- World Wide
- Content Extraction
- Service and Solution
- for Enterprises
- and Individuals

Webstructor Mine: Goals and Use Cases

The **ABC company** has completed spin off of **subsidiary** **XYZ company** under the contract of **5 million dollars**.
CEO of the **ABC** company **John Doe** yesterday said that they are going to …

Today, in its **Georgetown, TX** headquarters, **DEF corporation** has announced its new **spin-off UVW**. The sum of deal is announced to be **8.5 million dollars**.

Extract

Train

Webstructor Mine

ABC Subsidiary XYZ Spin-Off $5M

Theme:	**Spin-Off**
Parent Company:	**ABC**
Child Company:	**XYZ**
Asset:	**Subsidiary**
Amount:	**5**

Explore

DEF Subsidiary UVF Spin-Off $8.5M

Theme:	**Spin-Off**
Parent Company:	**DEF**
Child Company:	**UVF**
Asset:	**Subsidiary**
Amount:	**8.5**

Webstructor Mine: Content Extraction Example

Webstructor Mine: Document View UI

Webstructor Mine: Document View

Asset: `My Asset 1`   View: `Documents`   ☒ Categories & Features   ☐ Features & Categories   ☒ Objects & Properties

Documents for `DocSrc1` with `yesterday`

where `State` is `NY` and `___`

Name
```
http:\\1.2.3.4\news\doc77.txt
\\myhost\data\press\doc35.txt
c:\documents\document89.html
```

`<New Document>` Add Delete Update

Document   ◉ Highlights ○ Original

The company ABC has **spin off subsidiary** company XYZ for 5000000 USD. CEO of the company John Belle yesterday said that they are going to compete with PQR subsidiary of DEF corp. spun off for 4 million euro year before.

Categories for Document and Domain `Dom2`

Name	Relevance	Confirm
SpinOff	▫▫▫▫▫	+
Merge	■	

`<New Category>` Add Confirm

Features for Document Category

Name	Relevance	Confirm
spin off	▫▫▫▫	+
spun	■■■	+
subsidiary	■■■	+

`<New Feature>` Add Confirm

Objects for Document Category

Comp	AssetType	Spun	Amount	Unit	Curr
ABC	subsidiary	XYZ	5	million	USD
DEF	subsidiary	PQR	4	million	EUR

Add Delete

Object Properties

Property	Domain	Category
Company	Company	ABC
Asset Type	Asset Type	Subsidiary
SpunOffCompany	Company	XYZ
Amount		
Unit		

`<New Domain>` `<New Property>` Add Delete

Property Categories

Name	Relevance	Confirm
Subsidiary	▫▫▫▫	+
Property	■	
Facility	■	

`<New Category>` Add Confirm

Document-centric user interface allows online users to search, browse and view existent documents, explore categories assigned to them and objects created from them or assign new categories or create new objects.

© Anton Kolonin 2008   4

Webstructor Mine: Category View UI

Webstructor Mine: Category View

Asset [My Asset 1]   View [Categories]   [X] Categories & Features   [X] Documents   [X] Objects & Properties

Categories for [Asset Type] **with** [sub pro fac]

Name
Subsidiary
Property
Facility

[<New Category>] Add Delete Update

Features for Category

Name	Relevance	Confirm
spin off	■■■■	+
spun	■■■	+
subsidiary	■■■	+

[<New Feature>] Add Confirm

Documents for [State]

[DocSrc1] [NY]

Name		Relevance	Confirm
http:\\1.2.3.4\news\doc77.txt	■■■		+
\\myhost\data\press\doc35.txt	■■		+
c:\documents\document89.html	■■		+

[<New Document>] Add Confirm

Document   ◉ Highlights ○ Original

The company ABC has **spin off subsidiary** company XYZ for 5000000 USD. CEO of the company John Belle yesterday said that they are going to compete with PQR subsidiary of DEF corp. spun off for 4 million euro year before.

Objects for Category Document

Comp	AssetType	Spun	Amount	Unit	Curr
ABC	subsidiary	XYZ	5	million	USD
DEF	subsidiary	PQR	4	million	EUR

Add Delete

Object Properties

Property	Domain	Category
Company	Company	ABC
Asset Type	Asset Type	Subsidiary
SpunOffCompany	Company	XYZ
Amount		
Unit		

[<New Domain>] [<New Property>] Add Delete

Property Categories

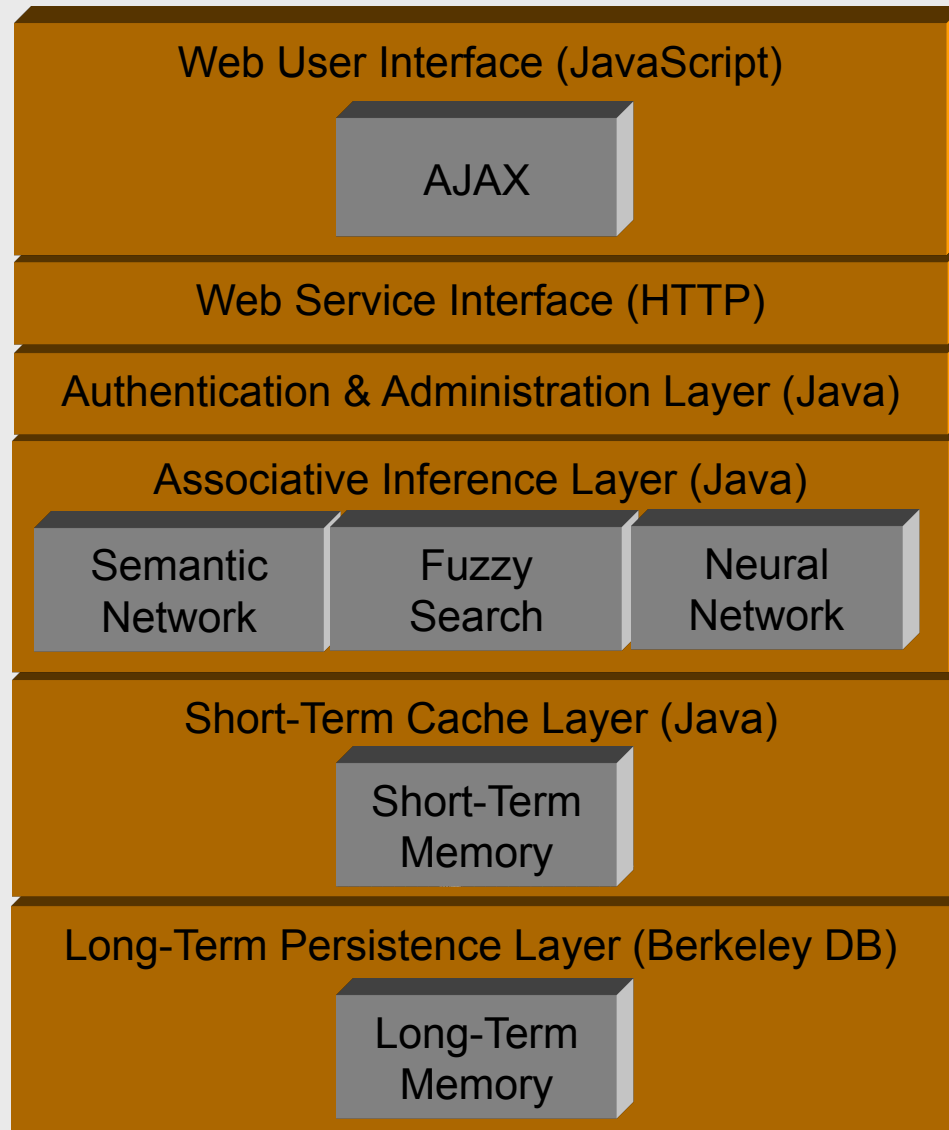| Name	Relevance	Confirm
Subsidiary	■■■■	+
Property	■	
Facility	■	

[<New Category>] Add Confirm

> Category-centric user interface allows online users to search, browse and view existent categories, explore documents they are assigned to and objects involving them or associate new documents or create new objects.

Webstructor Mine: Content Extraction Precision

- 76%-96% Software matches Human
- 1-8% Software performs better than Human
- 1-8% Human typos in training corpus
- 1-8% Lack of data in training corpus
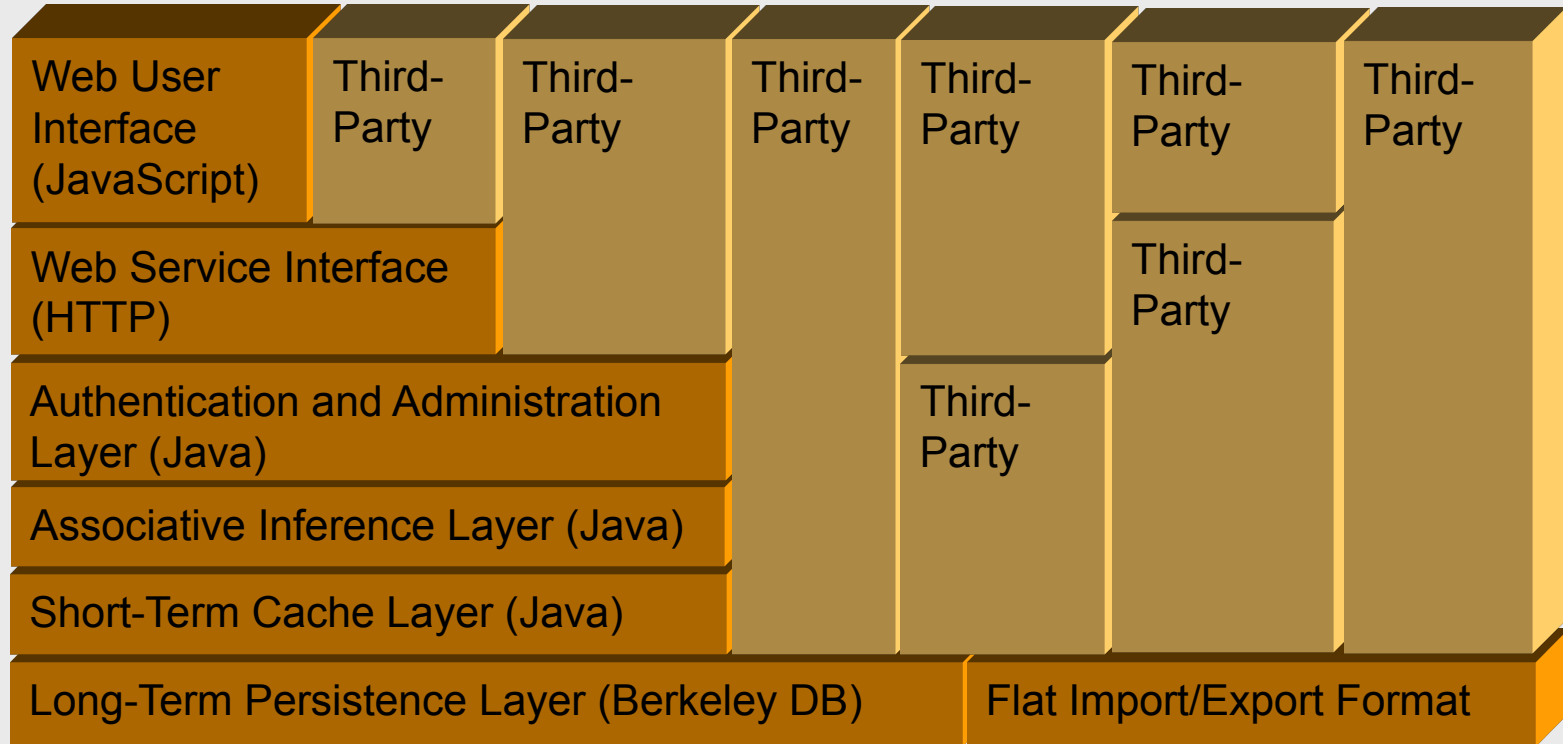- 1-8% Software lacks Human-level intelligence
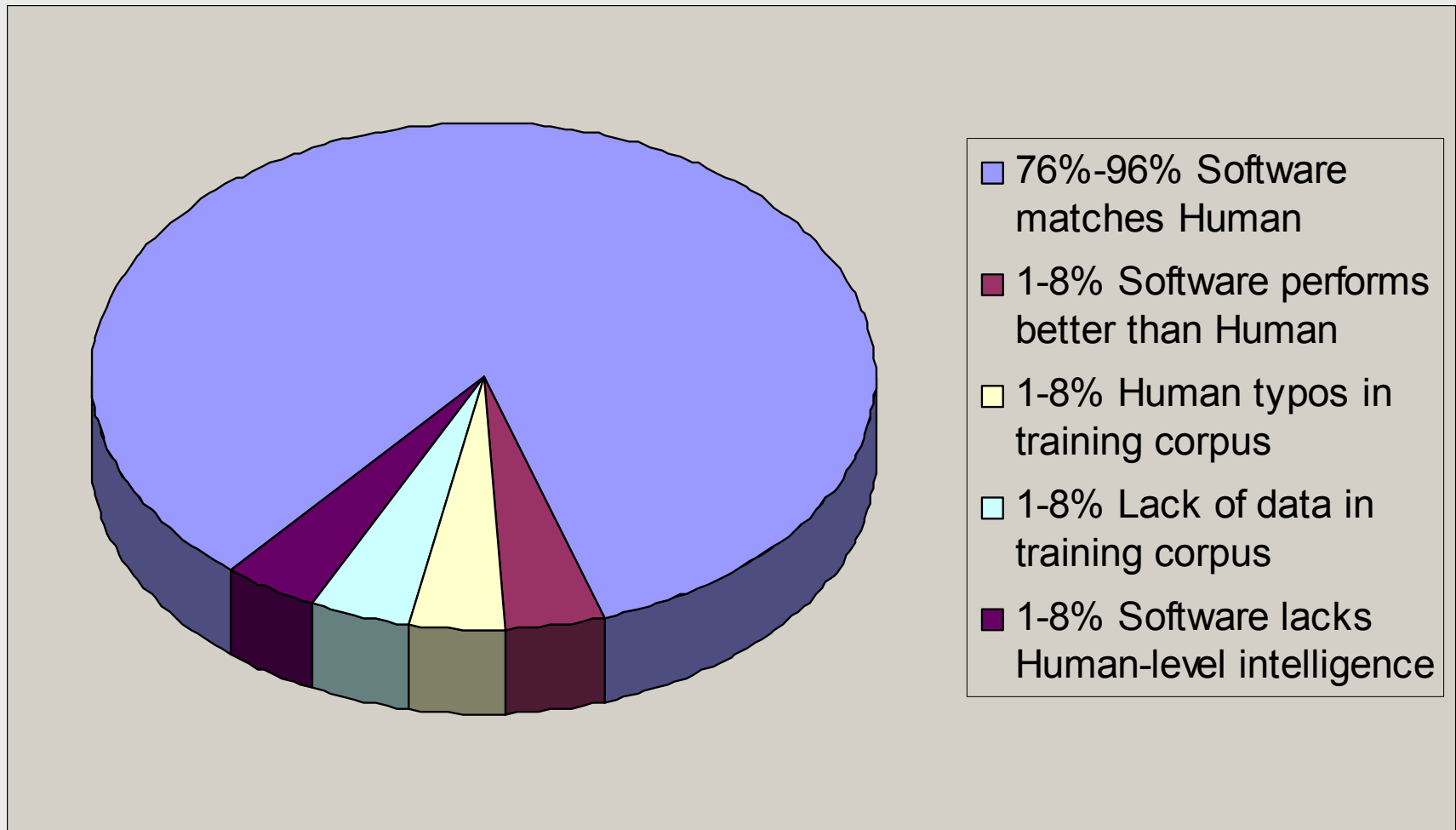
The approach has been tested with various training and testing sets of size such as 25 domains, 500 thousand categories, 2 million objects and documents, 25 million ternary and binary associations. Testing results unveil the automated extraction performance is comparable or better than performance of human specialists.

Webstructor Mine: Know-How Ingredients

Web User Interface (JavaScript)

AJAX

Web Service Interface (HTTP)

Authentication & Administration Layer (Java)

Associative Inference Layer (Java)

Semantic Network

Fuzzy Search

Neural Network

Short-Term Cache Layer (Java)

Short-Term Memory

Long-Term Persistence Layer (Berkeley DB)
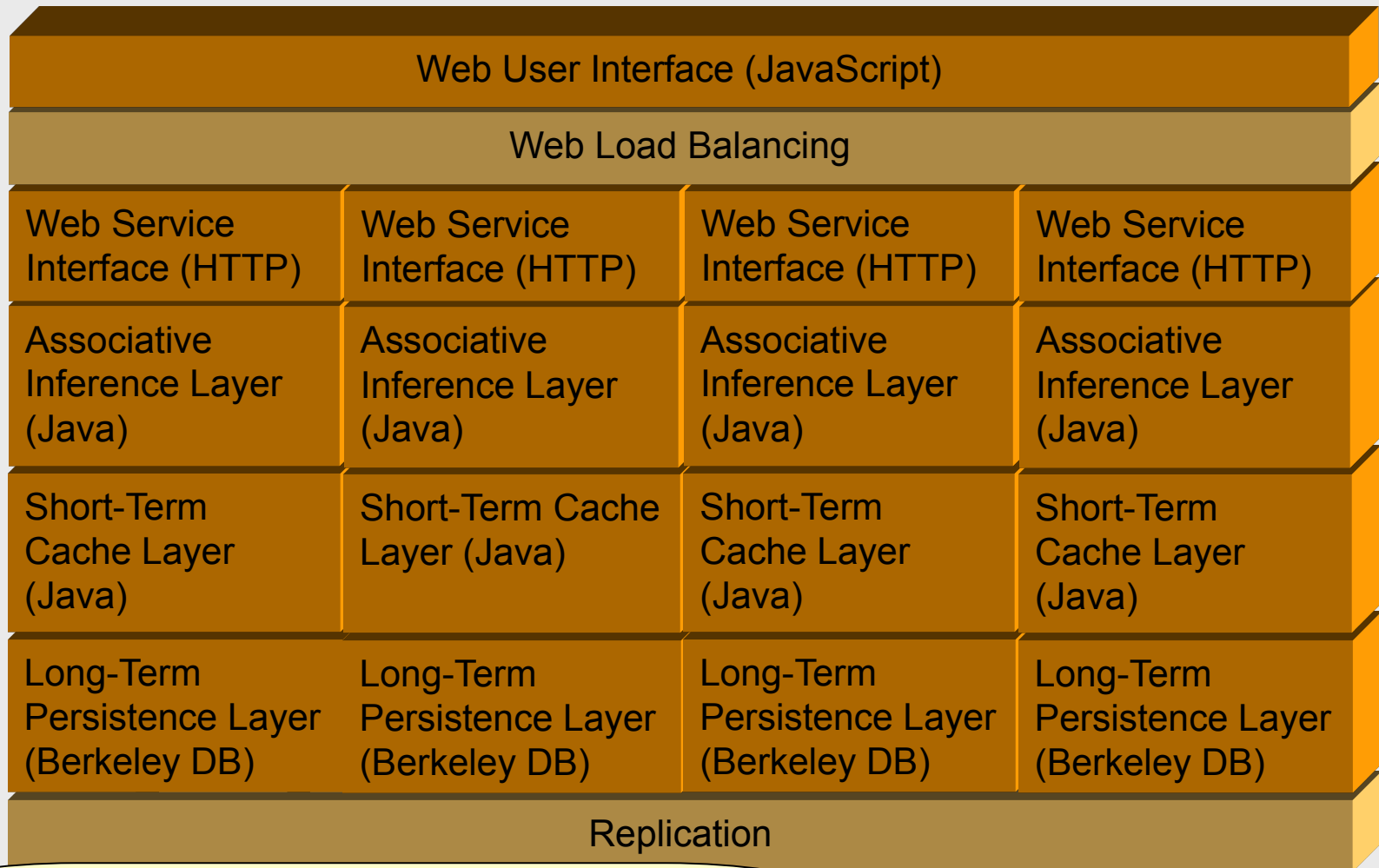
Long-Term Memory

Represents synergetic combination of modern web development technologies and advances in artificial intelligence, knowledge management, fuzzy computing, text mining and neurophysiologic simulation.

© Anton Kolonin 2008

Webstructor Mine: Platform Architecture

Web User Interface (JavaScript)

Third-Party

Third-Party

Third-Party

Third-Party

Third-Party

Third-Party

Web Service Interface (HTTP)

Third-Party

Authentication and Administration Layer (Java)

Third-Party

Associative Inference Layer (Java)

Short-Term Cache Layer (Java)

Long-Term Persistence Layer (Berkeley DB)
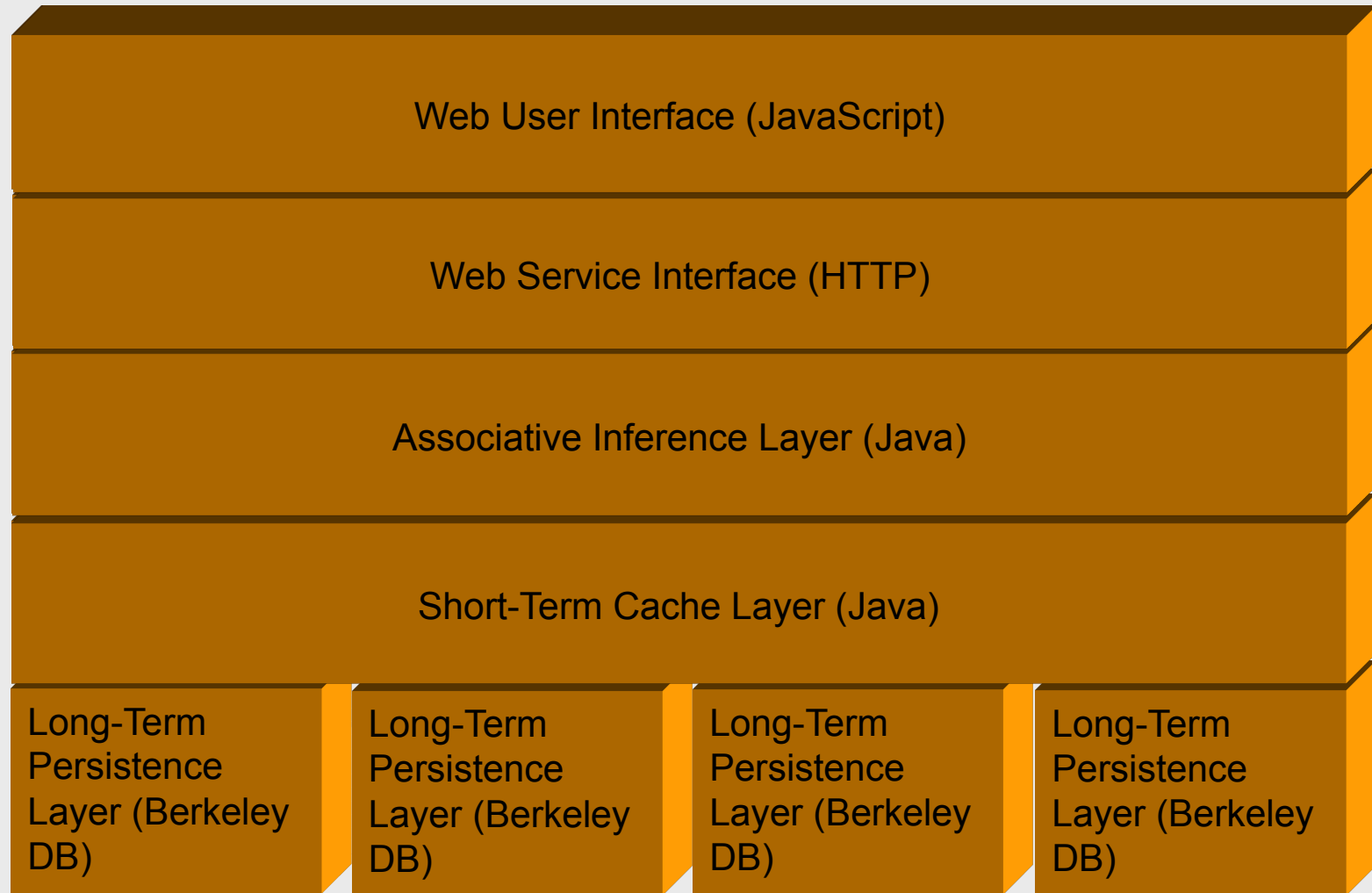
Flat Import/Export Format

The employed platform architecture involves four isolated layers and enables seamless integration of the technology with and into any third-party solutions at any layer.

Webstructor Mine: Clustering Approach

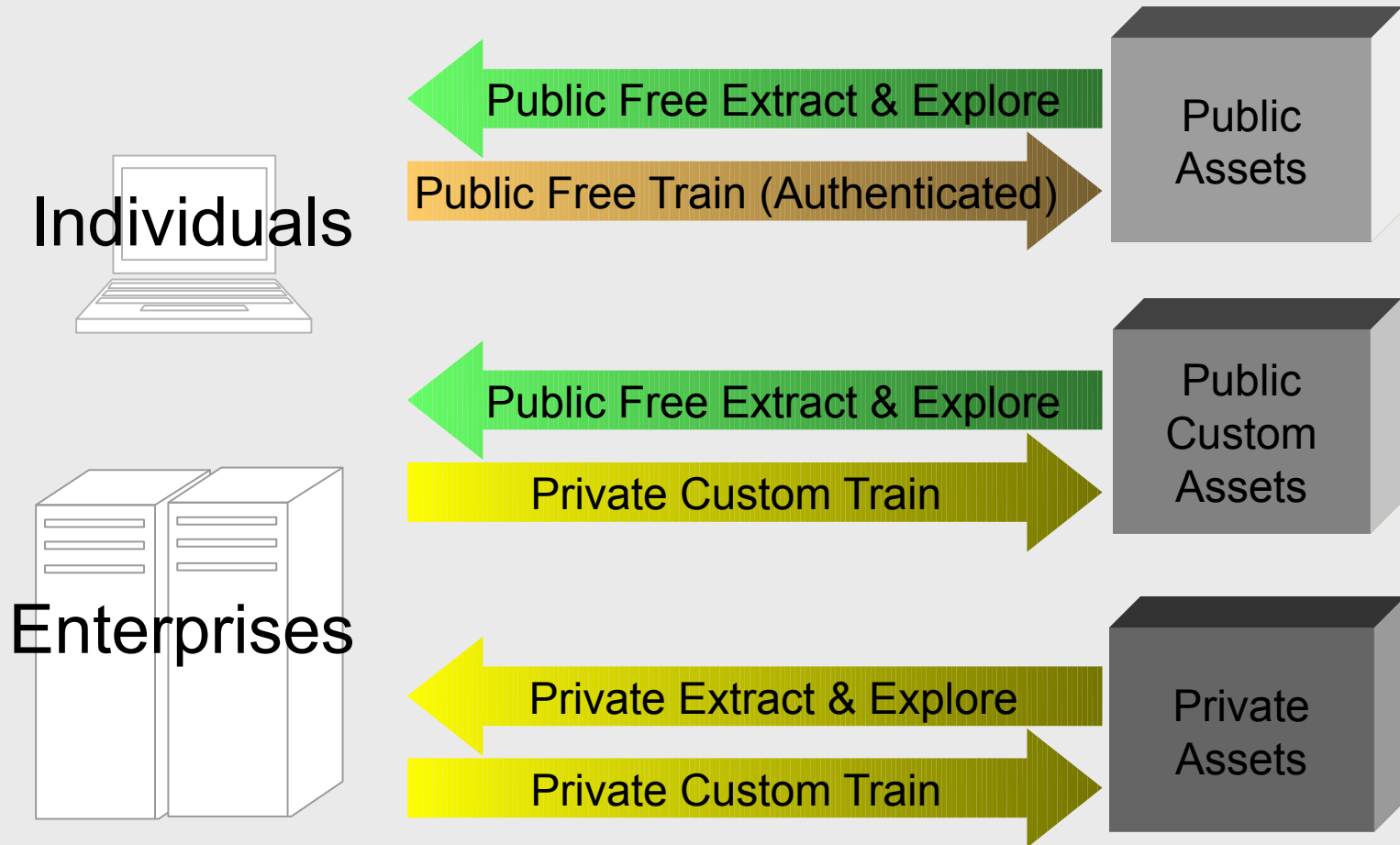| Web User Interface (JavaScript)			
Web Load Balancing			
Web Service Interface (HTTP)	Web Service Interface (HTTP)	Web Service Interface (HTTP)	Web Service Interface (HTTP)
Associative Inference Layer (Java)	Associative Inference Layer (Java)	Associative Inference Layer (Java)	Associative Inference Layer (Java)
Short-Term Cache Layer (Java)	Short-Term Cache Layer (Java)	Short-Term Cache Layer (Java)	Short-Term Cache Layer (Java)
Long-Term Persistence Layer (Berkeley DB)	Long-Term Persistence Layer (Berkeley DB)	Long-Term Persistence Layer (Berkeley DB)	Long-Term Persistence Layer (Berkeley DB)
Replication			

Industry-compliant clustering approach enables almost linear scalability of online solution based on web-level load balancing supported with low-level replication of the underlying data.

© Anton Kolonin 2008   9

Webstructor Mine: Partitioning Approach

Web User Interface (JavaScript)

Web Service Interface (HTTP)

Associative Inference Layer (Java)

Short-Term Cache Layer (Java)

Long-Term Persistence Layer (Berkeley DB)

Long-Term Persistence Layer (Berkeley DB)

Long-Term Persistence Layer (Berkeley DB)

Long-Term Persistence Layer (Berkeley DB)

Low-level data partitioning technology makes it possible to face the problem with enormous data sets appearing once full web-scale document sets are involved in operation.

Webstructor Mine: Business Models

Individuals

Public Free Extract & Explore ◄

Public Free Train (Authenticated) ►

Public Assets

Public Free Extract & Explore ◄

Private Custom Train ►

Public Custom Assets

Enterprises

Private Extract & Explore ◄

Private Custom Train ►

Private Assets

Several business models for usage of the service and solution make it cheap-to-free for both enterprises and individuals to use them for their specific and proprietary needs.

Thank you for your time!
Visit us at

http://www.webstructor.net/mine

E-mail us to
contact@webstructor.net