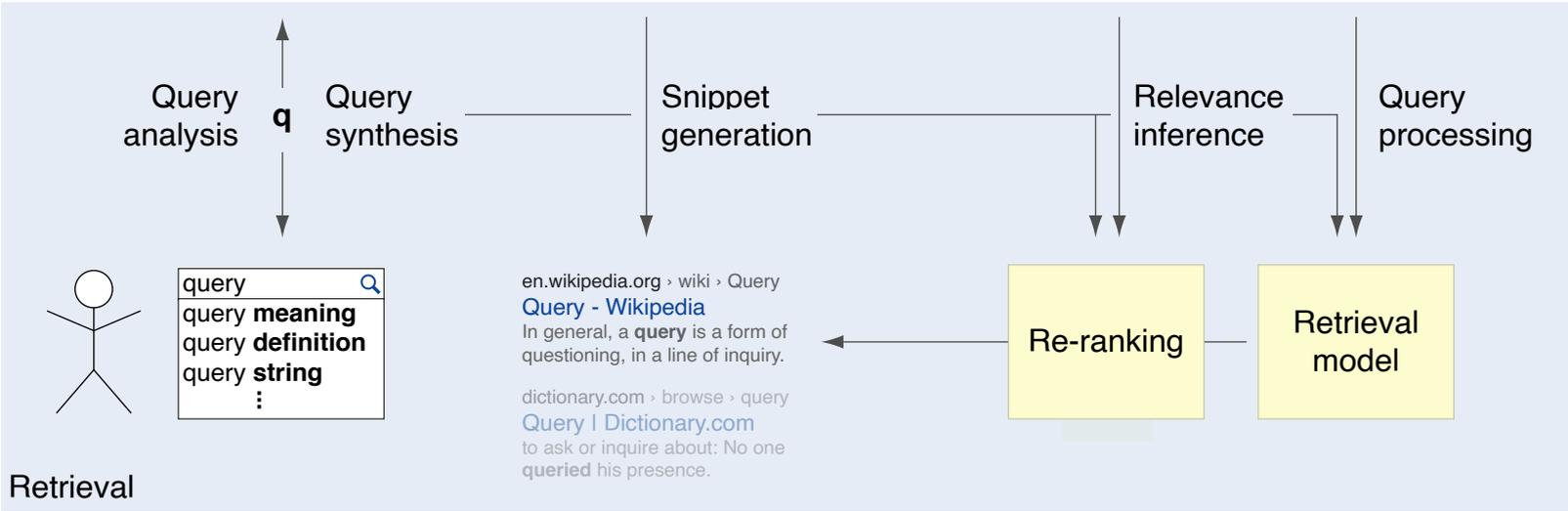
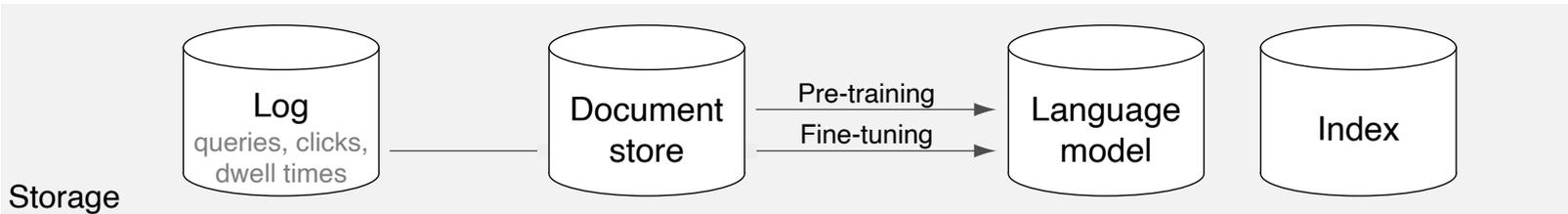
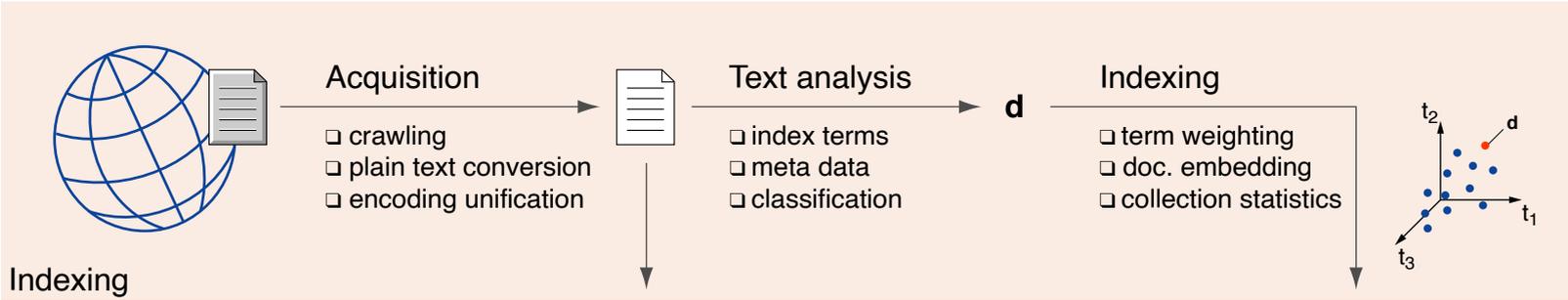
# Web Search Architecture



# 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
<b>∑ 550</b>	<b>1,010.0 M</b>	<b>356.5 M</b>	<b>64.5 M</b>	<b>166.4 M</b>	<b>1,709.0 M</b>

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



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.

[1] [How to Pass the Turing Artificial Intelligence Test | WIRED](#)

www.wired.com/.../pass-turing-ai-test/  
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 the ability of a machine to exhibit intelligent behavior.

[2] [Machine Intelligence - Shivon Zilis](#)

www.shivonzilis.com/  
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.

[3] [Turing Test in Artificial Intelligence - GeeksforGeeks](#)

www.geeksforgeeks.org/turing-test-artificial-intelligence/  
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 skills of humans.

Show more...



# Big Data Analytics Stack

Data  
Consumption  
Layer

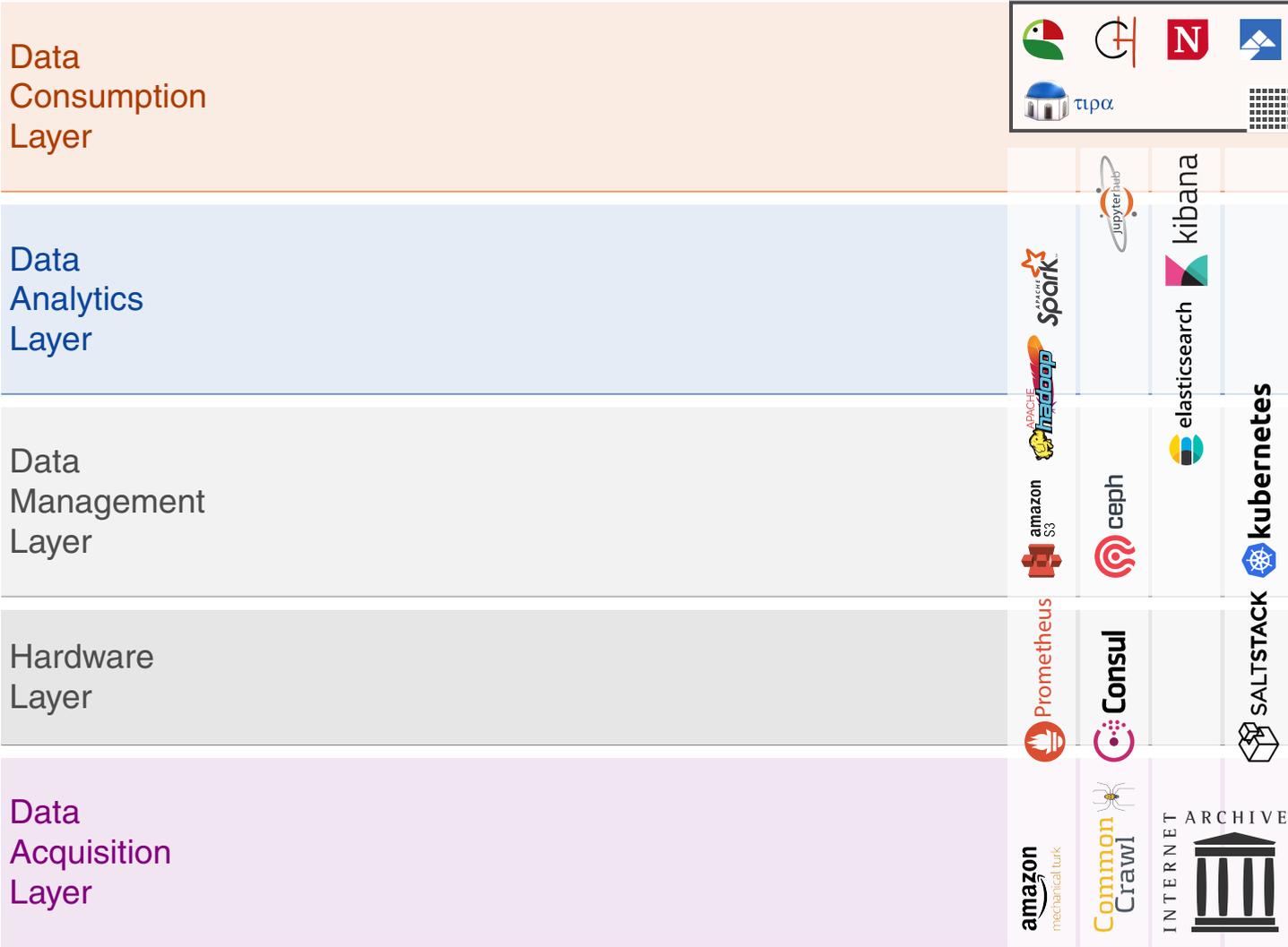
Data  
Analytics  
Layer

Data  
Management  
Layer

Hardware  
Layer

Data  
Acquisition  
Layer

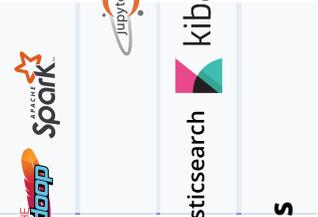
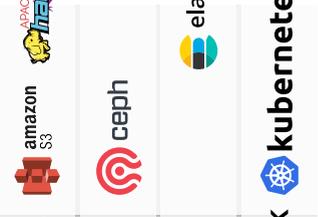
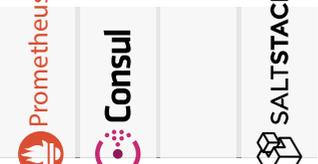
# Vendor stack



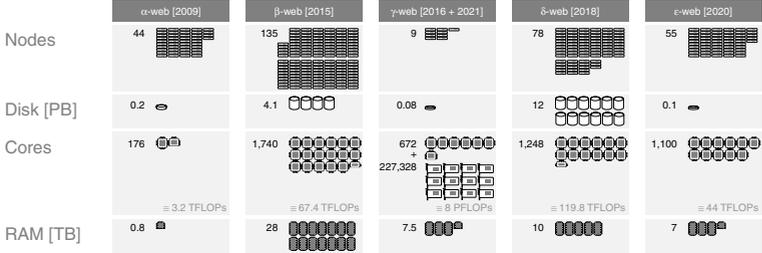


	Task Stack	Technology stack	Vendor stack
<b>Data Consumption Layer</b>	<ul style="list-style-type: none"> <li>- Query and explore</li> <li>- Visualize and interact</li> <li>- Explain and justify</li> </ul>	<ul style="list-style-type: none"> <li>- Visual analytics</li> <li>- Immersive technologies</li>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



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

# 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 ≈ 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

The Webis clusters are our computing foundation.



The web archive data is our data foundation.

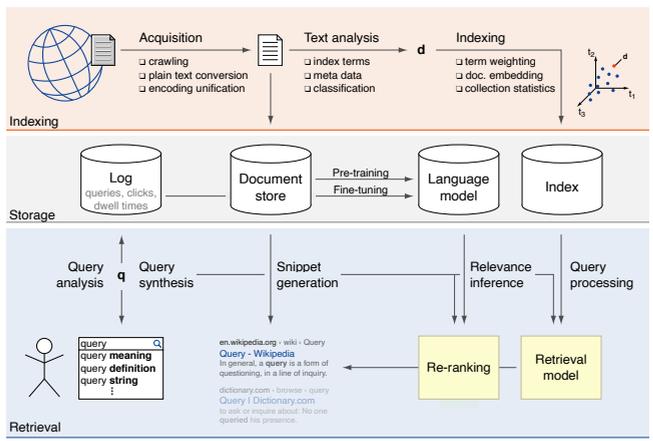
# Summary

	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

The Webs clusters are our computing foundation.



The web archive data is our data foundation.



We develop open search engines that scale to the web.



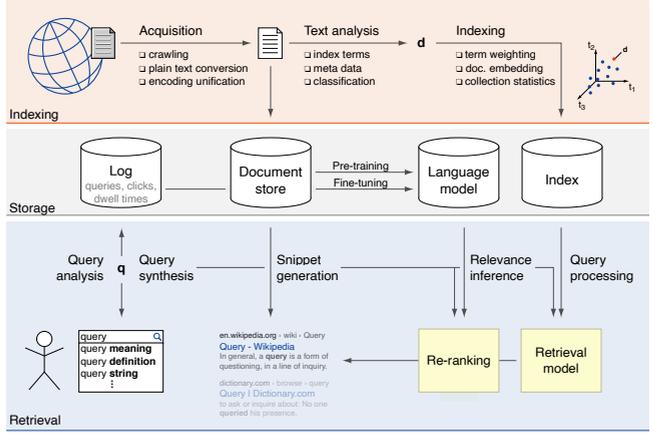
# Summary

	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	672 + 227,328	1,248	1,100
RAM [TB]	0.8	28	7.5	10	7

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.

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

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

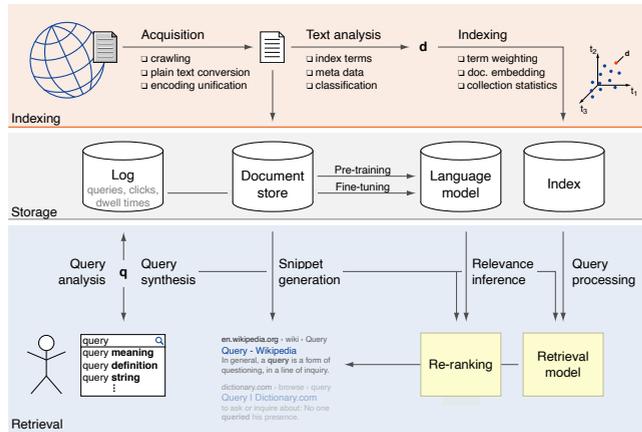


	α-web [2009]	β-web [2015]	γ-web [2016 + 2021]	δ-web [2018]	ε-web [2020]
Nodes	44	135	9	78	55
Disk [PB]	0.2	4.1	0.08	12	0.1
Cores	176	1,740	672 + 227,328	1,248	1,100
RAM [TB]	0.8	28	7.5	10	7

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.

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

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