LaSEWeb: Automating Search Strategies over Semi-Structured Web Data

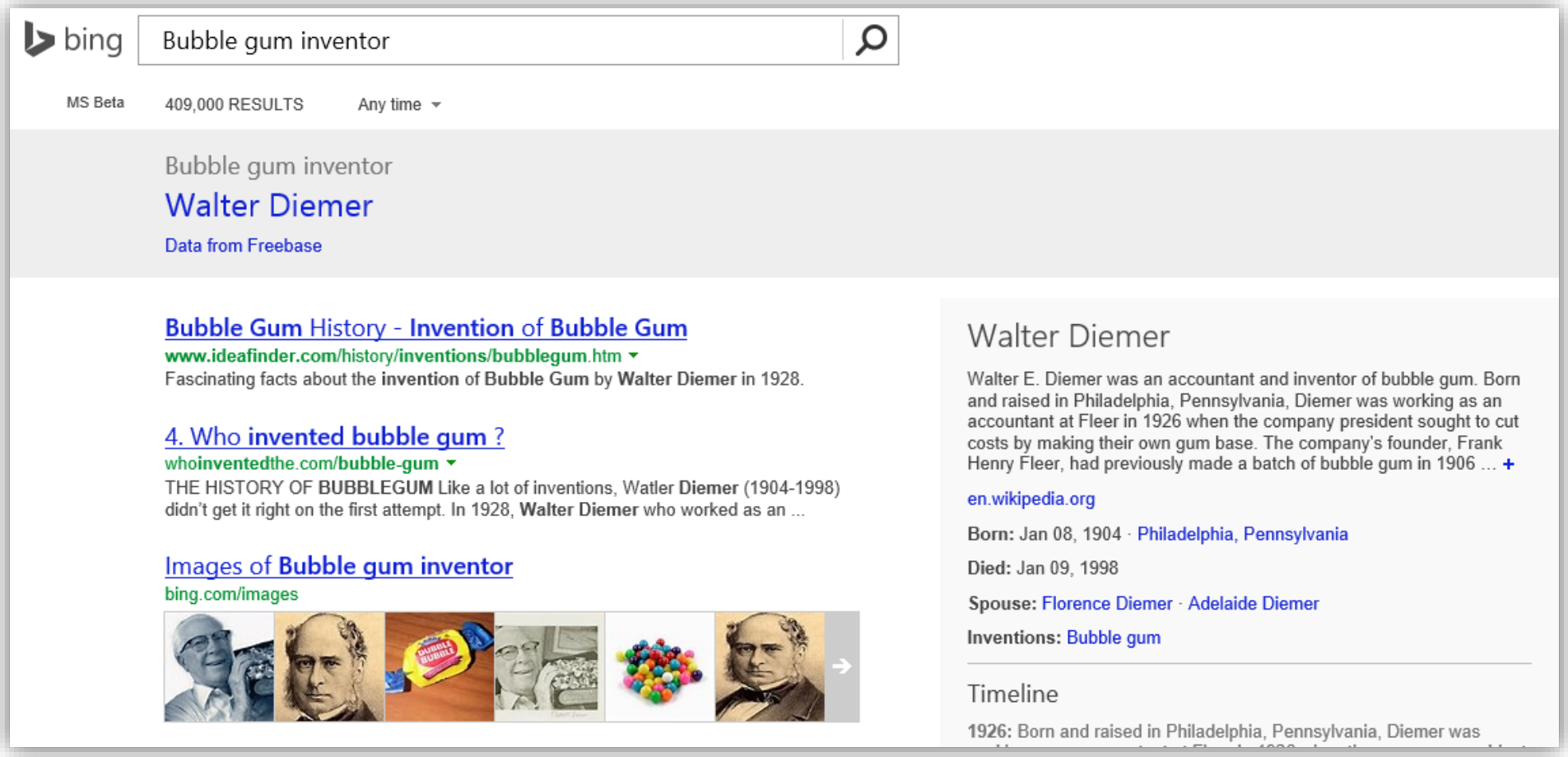
Oleksandr Polozov
University of Washington
polozov@cs.washington.edu

Sumit Gulwani
Microsoft Research
sumitg@microsoft.com

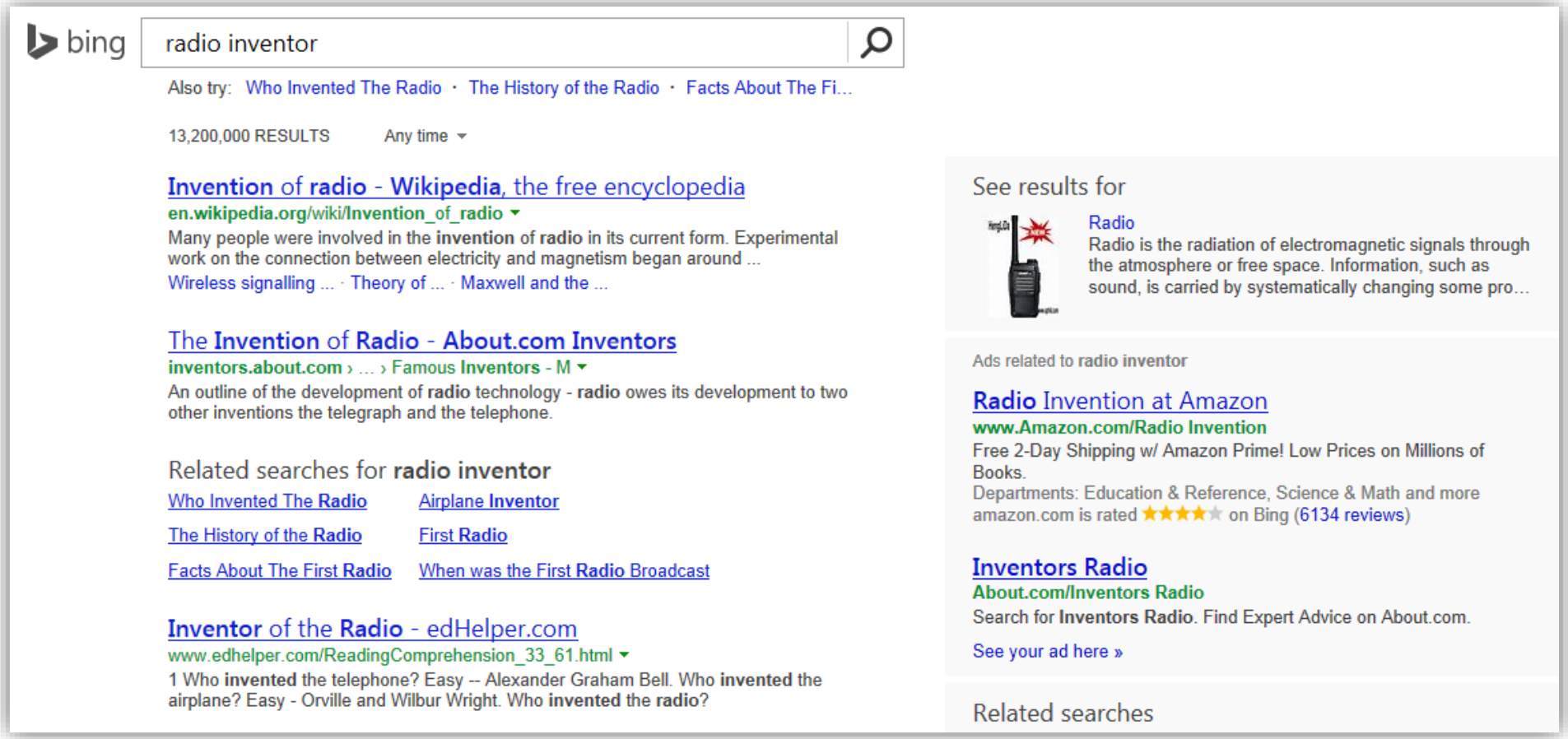
Microsoft® Research

KDD 2014 — August 27, 2014

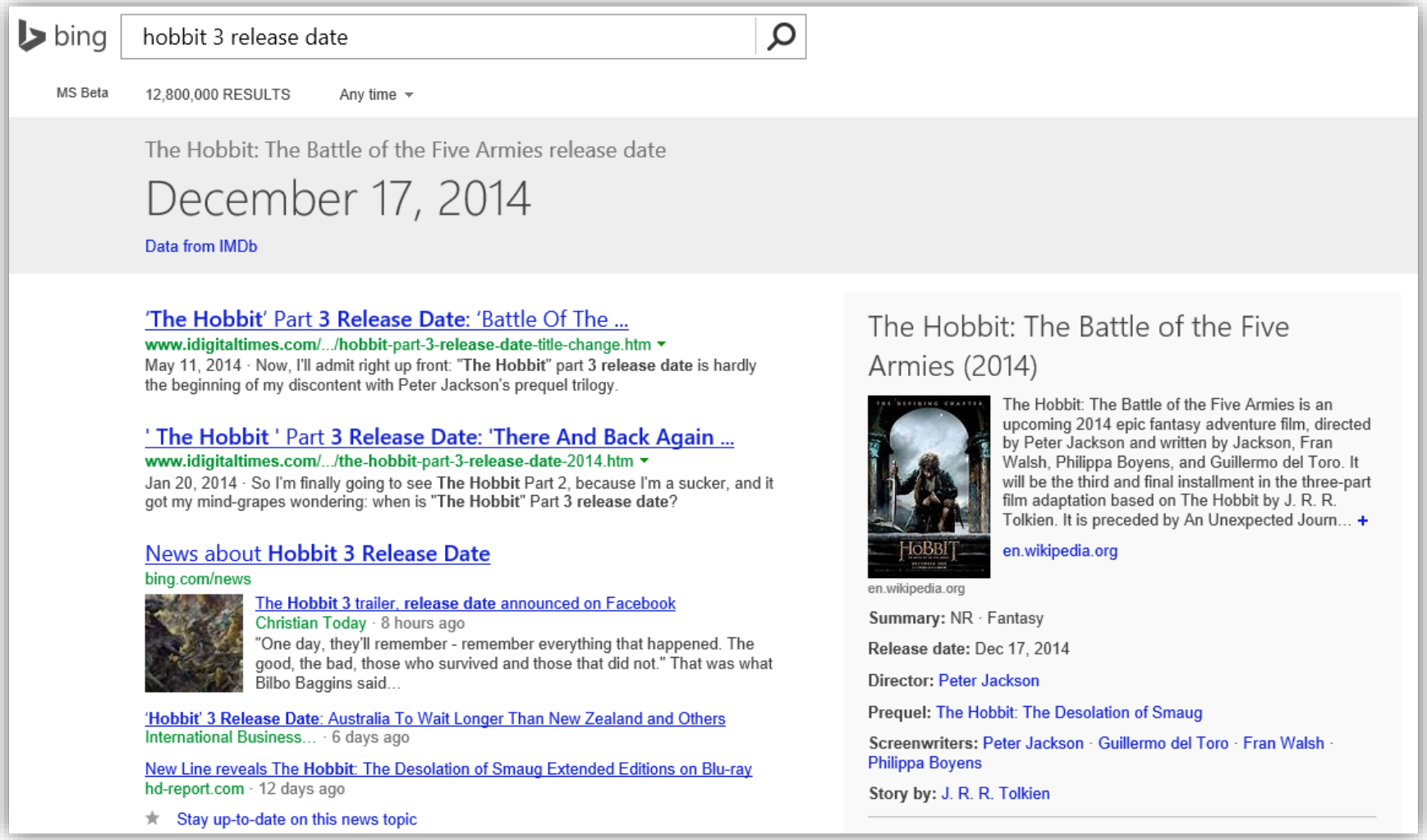
Motivation: search engine micro-segments



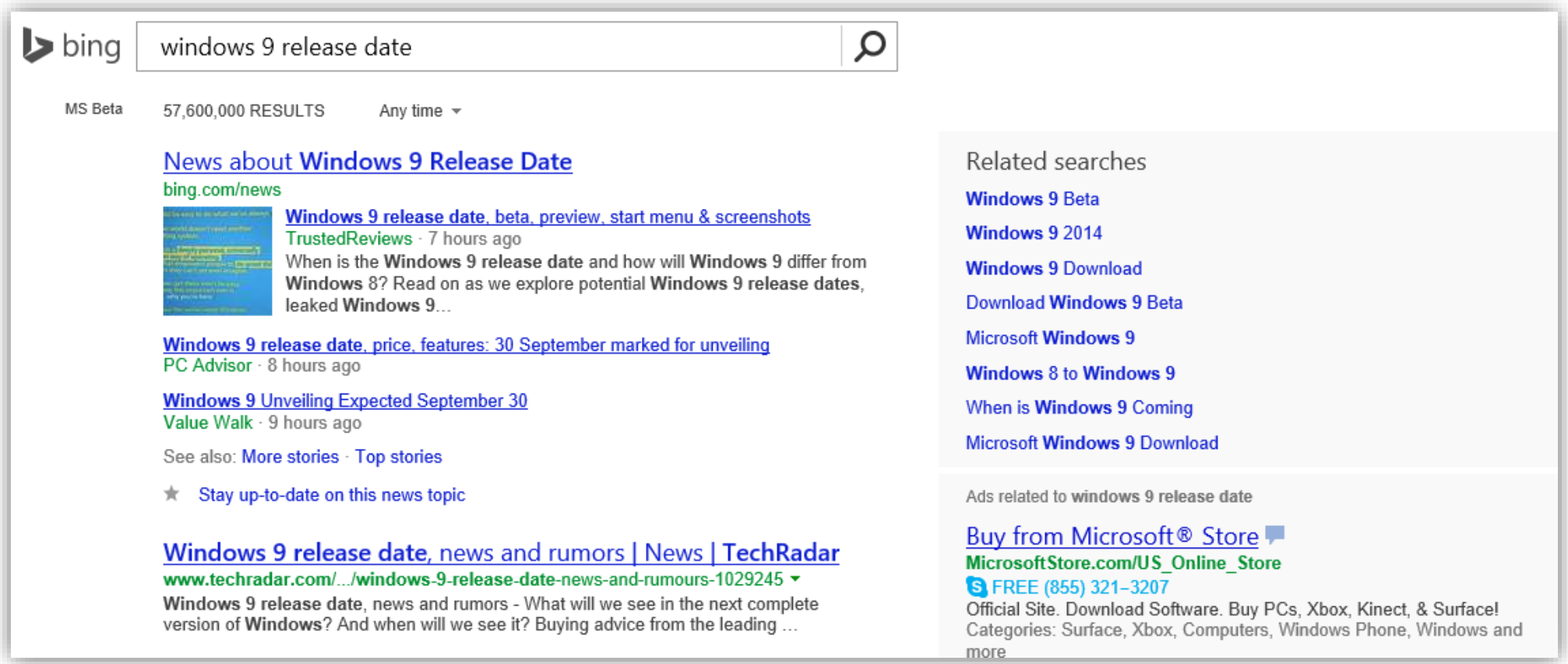
Motivation: search engine micro-segments



Motivation: search engine micro-segments



Motivation: search engine micro-segments



Repetitive search tasks

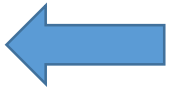
Structured databases

- Precise, but limited in content
- No time-sensitive information
- Provide no context (sources)

Repetitive search tasks

Structured databases

- Precise, but limited in content
- No time-sensitive information
- Provide no context (sources)



Web mining scripts

- Two extremes:
  - Powerful ML, which has to be re-learned for each micro-segment
  - Fragile HTML layout parser
- Inaccessible for end-users

LaSEWeb Query Language

- A semantic scripting language for semi-structural information extraction from the Web
- Models natural patterns from the humans' search strategies

LaSEWeb interpreter

- Explores multiple webpages, clusters different answer candidates, and provides context for each answer
- Makes use of state-of-the-art NLP/ML/PL algorithms

Example: phone number

Sumit Gulwani
Primary Affiliation: Senior Researcher @ Microsoft Research (Redmond Lab)
Other Affiliations: Adjunct Faculty @ IIT Kanpur
Affiliate Faculty @ Univ. of Washington

Address: Microsoft Corporation, One Microsoft Way, Redmond, WA, 98052
Email: sumitg at microsoft dot com
Phone: (425) 706-7709
Fax: (425) 936 7329 (attn: sumitg)

$\vec{v}$ = (“Sumit Gulwani”)

let $\eta_t = Emphasized(v_1)$ **in**
let $\eta_b = AttributeLookup(Syn(\text{"phone"}), \ell_a)$ **in**
$Union(\eta_t, \eta_b)$
where $Regex(\ell_a, \text{"(\d+)\W*\d+\W*\d+"})$
where $Layout(\eta_t, \eta_b, \text{Down})$ **and** $Nearby(\eta_t, \eta_b)$

Example: phone number

Sumit Gulwani
Primary Affiliation: Senior Researcher @ Microsoft Research (Redmond Lab)
Other Affiliations: Adjunct Faculty @ IIT Kanpur
Affiliate Faculty @ Univ. of Washington

Address: Microsoft Corporation, One Microsoft Way, Redmond, WA, 98052
Email: sumitg at microsoft dot com
Phone: (425) 706-7709
Fax: (425) 936 7329 (attn: sumitg)

$\vec{v}$ = (“Sumit Gulwani”)

let $\eta_t = Emphasized(v_1)$ **in**
let $\eta_b = AttributeLookup(Syn(\text{"phone"}), \ell_a)$ **in**
$Union(\eta_t, \eta_b)$
where $Regex(\ell_a, \text{"(\d+)\W*\d+\W*\d+"})$
where $Layout(\eta_t, \eta_b, \text{Down})$ **and** $Nearby(\eta_t, \eta_b)$

- Visual attributes

Example: phone number

Sumit Gulwani

Primary Affiliation: Senior Researcher @ Microsoft Research (Redmond Lab)
Other Affiliations: Adjunct Faculty @ IIT Kanpur
Affiliate Faculty @ Univ. of Washington

Address: Microsoft Corporation, One Microsoft Way, Redmond, WA, 98052
Email: sumitg at microsoft dot com
Phone: (425) 706-7709
Fax: (425) 936 7329 (attn: sumitg)

$\vec{v}$ = ("Sumit Gulwani")

let $\eta_t = Emphasized(v_1)$ **in**
let $\eta_b = AttributeLookup(Syn(\text{"phone"}), \ell_a)$ **in**
$Union(\eta_t, \eta_b)$
where $Regex(\ell_a, \text{"(\d+)\W*\d+\W*\d+"})$
where $Layout(\eta_t, \eta_b, \text{Down})$ **and** $Nearby(\eta_t, \eta_b)$

- Visual attributes
- Implicit table detection

Example: phone number

Sumit Gulwani
Primary Affiliation: Senior Researcher @ Microsoft Research (Redmond Lab)
Other Affiliations: Adjunct Faculty @ IIT Kanpur
Affiliate Faculty @ Univ. of Washington

Address: Microsoft Corporation, One Microsoft Way, Redmond, WA, 98052
Email: sumitg at microsoft dot com
Phone: (425) 706-7709
Fax: (425) 936 7329 (attn: sumitg)

$\vec{v}$ = ("Sumit Gulwani")

let $\eta_t = Emphasized(v_1)$ **in**
let $\eta_b = AttributeLookup(Syn(\text{"phone"}), \ell_a)$ **in**
$Union(\eta_t, \eta_b)$
where $Regex(\ell_a, \text{"(\d+)\W*\d+\W*\d+"})$
where $Layout(\eta_t, \eta_b, \text{Down})$ **and** $Nearby(\eta_t, \eta_b)$

- Visual attributes
- Implicit table detection
- Linguistic patterns

Example: phone number

Sumit Gulwani
Primary Affiliation: Senior Researcher @ Microsoft Research (Redmond Lab)
Other Affiliations: Adjunct Faculty @ IIT Kanpur
Affiliate Faculty @ Univ. of Washington

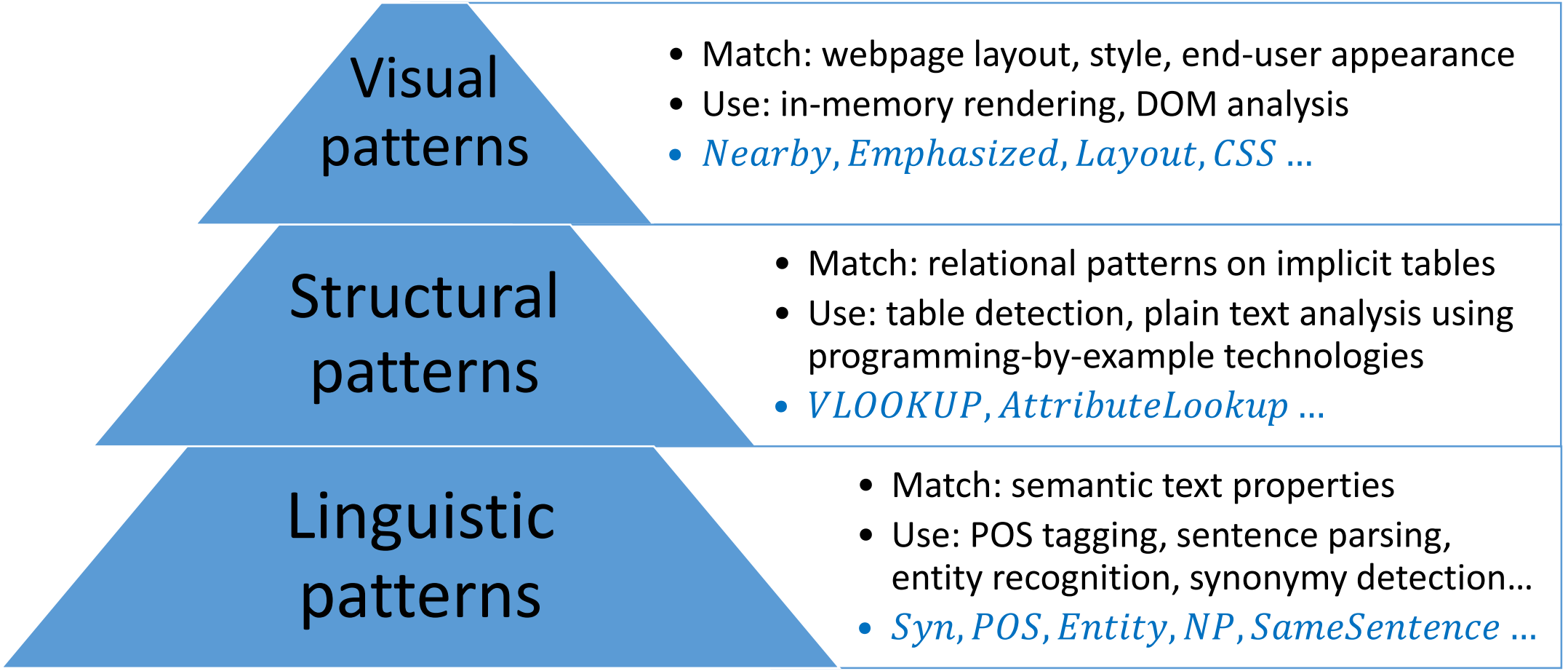
Address: Microsoft Corporation, One Microsoft Way, Redmond, WA, 98052
Email: sumitg at microsoft dot com
Phone: (425) 706-7709
Fax: (425) 936 7329 (attn: sumitg)

$\vec{v}$ = ("Sumit Gulwani")

let $\eta_t = Emphasized(v_1)$ **in**
let $\eta_b = AttributeLookup(Syn(\text{"phone"}), \ell_a)$ **in**
$Union(\eta_t, \eta_b)$
where $Regex(\ell_a, \text{"(\d+)\W*\d+\W*\d+"})$
where $Layout(\eta_t, \eta_b, \text{Down})$ **and** $Nearby(\eta_t, \eta_b)$

- Visual attributes
- Implicit table detection
- Linguistic patterns
- Clustering across webpages

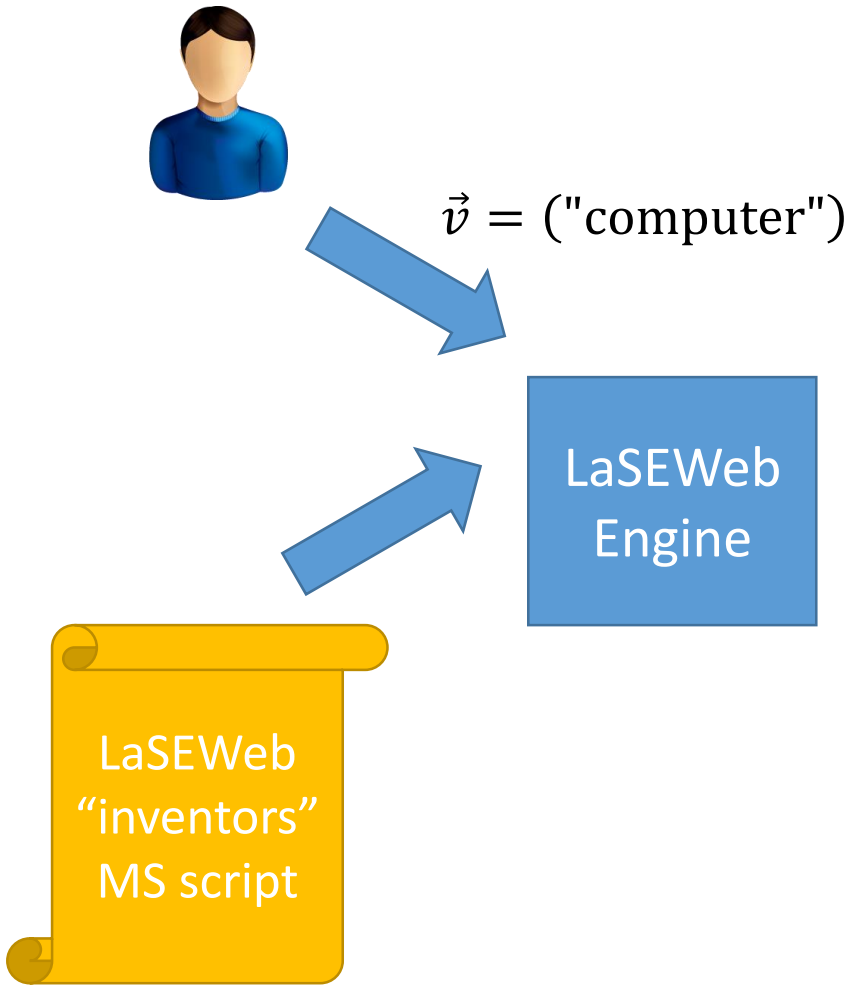
Language Structure



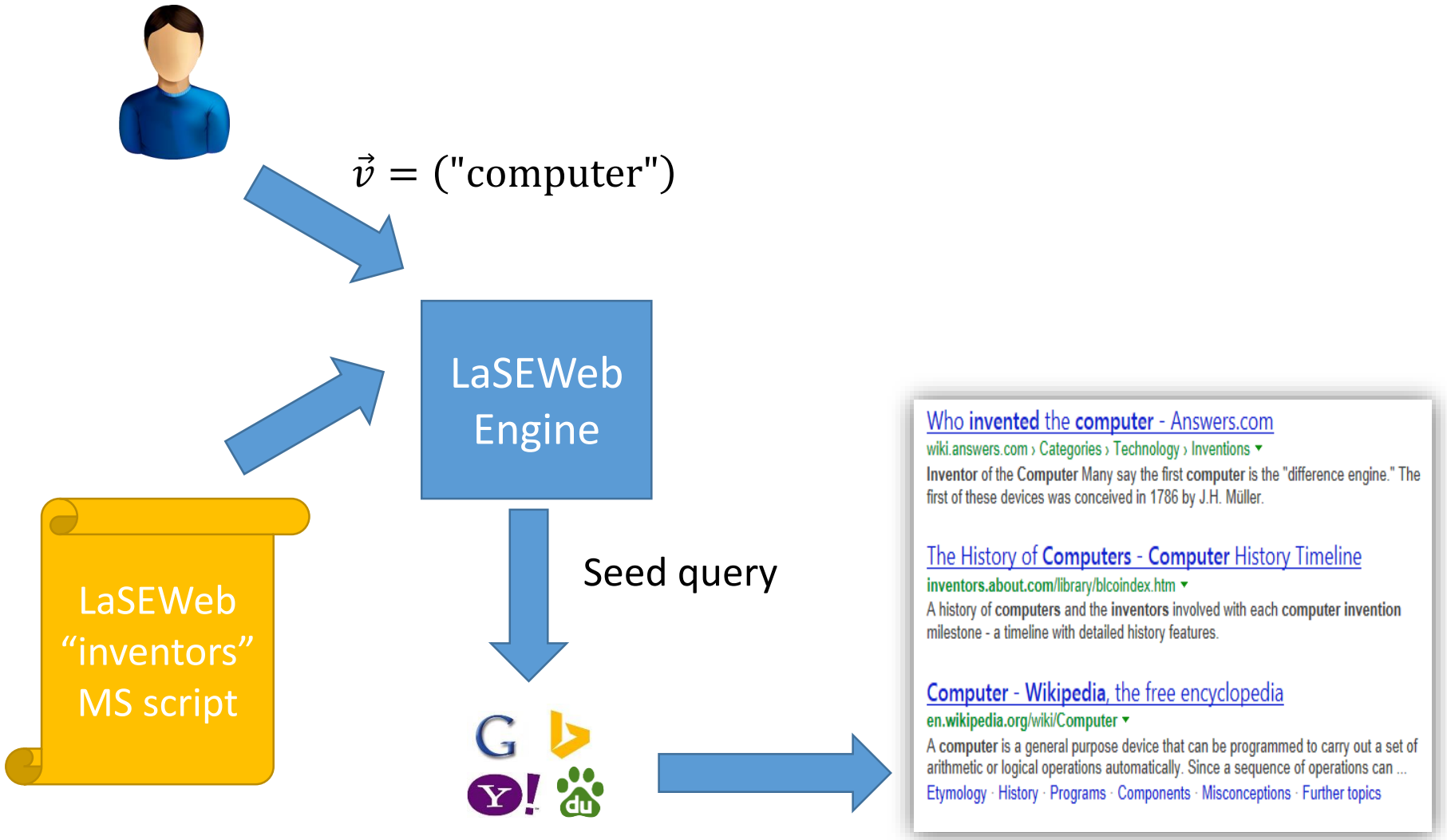
[1] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In ACL, 2005.
[2] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In ACL, 2003.
[3] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In HLT-NAACL, 2003.
[4] C. Quirk, P. Choudhury, J. Gao, H. Suzuki, K. Toutanova, M. Gamon, W.-t. Yih, L. Vanderwende, and C. Cherry. MSR SPLAT, a language analysis toolkit. In ACL, 2012.
[5] W.-t. Yih, G. Zweig, and J. C. Platt. Polarity inducing latent semantic analysis. In ACL, 2012.
[6] S. Gulwani. Automating string processing in spreadsheets using input-output examples. In POPL, 2011.
[7] M. J. Cafarella., A. Halevy, and J. Madhavan. Structured data on the web. In *CACM* 54.2 (2011): 72-79.

Program interpreter: “user emulation” algorithm

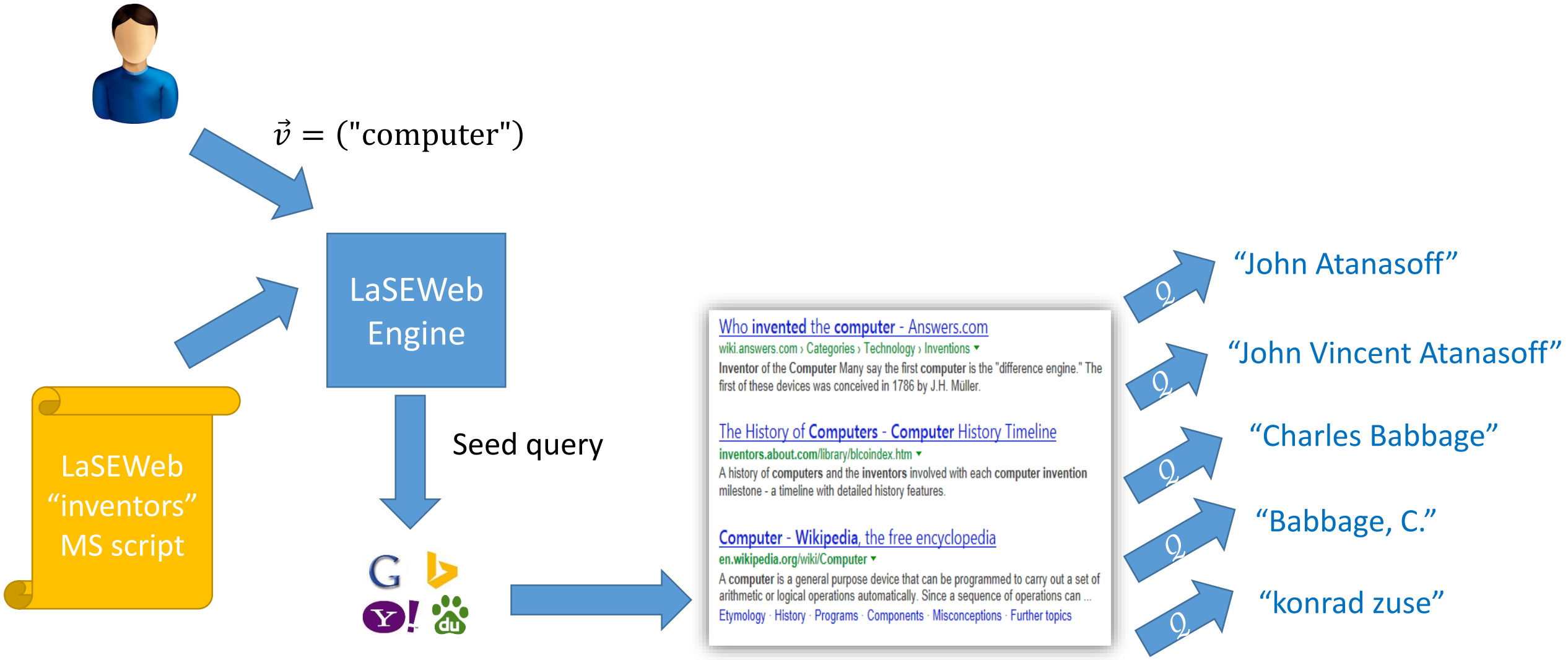
Program interpreter: “user emulation” algorithm



Program interpreter: “user emulation” algorithm



Program interpreter: “user emulation” algorithm



Program interpreter: “user emulation” algorithm

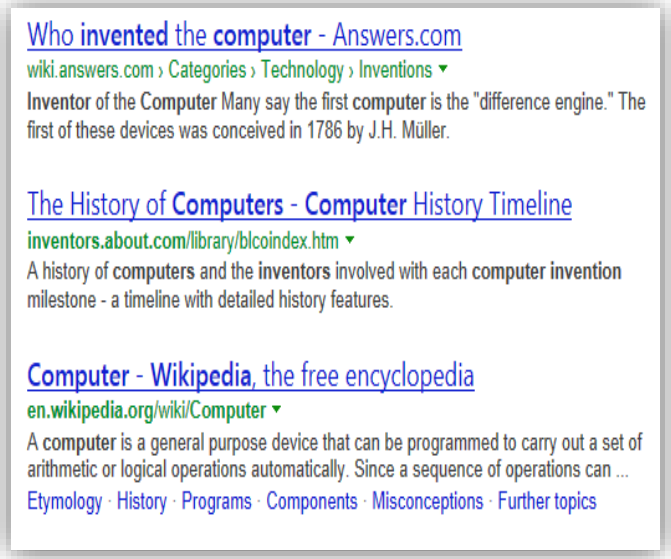
$\vec{v} = (\text{"computer"})$

$$score(C_i) = \frac{1}{|U|}\sum_{j=1}^{|U|}\sum_{s\in C_i}\frac{c(s,u_j)}{c(u_j)}$$

LaSEWeb Engine

LaSEWeb “inventors” MS script

Seed query



Q_1 “John Atanasoff”

Q_2 “John Vincent Atanasoff”

Q_3 “Charles Babbage”

Q_4 “Babbage, C.”

Q_5 “konrad zuse”

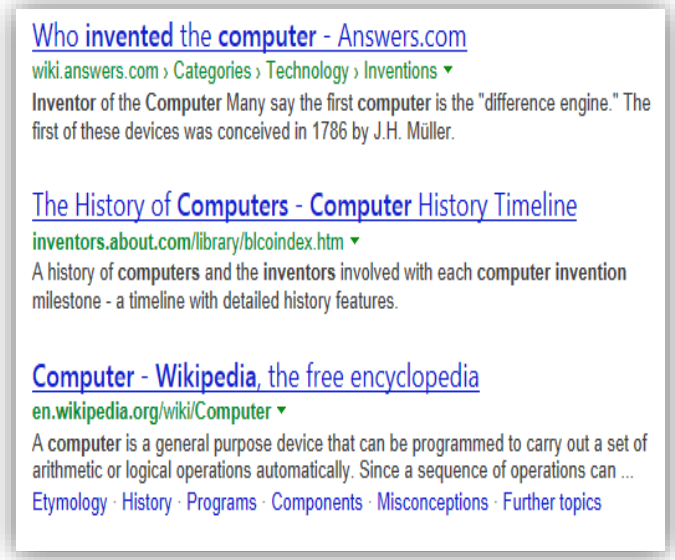
Program interpreter: “user emulation” algorithm

$\vec{v} = (\text{"computer"})$

LaSEWeb Engine

LaSEWeb “inventors” MS script

Seed query



Q_1 “John Atanasoff”

Q_2 “John Vincent Atanasoff”

Q_3 “Charles Babbage”

Q_4 “Babbage, C.”

Q_5 “konrad zuse”

John Atanasoff (14.5%)
http://www.computerhope.com
http://www.ehow.com
http://inventors.about.com

Charles Babbage (10.5%)
http://www.buzzle.com
http://www.ask.com

…

$$score(C_i) = \frac{1}{|U|}\sum_{j=1}^{|U|}\sum_{s\in C_i}\frac{c(s,u_j)}{c(u_j)}$$

Experiments

Micro-segment	# queries	Recall	Bing recall
ASCII code of a symbol	1,551	32.88%	0%
Calories in a food	9,207	71.80%	0%
Inventor of a product	8,994	75.91%	16.01%
Lyrics of a song	48,995	24.36%	0%
Phone number of a company	6,881	95.49%	0%
Population of a place	18,151	92.53%	57.58%
Release date of a product	12,339	97.24%	12.60%

Search task	Recall	Precision
Phone #	29/37	21/29
Affiliation	34/37	22/34
PhD institution	21/37	13/21
General chair	21/28	17/21
Invited talks	13/28	11/13
Average	**71%**	**73%**

- ~95% precision and 71% recall on factoid micro-segments
  - For micro-segments: Precision measured by random sampling, based on top-3 results
  - For end-user repetitive search tasks: Precision/recall measured manually
- Average execution time: ~5 sec/webpage
  - Depends on the rendering settings
- Current setting: offline deployment / database population

Summary & Future work

- Typical patterns of human search strategies in a scripting language for IE
  - Match semi-structured Web content
  - Existing cross-disciplinary technologies used as building blocks
  - Exploit information redundancy across multiple webpages
- Applications:
  1. Micro-segments of factoid questions in search engines
  2. Repeatable batch data extraction tasks for end-users
  3. Structured database population from free Web text
  4. English language comprehension problem generation
- Future work:
  - Automatic query execution plans in the language
  - Integration with “natural language → logic” engines

Summary & Future work

1. The principal characterized his pupils as _____ because they were pampered and spoiled by their indulgent parents.
2. The commentator characterized the electorate as _____ because it was unpredictable and given to constantly shifting moods.

(a) cosseted (b) disingenuous (c) corrosive (d) laconic (e) mercurial

2. Repeatable batch data extraction tasks for end-t
3. Structured database population from free Web text
4. English language comprehension problem generation

- Future work:
  - Automatic query execution plans in the language
  - Integration with “natural language → logic” engines

Summary & Future work

- Typical patterns of human search strategies in a scripting language for IE
  - Match semi-structured Web content
  - Existing cross-disciplinary technologies used as building blocks
  - Exploit information redundancy across multiple webpages
- Applications:
  1. Micro-segments of factoid questions in search engines
  2. Repeatable batch data extraction tasks for end-users
  3. Structured database population from free Web text
  4. English language comprehension problem generation
- Future work:
  - Automatic query execution plans in the language
  - Integration with “natural language → logic” engines

Thanks for listening!

Questions?