Microsoft Research
微软亚洲研究院

代码搜索技术的研究和实践

张洪宇
(Hongyu Zhang)

微软亚洲研究院 (Microsoft Research, Asia)

2016.7

Programming is hard…

- Programming is largely a manual effort, which is tedious and time consuming.
- Developers often wonder what code to write in order to complete a programming task
  - Unfamiliar problems
  - Unfamiliar APIs
  - …

- In a survey conducted at Microsoft, 67.6% respondents mentioned that they encountered problems in learning APIs.
  (M.P. Robillard, “What makes APIs hard to learn? Answers from developers,” IEEE Softw., vol. 26, no. 6, pp. 27–34, 2009)

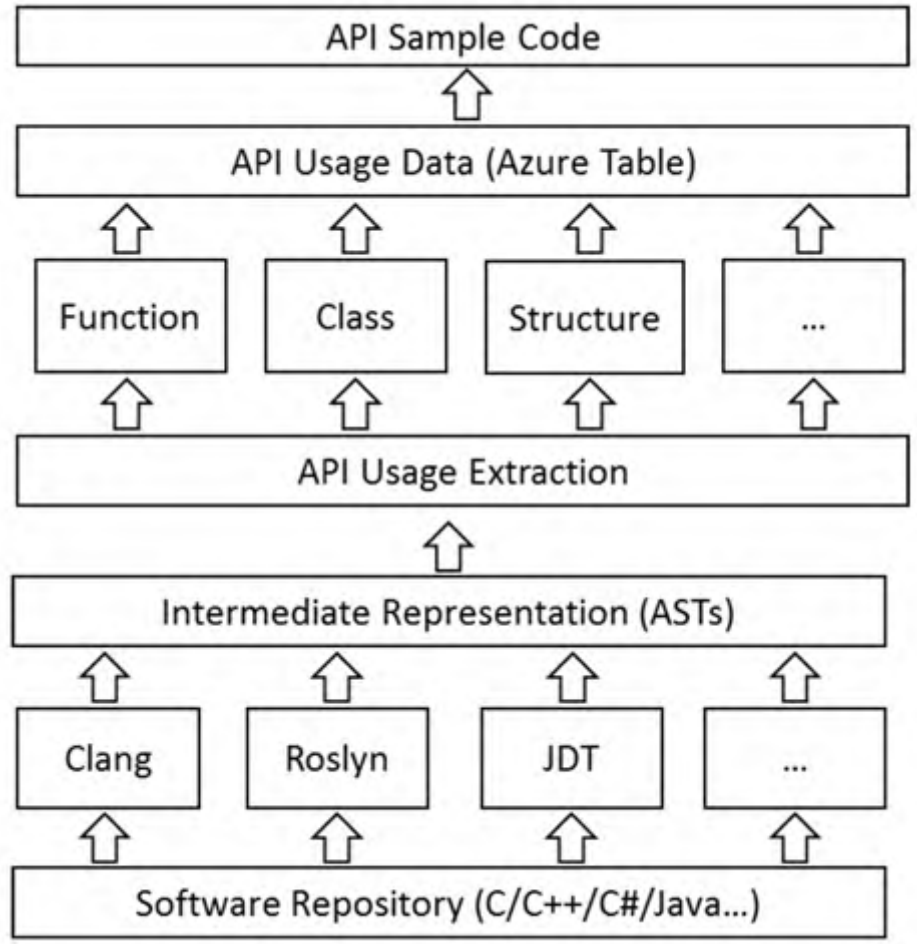
- Vision: *Improving Developer Productivity by Code Search*

Searching for API Usage

Question: how to reuse an API?

Our work: Recommending sample code about the API usage

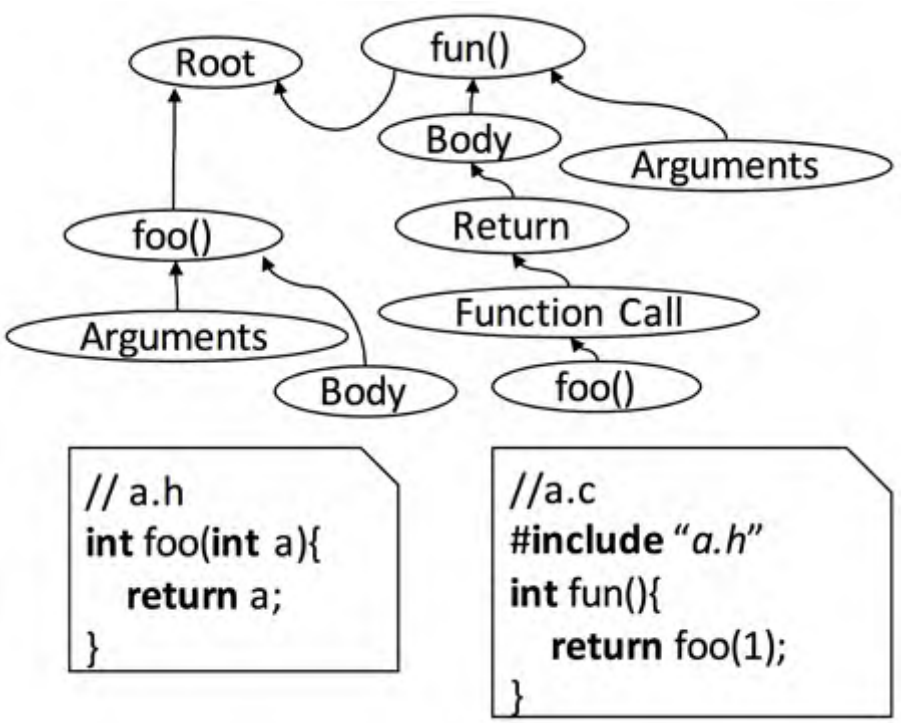
- Extract API usage information from a software repository
- Support C/C++/C#/Java/JavaScript languages.
- Support native and third-party APIs.
- Support a variety of data sources (including GitHub, MSDN, and Microsoft initiatives like Office APIs, Azure IOT, Universal Windows Platform, etc.)



Hongyu Zhang, Anuj Jain, Gaurav Khandelwal, Chandrashekhar Kaushik, Scott Ge and Wenxiang Hu, Bing Developer Assistant – Improving Developer Productivity by Recommending Sample Code, To appear: *FSE 2016*, industry track, Seattle, WA, USA, November 2016.

Static analysis of a program:

- Parse the source code and header files into an AST
- Performs an in-order traversal of the AST
- Process each node according to its type
- Challenges:
  - Third-party APIs
  - Type inference for weakly typed languages



An Example of an AST

Algorithm 1 Analysis of C/C++ files

```
1:  S ← One Source File
2:  H ← Head files
3:  AST ← ∅
4:
5:  function MAIN
6:      Sourcefile ← S
7:      Headfiles ← H
8:      AST ← PARSER(Sourcefile, Headfiles)
9:      for Node ∈ AST root's Children do
10:         VISIT(Node)
11:     end for
12: end function
13:
14: function VISIT(Node)
15:     if Node ∈ S then
16:         switch Node do
17:             case FunctionCall
18:                 PROCESS_FUNCALL(Node)
19:             case VariableDeclaration
20:                 PROCESS_VARDECL(Node)
21:             case Otherkindsofnodes
22:                 ...
23:         for Child ∈ Node's Children do
24:             VISIT(Child)
25:         end for
26:     end if
27: end function
28:
29: function PROCESS_FUNCALL(Node)
30:     Callee ← EXTRACT_CALLEE(Node)
31:     Decl ← FIND_DECLARATION(Callee)
32:     if Decl ≠ ∅ then
33:         File = WHERE_IS_FROM(Decl)
34:         if File ≠ S then
35:             LibName =WHICH_LIBRARY(File)
36:             LocationInfo = GET_LOCATIONINFO(Node)
37:             SAVE_DATA(Callee, LocationInfo, LibName)
38:         end if
39:     end if
40: end function
```

Extracted API Usage Data

- We construct a large-scale codebase by crawling projects from MSDN and GitHub： 65,253 projects, 437 GB, 3.5 million source code files.
- In total, 78 million code snippets are obtained.
- The API usage data and sample code are stored in and managed by Microsoft Azure Table.

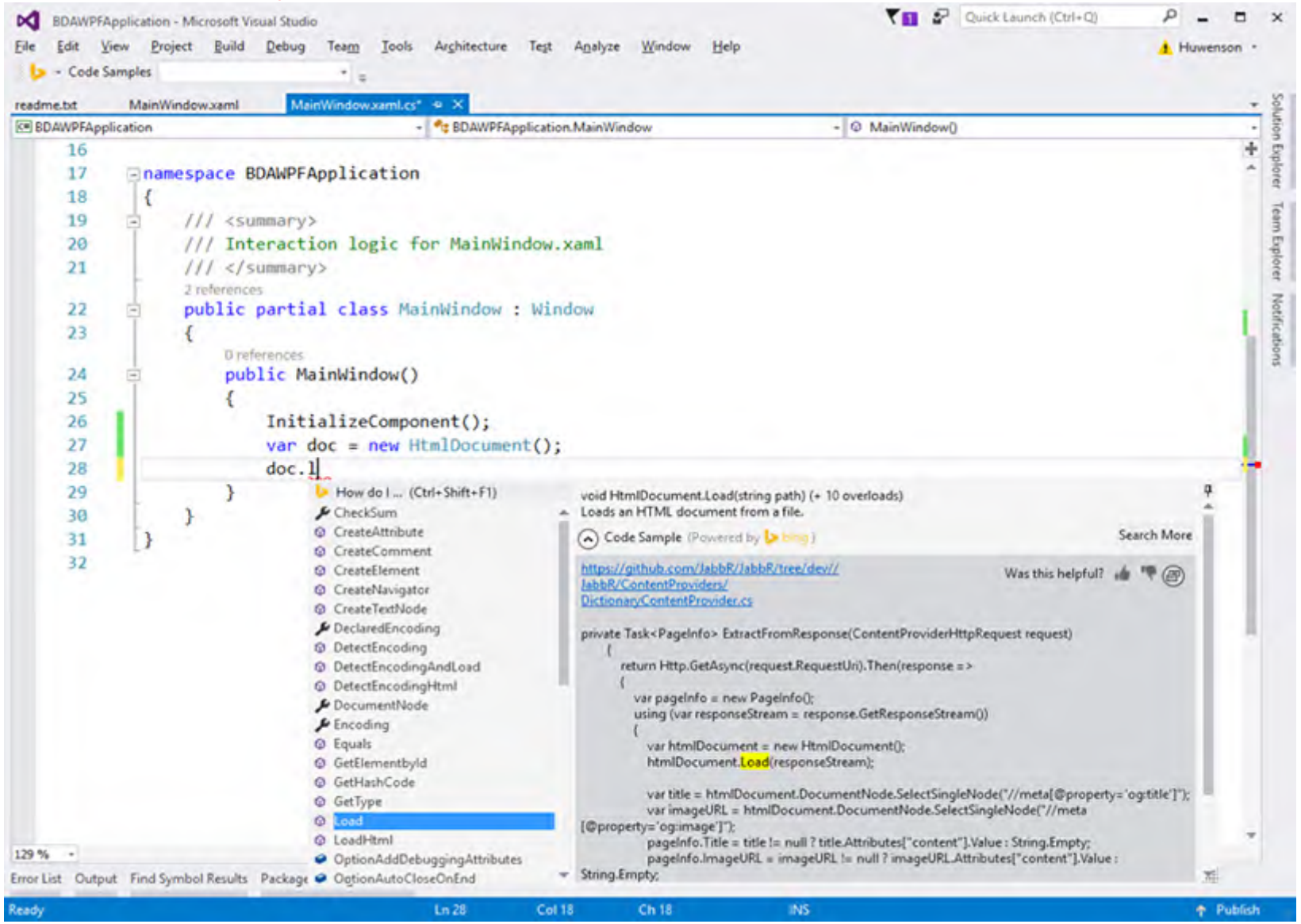
Language	#Projects	#Files	#Unique APIs	#Code Snippets
C/C++	12,162	730,084	2,980,543	43,944,234
C#	26,322	907,632	698,651	12,336,251
JavaScript	17,115	453,449	21,280	4,205,998
Java	9,654	1,398,695	68,958	17,679,077
Total	65,253	3,489,860	3,769,432	78,165,560

The statistics of the obtained API usage data

Bing Developer Assistant (BDA)

http://aka.ms/devassistant

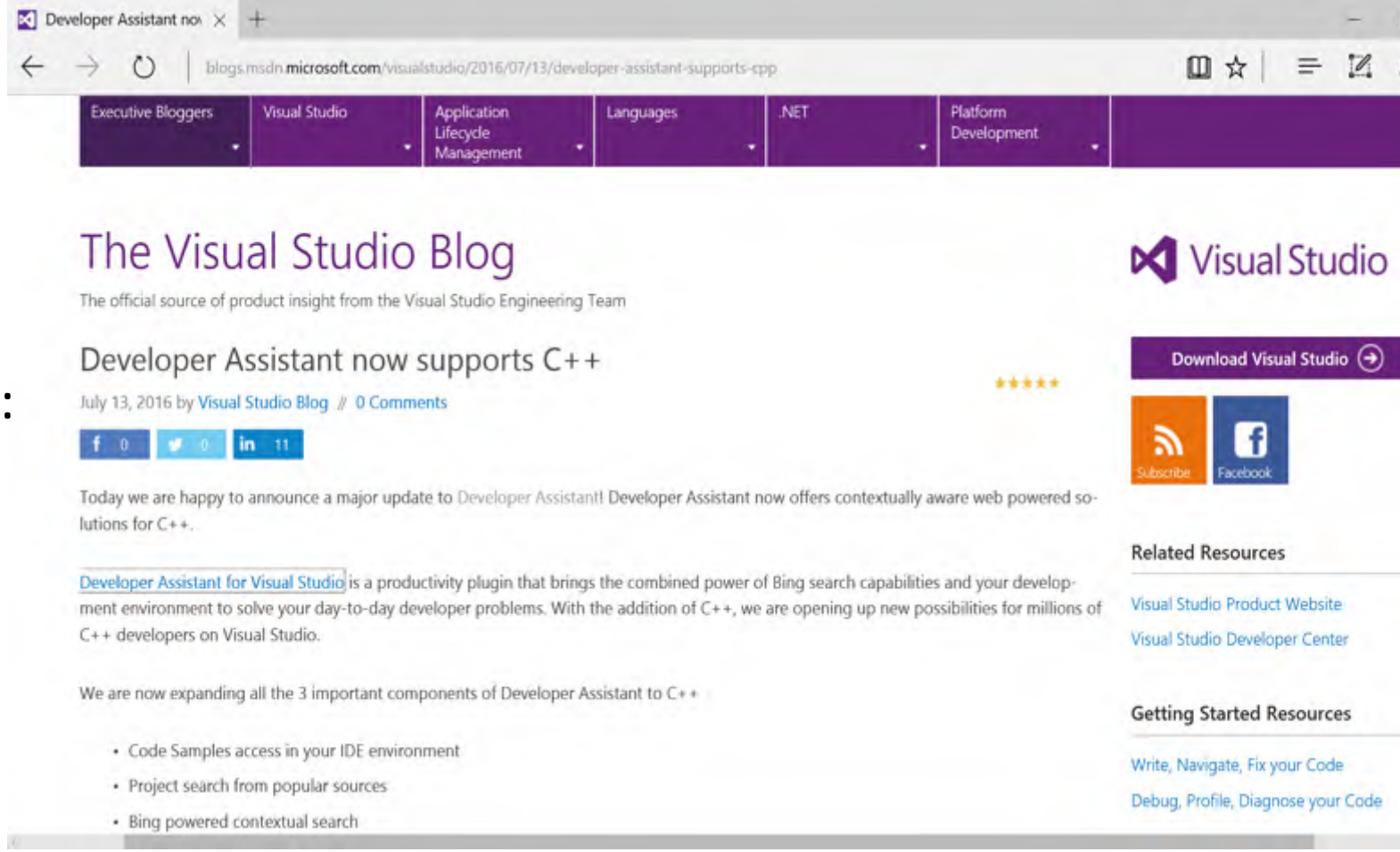
- Client side: an extension of Microsoft Visual Studio.
- Server side: Microsoft Azure servers located around the world.
- The APIs provided by IntelliSense of VS trigger the backend BDA service.
- The returned sample code is displayed within Visual Studio.



Technology Transfer

- Transferred to Bing Developer Assistant (BDA) in 2015.
- Status Updates (as of July 2016):
  - Received more than 450K downloads
  - ~2.1 million queries per month
  - Build 2016 presence
  - New release of BDA on July 13, 2016 (with C/C++ support)

Available at: http://aka.ms/devassistant



Visual Studio Blog Visual C++ Team Blog

Collaborators:

Bing Tech Segment
Hyderabad

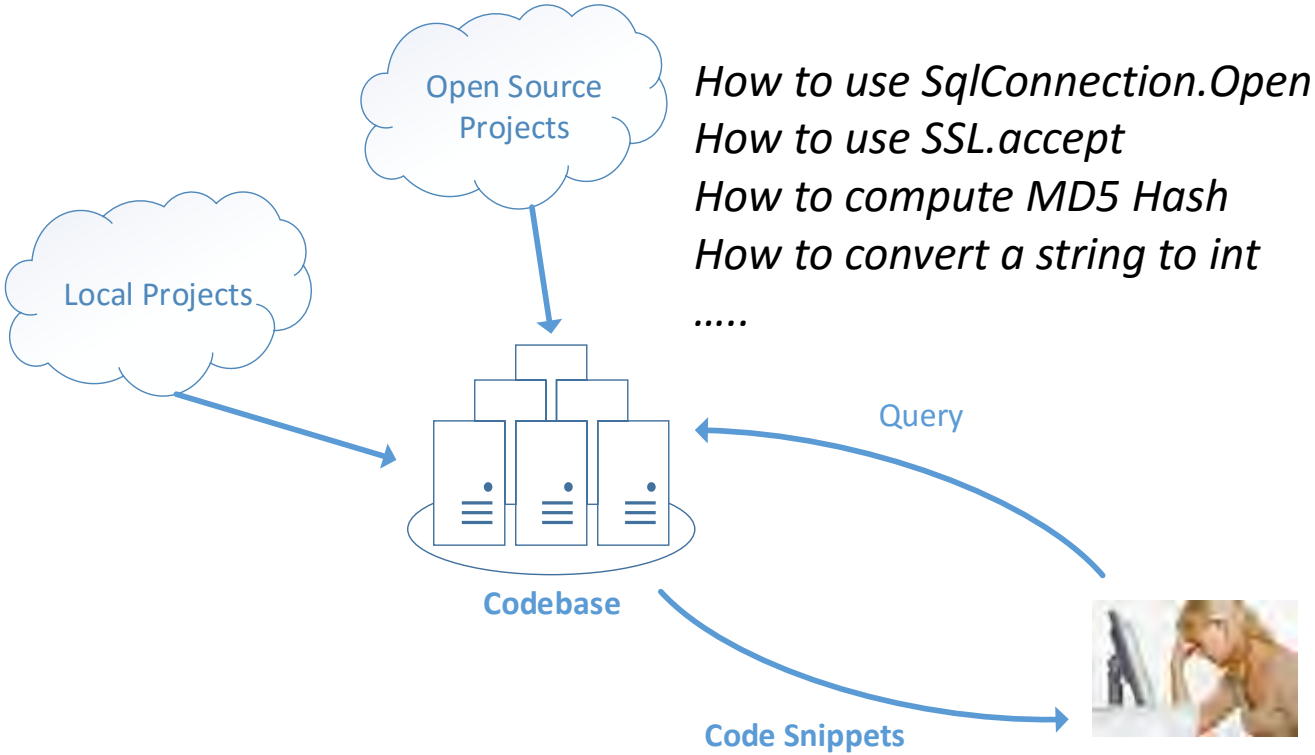
DX Audience Online Engines
Redmond

Visual Studio
Redmond

Searching for Reusable Code

We propose **CodeHow**, a method for searching reusable code based on free-form queries

- Given a user query, CodeHow searches codebases and returns the relevant code snippets that match the query.
- Consider both text similarity and program semantics in code retrieval
- Build the tool on top of ElasticSearch and Microsoft Azure
- Indexed ~50K C#/Java/C++/VB projects collected from GitHub and Codeplex



Fei Lv, Hongyu Zhang, Jian-guang Lou, Shaowei Wang, Dongmei Zhang, and Jianjun Zhao, "CodeHow: Effective Code Search based on API Understanding and Extended Boolean Model", in *Proc. ASE 2015*, Lincoln, Nebraska, Nov 2015.

Natural Language Query - Query Formulation

- Text Similarity Query q_{text}
  - For retrieving code snippets that match the query in terms of text similarity

- API Query q_{api_i}
  - Obtain the top *k* APIs that are potentially relevant to the query
  - For retrieving code snippets that contain the APIs

- Expanded Query: incorporate both API similarity and text similarity

$$q_{expand} = (q_{api_1}, q_{api_2}, \cdots, q_{api_k}, q_{text})$$

Natural Language Query - Retrieval

- Retrieve code snippets based on the similarity between the expanded query q_{expand} and a code snippet d:

$$sim(q_{expand}, d) = \sum_{i=1}^{k} sim(q_{api_i}, d) + sim(q_{text}, d)$$

The similarity value is computed using Extended Boolean model:

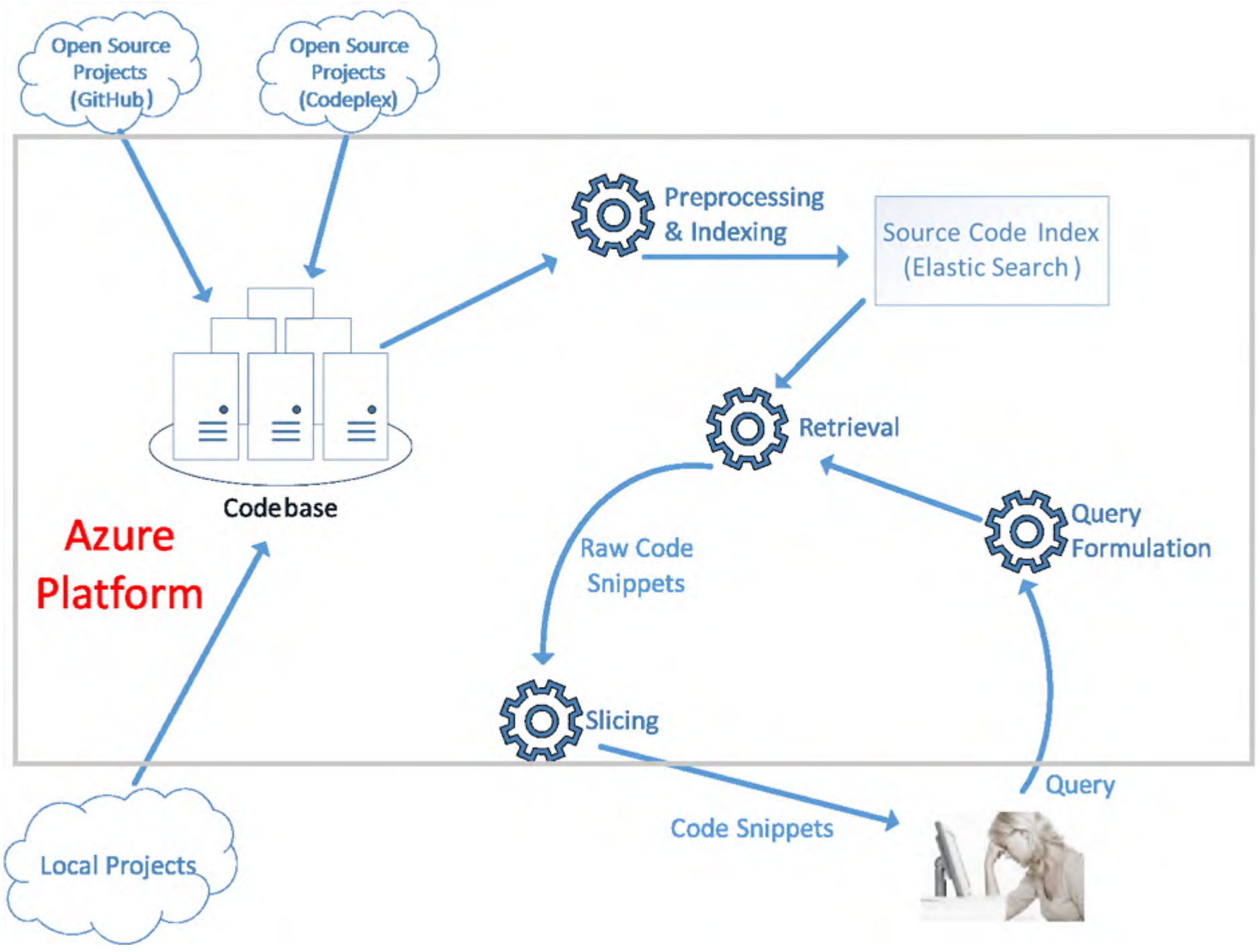
$$sim(q_{or}, d) = \left(\frac{w_{t_1,q}^p w_{t_1,d}^p + w_{t_2,q}^p w_{t_2,d}^p + \ldots + w_{t_n,q}^p w_{t_n,d}^p}{w_{t_1,q}^p + w_{t_2,q}^p + \ldots + w_{t_n,q}^p}\right)^{\frac{1}{p}}$$

$$sim(q_{and}, d) = 1 - \left(\frac{w_{t_1,q}^p (1 - w_{t_1,d})^p + \ldots + w_{t_n,q}^p (1 - w_{t_n,d})^p}{w_{t_1,q}^p + w_{t_2,q}^p + \ldots + w_{t_n,q}^p}\right)^{\frac{1}{p}}$$

Fei Lv, Hongyu Zhang, Jian-guang Lou, Shaowei Wang, Dongmei Zhang, and Jianjun Zhao, "CodeHow: Effective Code Search based on API Understanding and Extended Boolean Model", *Proc. ASE'15*, Lincoln, USA, Nov 2015.

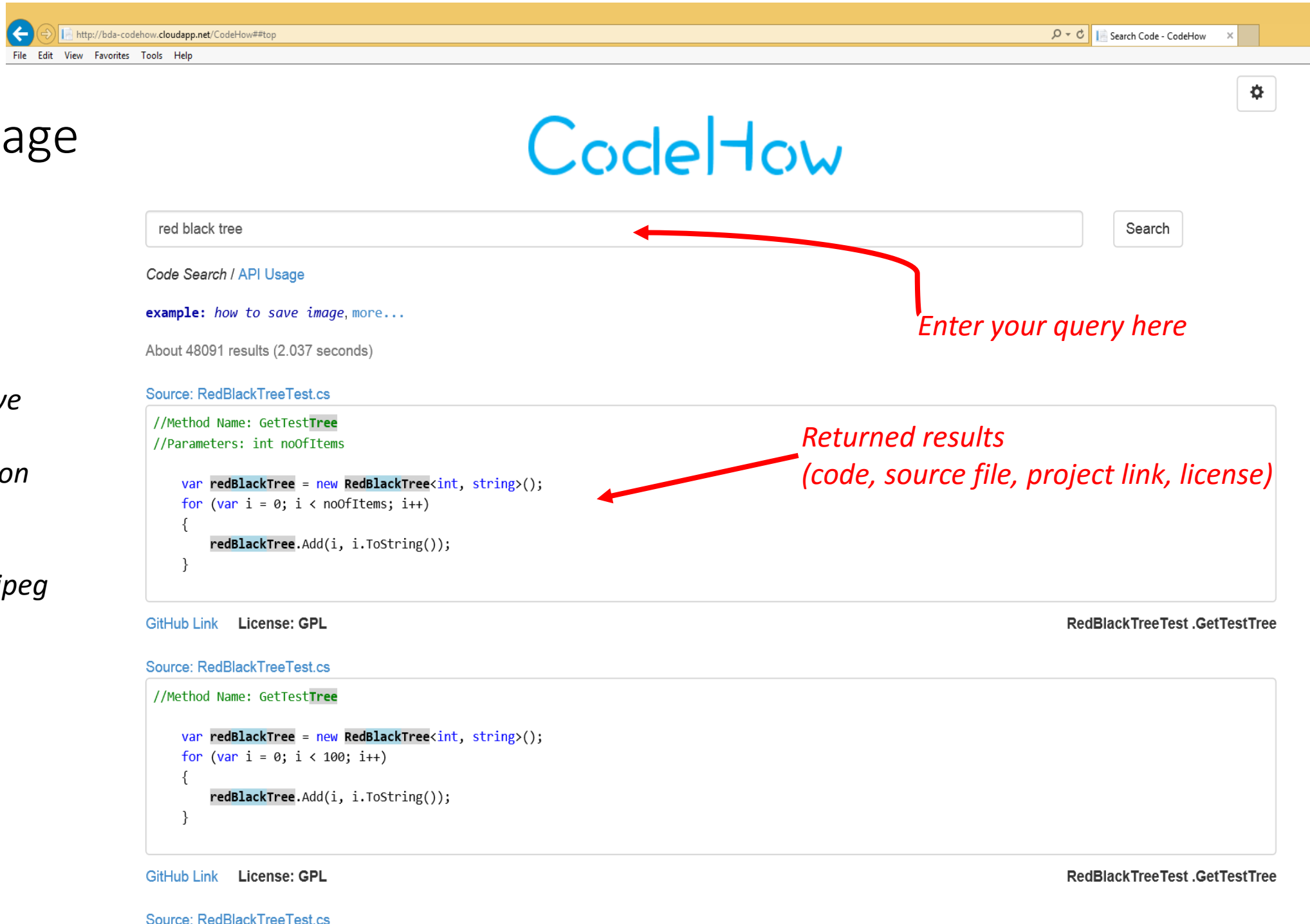
System Structure

- ElasticSearch (1 master and 4 workers)
- Azure platform



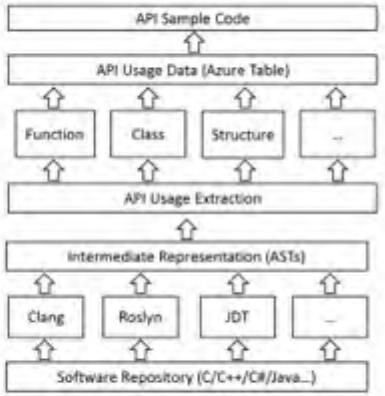
Natural Language Query

Query Examples:
red black tree
get free space for disk drive
parse an xml document
compute standard deviation
quick sort
MD5 hash
how to save an image as jpeg
how to export to excel



Searching for API Usage

- Question: how to use an API?
- Our work: Recommending sample code about the API usage
  - Extract API usage information from a software repository
  - Support C/C++/C#/Java/JavaScript languages.
  - Support native and third-party APIs.
  - Support a variety of data sources (including GitHub, MSDN, and Microsoft initiatives like Office APIs, Azure IOT, Universal Windows Platform, etc.)



Technology Transfer

- Transferred to Bing Developer Assistant (BDA)
- Status Updates (as of July 2016):
  - Received more than 450K downloads
  - ~2.1 million queries per month
  - Build 2016 presence
  - New release of BDA on July 13, 2016

Available at: http://aka.ms/devassistant

https://blogs.msdn.microsoft.com/visualstudio/2016/07/13/developer-assistant-supports-cpp/

Collaborators:
Bing Tech Segment
Hyderabad

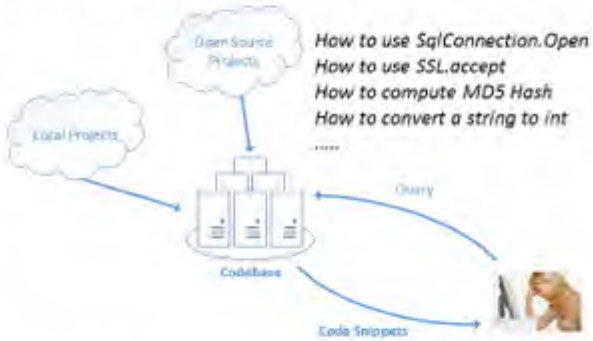
DX Audience Online Engines
Redmond

Visual Studio
Redmond

Searching for Reusable Code

We propose **CodeHow**, a method for searching reusable code based on free-form queries

- Given a user query, CodeHow searches codebases and returns the relevant code snippets that match the query.
- Consider both text similarity and program semantics in code retrieval
- Build the tool on top of ElasticSearch and Microsoft Azure
- Indexed ~50K C#/Java/C++/VB projects collected from GitHub and Codeplex



Fei Lv, Hongyu Zhang, Jian-guang Lou, Shaowei Wang, Dongmei Zhang, and Jianjun Zhao, "CodeHow: Effective Code Search based on API Understanding and Extended Boolean Model", in *Proc. ASE 2015*, Lincoln, Nebraska, Nov 2015.

Natural Language Query

Query Examples:
red block tree
get free space for disk drive
parse an xml document
compute standard deviation
quick sort
MD5 hash
how to save an image as jpeg
how to export to excel

Microsoft Research
微软亚洲研究院

Thanks!

Hongyu Zhang
Microsoft Research
No.5 Danling Street
Beijing 100080, China
Email: honzhang@microsoft.com
URL: http://hongyujohn.github.io/