KenLM: Faster and Smaller Language Model Queries

Kenneth Heafield
heafield@cs.cmu.edu

Carnegie Mellon

July 30, 2011

kheafield.com/code/kenlm

What KenLM Does

Answer language model queries using less time and memory.

$$\begin{aligned}
\log p(\text{<s>} &\rightarrow \text{iran}) &&= \text{-3.33437}\\
\log p(\text{<s> iran} &\rightarrow \text{is}) &&= \text{-1.05931}\\
\log p(\text{<s> iran is} &\rightarrow \text{one}) &&= \text{-1.80743}\\
\log p(\text{<s> iran is one} &\rightarrow \text{of}) &&= \text{-0.03705}\\
\log p(\text{iran is one of} &\rightarrow \text{the}) &&= \text{-0.08317}\\
\log p(\text{is one of the} &\rightarrow \text{few}) &&= \text{-1.20788}
\end{aligned}$$

Related Work

Downloadable Baselines

SRI	Popular and considered fast but high-memory
IRST	Open source, low-memory, single-threaded
Rand	Low-memory lossy compression
MIT	Mostly estimates models but also does queries

Papers Without Code

TPT	Better memory locality
Sheffield	Lossy compression techniques

Related Work

Downloadable Baselines

SRI	Popular and considered fast but high-memory
IRST	Open source, low-memory, single-threaded
Rand	Low-memory lossy compression
MIT	Mostly estimates models but also does queries

Papers Without Code

TPT	Better memory locality
Sheffield	Lossy compression techniques

After KenLM's Public Release

Berkeley	Java; slower and larger than KenLM

Why I Wrote KenLM

Decoding takes too long

- Answer queries quickly
- Load quickly with memory mapping
- Thread-safe

Why I Wrote KenLM

Decoding takes too long

- Answer queries quickly
- Load quickly with memory mapping
- Thread-safe

Bigger models

- Conserve memory

Why I Wrote KenLM

Decoding takes too long

- Answer queries quickly
- Load quickly with memory mapping
- Thread-safe

Bigger models

- Conserve memory

SRI doesn't compile

- Distribute and compile with decoders

Outline

Example Language Model

Unigrams

Words	log p	Back
<s>	-∞	-2.0
iran	-4.1	-0.8
is	-2.5	-1.4
one	-3.3	-0.9
of	-2.5	-1.1

Bigrams

Words	log p	Back
<s> iran	-3.3	-1.2
iran is	-1.7	-0.4
is one	-2.0	-0.9
one of	-1.4	-0.6

Trigrams

Words	log p
<s> iran is	-1.1
iran is one	-2.0
is one of	-0.3

Example Queries

Unigrams

Words	log p	Back
<s>	-∞	-2.0
iran	-4.1	-0.8
is	-2.5	-1.4
one	-3.3	-0.9
of	-2.5	-1.1

Bigrams

Words	log p	Back
<s> iran	-3.3	-1.2
iran is	-1.7	-0.4
is one	-2.0	-0.9
one of	-1.4	-0.6

Trigrams

Words	log p
<s> iran is	-1.1
iran is one	-2.0
is one of	-0.3

Query: <s> iran is

$$\log p(\text{<s> iran} \rightarrow \text{is}) = \text{-1.1}$$

Query: iran is of

$$\begin{aligned} &\log p(\text{of}) && \text{-2.5} \\ &\text{Backoff(is)} && \text{-1.4} \\ &\text{Backoff(iran is)} && + \text{-0.4} \\ \hline &\log p(\text{iran is} \rightarrow \text{of}) && = \text{-4.3} \end{aligned}$$

Lookups Performed by Queries

<s> iran is

Lookup

1. is
2. iran is
3. <s> iran is

Score

$\log p(\text{<s> iran} \rightarrow \text{is}) = \text{-1.1}$

iran is of

Lookup

1. of
2. is of (not found)
3. is
4. iran is

Score

$\log p(\text{of})$	-2.5
Backoff(is)	-1.4
Backoff(iran is)	+ -0.4
$\log p(\text{iran is} \rightarrow \text{of})$	= -4.3

Lookups Performed by Queries

<s> iran is

Lookup

1. is
2. iran is
3. <s> iran is

Score

$\log p(\text{<s> iran} \rightarrow \text{is}) = \text{-1.1}$

iran is of

Lookup

1. of
2. is of (not found)
3. is
4. iran is

Score

$\log p(\text{of})$	-2.5
Backoff(is)	-1.4
Backoff(iran is)	+ -0.4
$\log p(\text{iran is} \rightarrow \text{of}) =$	-4.3

Lookups Performed by Queries

<s> iran is

Lookup

1. is
2. iran is
3. <s> iran is

State

Backoff(is)
Backoff(iran is)

Score

$\log p(\text{<s> iran} \rightarrow \text{is}) = \text{-1.1}$

iran is of

Lookup

1. of
2. is of (not found)
3. ~~is~~
4. ~~iran is~~

Score

$\log p(\text{of})$	-2.5
Backoff(is)	-1.4
Backoff(iran is)	+ -0.4
$\log p(\text{iran is} \rightarrow \text{of}) =$	-4.3

Stateful Query Pattern

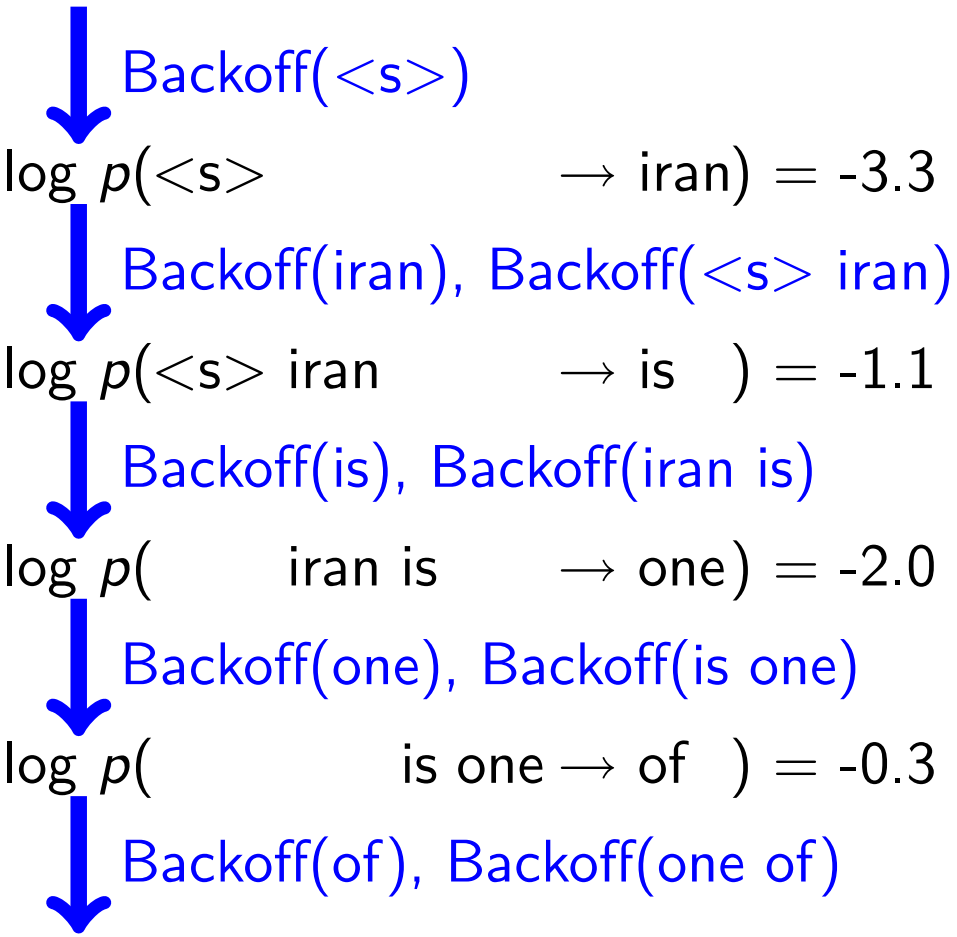
$$\log p(\text{<s>} \rightarrow \text{iran}) = \text{-3.3}$$

$$\log p(\text{<s> iran} \rightarrow \text{is}) = \text{-1.1}$$

$$\log p(\text{iran is} \rightarrow \text{one}) = \text{-2.0}$$

$$\log p(\text{is one} \rightarrow \text{of}) = \text{-0.3}$$

Stateful Query Pattern



Data Structures

Probing Fast. Uses hash tables.

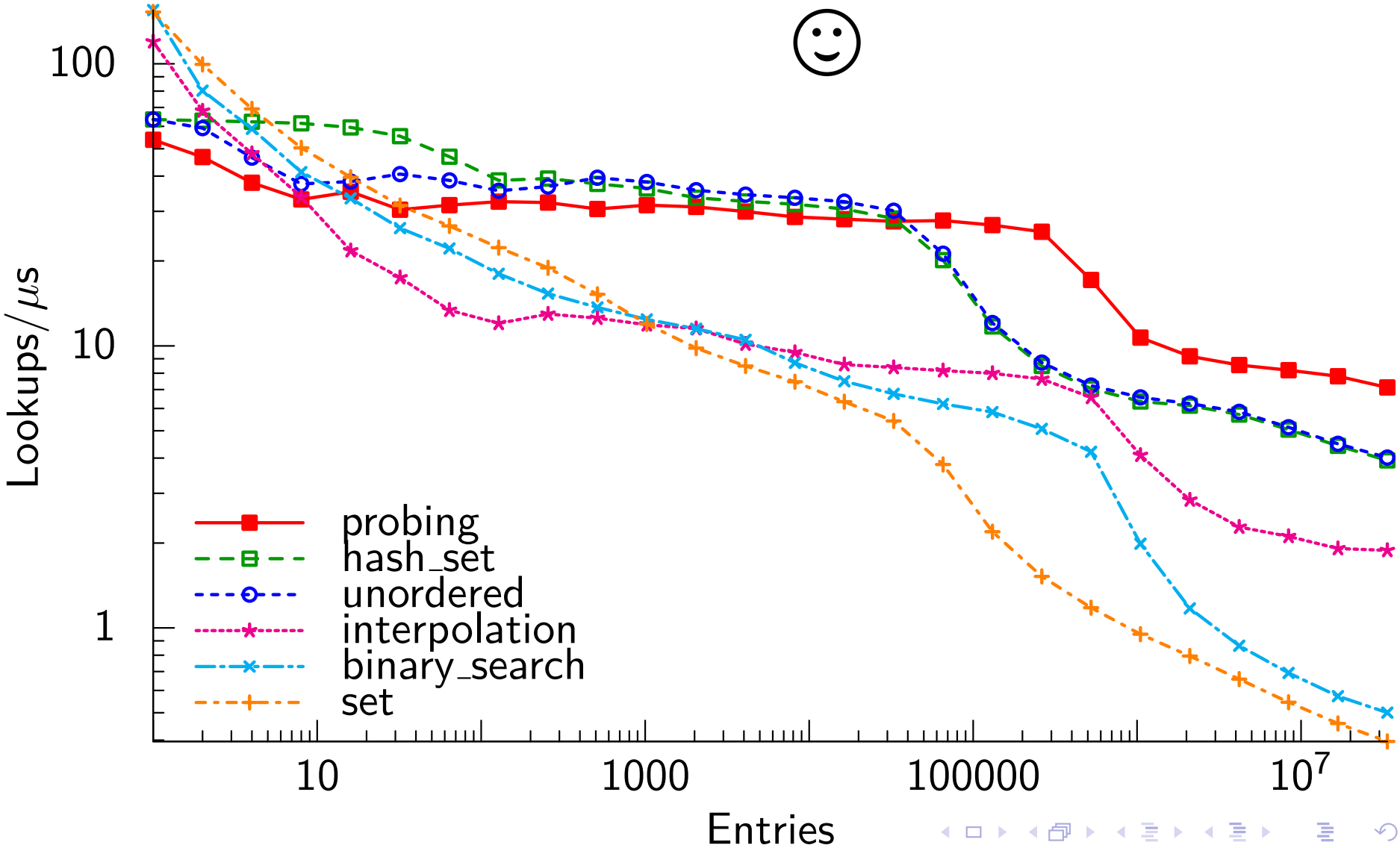
Trie Small. Uses sorted arrays.

Chop Smaller. Trie with compressed pointers.

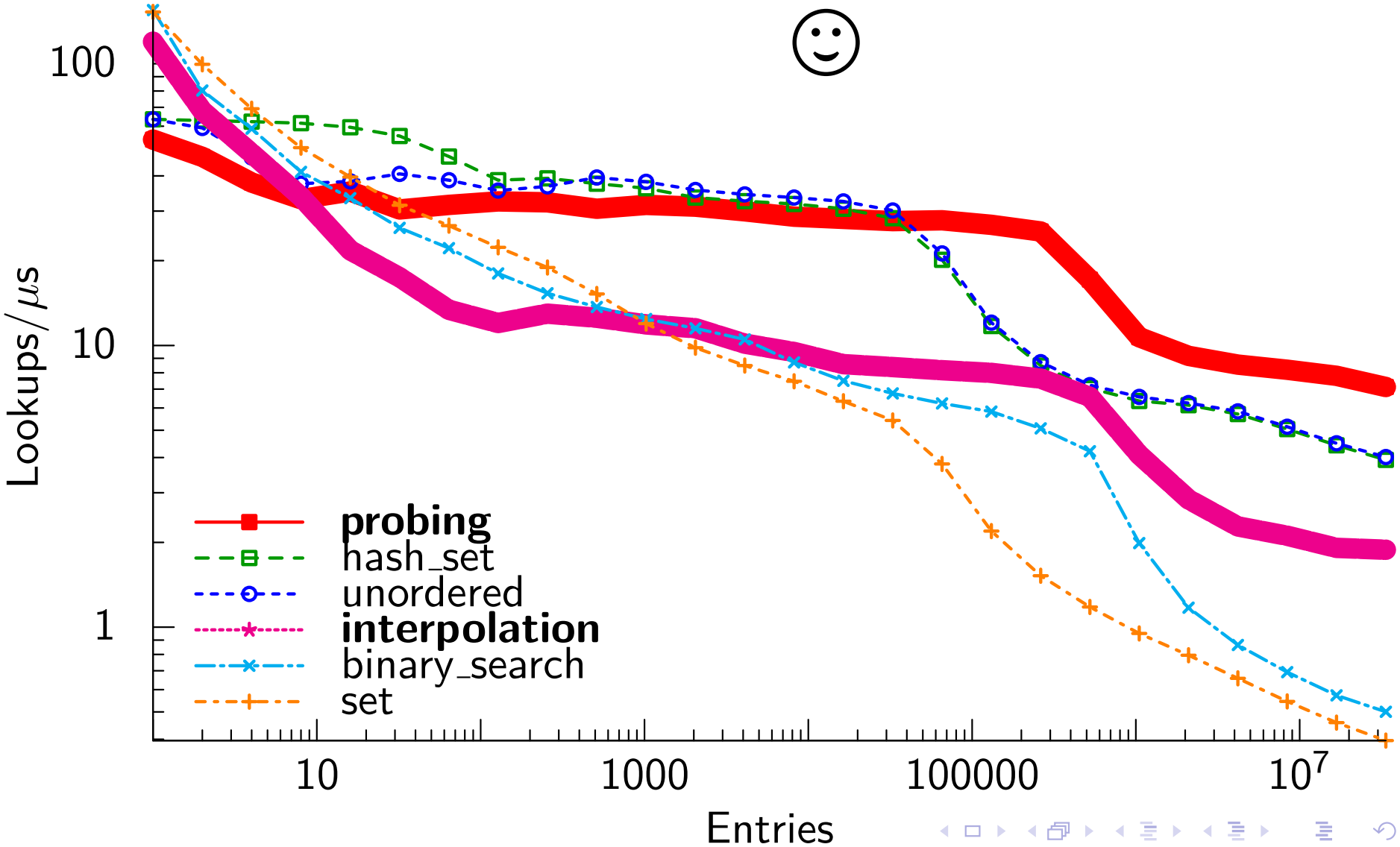
Key Subproblem

Sparse lookup: efficiently retrieve values for sparse keys

Sparse Lookup Speed



Sparse Lookup Speed



Linear Probing Hash Table

Store 64-bit hashes and ignore collisions.

	Bigrams		
Words	**Hash**	**log p**	**Back**
<s> iran	0xf0ae9c2442c6920e	-3.3	-1.2
iran is	0x959e48455f4a2e90	-1.7	-0.4
is one	0x186a7caef34acf16	-2.0	-0.9
one of	0xac66610314db8dac	-1.4	-0.6

Linear Probing Hash Table

- 1.5 buckets/entry (so buckets = 6).
- Ideal bucket = hash mod buckets.
- Resolve bucket collisions using the next free bucket.

Bigrams

Words	**Ideal**	**Hash**	**log *p***	**Back**
iran is	0	0x959e48455f4a2e90	-1.7	-0.4
		0x0	0	0
is one	2	0x186a7caef34acf16	-2.0	-0.9
one of	2	0xac66610314db8dac	-1.4	-0.6
<s> iran	4	0xf0ae9c2442c6920e	-3.3	-1.2
		0x0	0	0

Array

Probing Data Structure

Unigrams

Words	log p	Back
<s>	-∞	-2.0
iran	-4.1	-0.8
is	-2.5	-1.4
one	-3.3	-0.9
of	-2.5	-1.1

Array

Bigrams

Words	log p	Back
<s> iran	-3.3	-1.2
iran is	-1.7	-0.4
is one	-2.0	-0.9
one of	-1.4	-0.6

Probing Hash Table

Trigrams

Words	log p
<s> iran is	-1.1
iran is one	-2.0
is one of	-0.3

Probing Hash Table

Probing Hash Table Summary

Hash tables are fast. But memory is 24 bytes/entry.

Next: Saving memory with Trie.

Trie Uses Sorted Arrays

Sort in suffix order.

Unigrams

Words	log p	Back
<s>	-∞	-2.0
iran	-4.1	-0.8
is	-2.5	-1.4
one	-3.3	-0.9
of	-2.5	-1.1

Bigrams

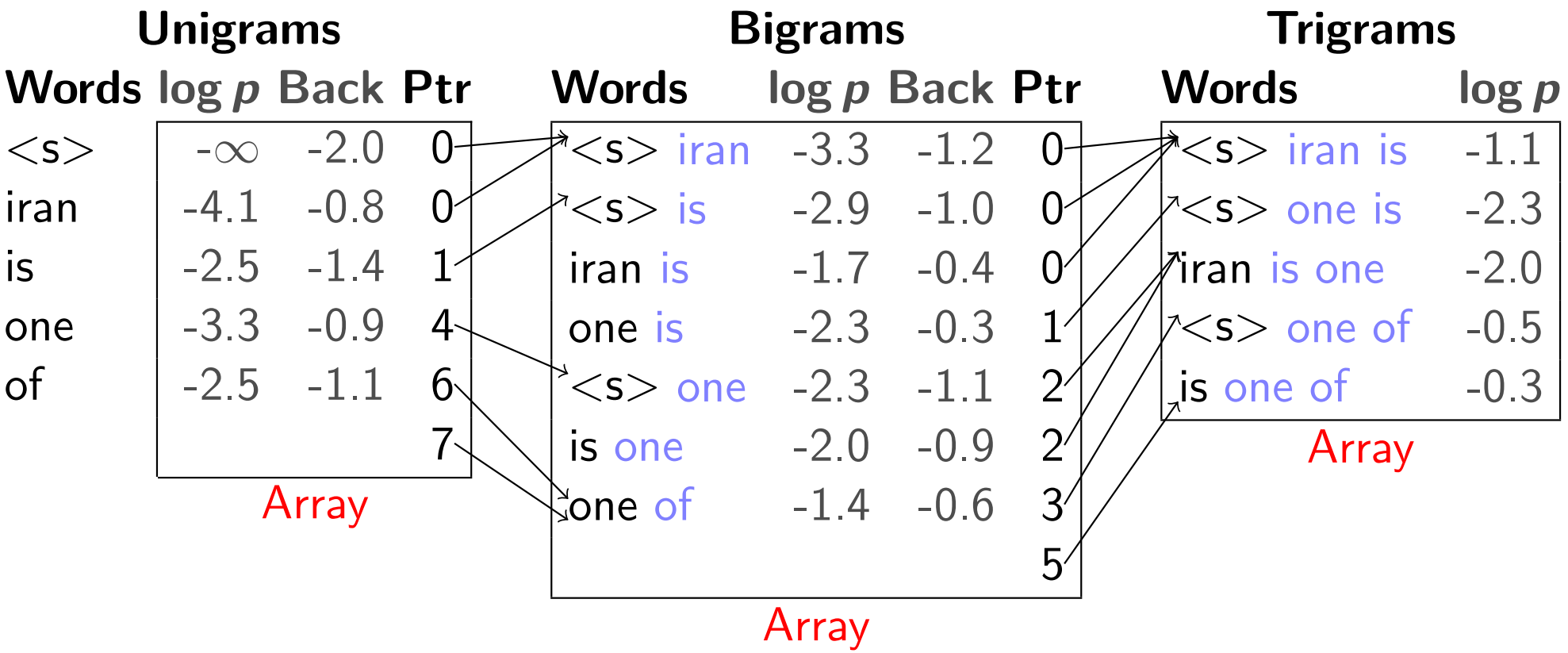
Words	log p	Back
<s> iran	-3.3	-1.2
iran is	-1.7	-0.4
one is	-2.3	-0.3
<s> one	-2.3	-1.1
is one	-2.0	-0.9
one of	-1.4	-0.6

Trigrams

Words	log p
<s> iran is	-1.1
<s> one is	-2.3
iran is one	-2.0
<s> one of	-0.5
is one of	-0.3

Trie

Sort in suffix order. Encode suffix using pointers.



Interpolation Search In Trie

Each trie node is a sorted array.

Bigrams: * is

Words	**log *p***	**Back**	**Ptr**
<s> is	-2.9	-1.0	0
iran is	-1.7	-0.4	0
one is	-2.3	-0.3	1

Interpolation Search $O(\log \log n)$

$$pivot = |A| \frac{key - A.first}{A.last - A.first}$$

Binary Search: $O(logn)$

$$pivot = \frac{|A|}{2}$$

Saving Memory with Trie

Bit-Level Packing

Store word index and pointer using the minimum number of bits.

Optional Quantization

Cluster floats into 2^q bins, store q bits/float (same as IRSTLM).

Chop: Compress Trie Pointers

Bigrams

Words	**log *p***	**Back**	**Ptr**
<s> iran	-3.3	-1.2	0
iran is	-1.7	-0.4	0
one is	-2.3	-0.3	1
<s> one	-2.3	-1.1	2
is one	-2.0	-0.9	2
one of	-1.4	-0.6	3
			5

Chop: Compress Trie Pointers

Offset	Ptr	Binary
0	0	000
1	0	000
2	1	001
3	2	010
4	2	010
5	3	011
6	5	101

Raj and Whittaker (2003)

Chop: Compress Trie Pointers

Offset	Ptr	Binary
0	0	000
1	0	000
2	1	001
3	2	010
4	2	010
5	3	011
6	5	101

Chopped	Offset
1	6

Raj and Whittaker (2003)

Chop: Compress Trie Pointers

Offset	Ptr	Binary
0	0	000
1	0	000
2	1	001
3	2	010
4	2	010
5	3	011
6	5	101

Chopped	Offset
01	3
10	6

Raj and Whittaker (2003)

Trie/Chop Summary

Save memory: bit packing, quantization, and pointer compression.

Outline

1. Backoff Models
   - State
2. Data Structures
   - Probing
   - Trie
   - Chop
3. Results
   - Perplexity
   - Translation

Perplexity Task

Score the English Gigaword corpus.

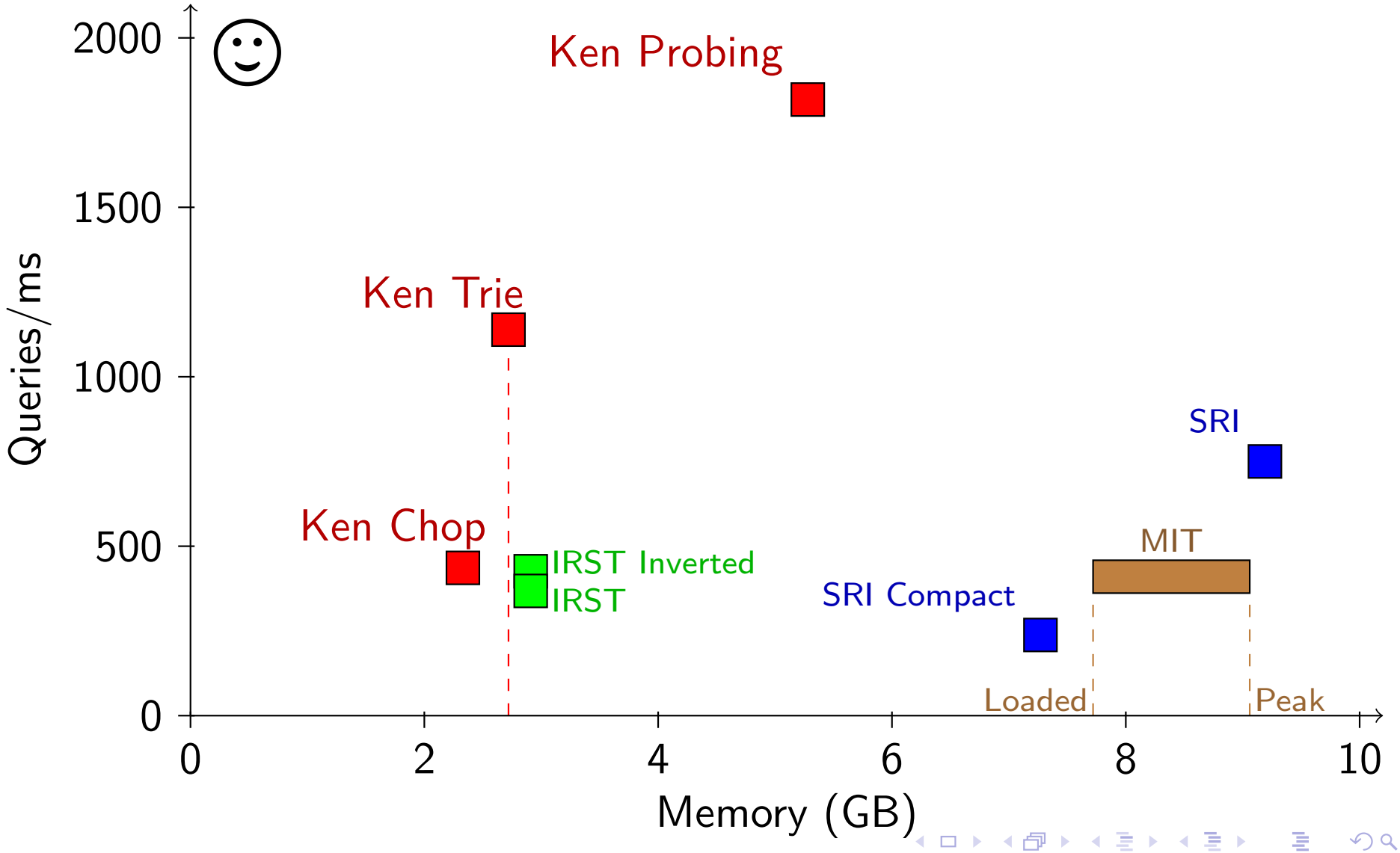
Model

SRILM 5-gram from Europarl + De-duplicated News Crawl

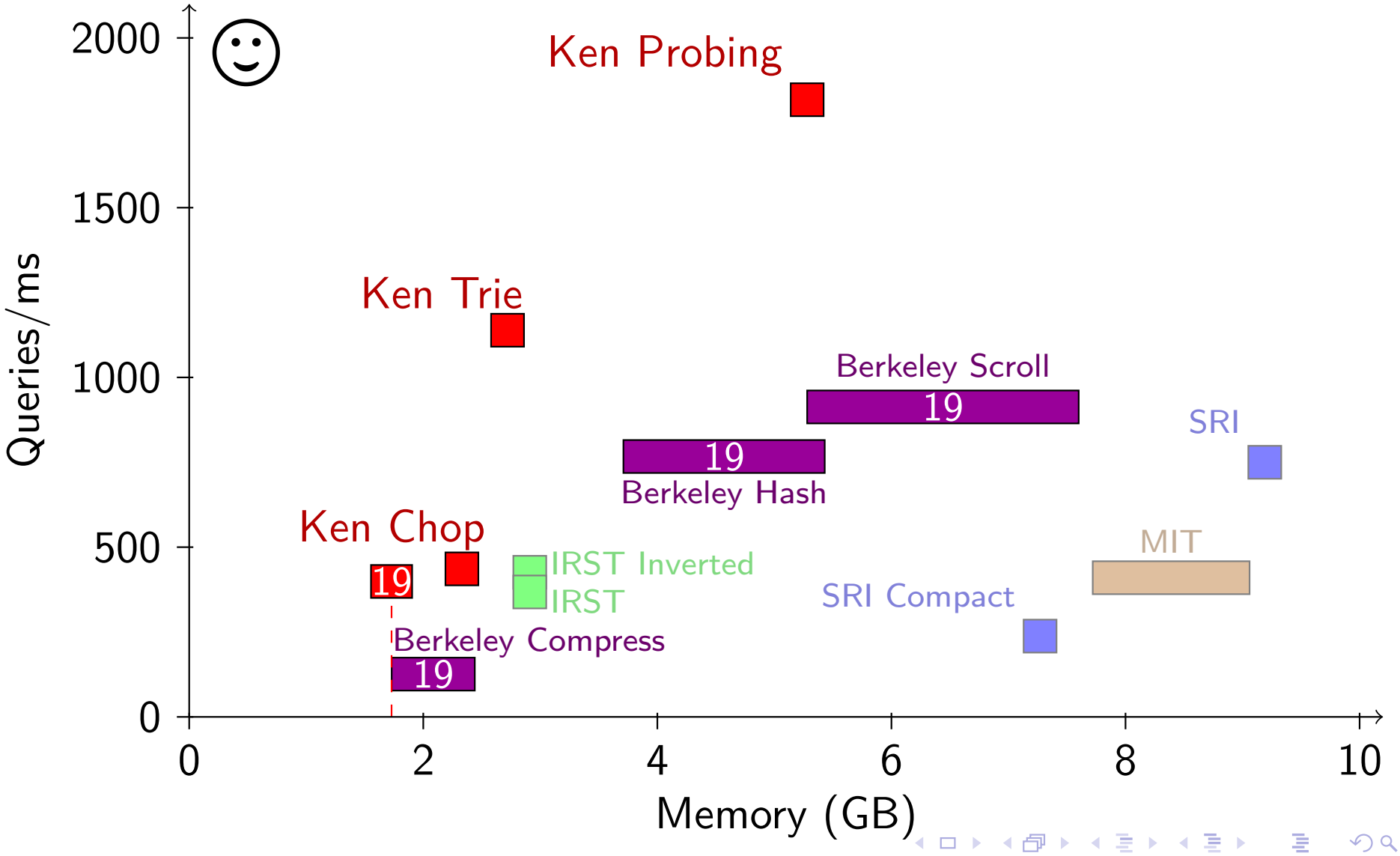
Measurements

Queries/ms	Excludes loading and file reading time
Loaded Memory	Resident after loading
Peak Memory	Peak virtual after scoring

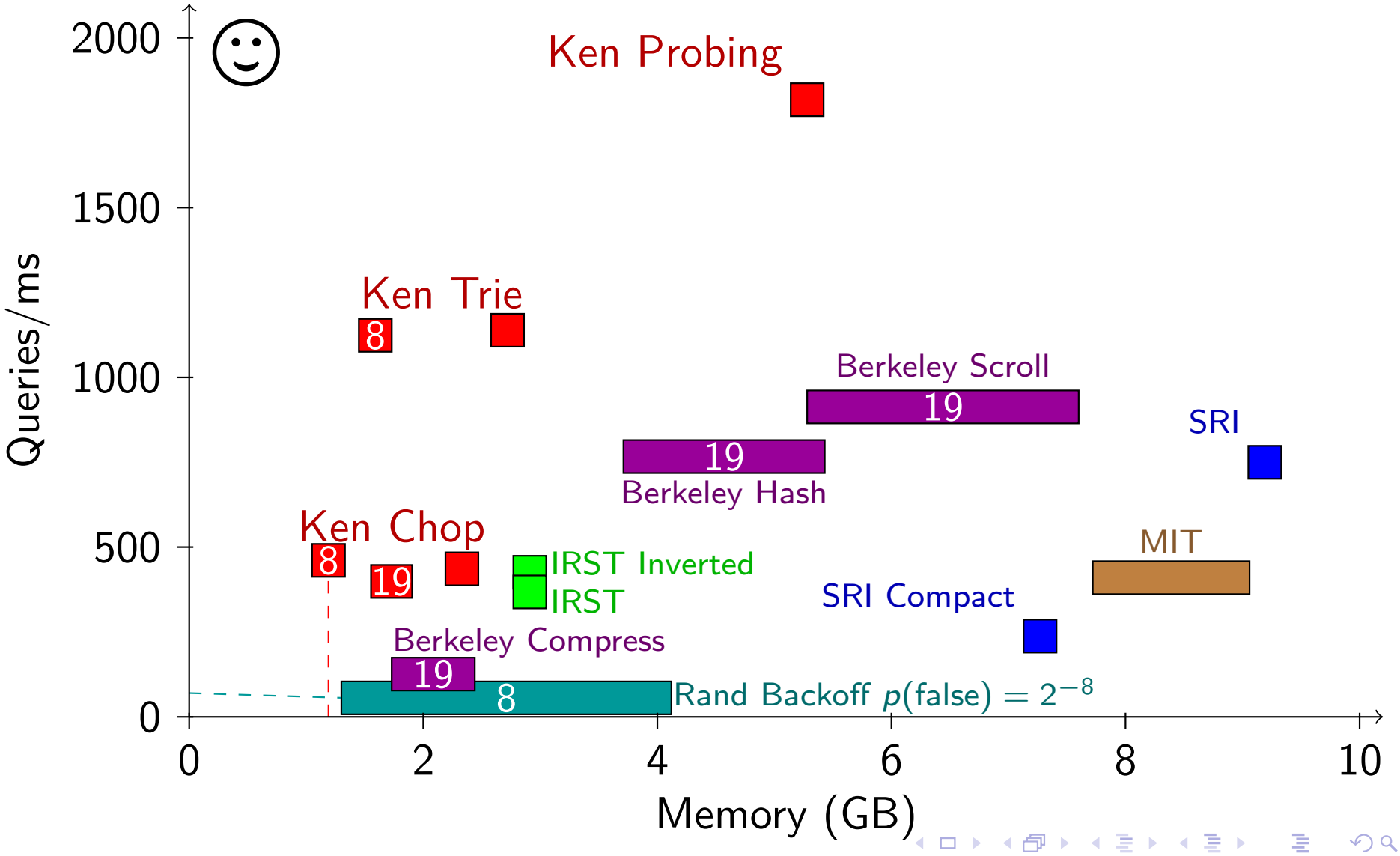
Perplexity Task: Exact Models



Perplexity Task: Berkeley Always Quantizes to 19 bits



Perplexity Task: RandLM from an ARPA file



Translation Task

Translate 3003 sentences using Moses.

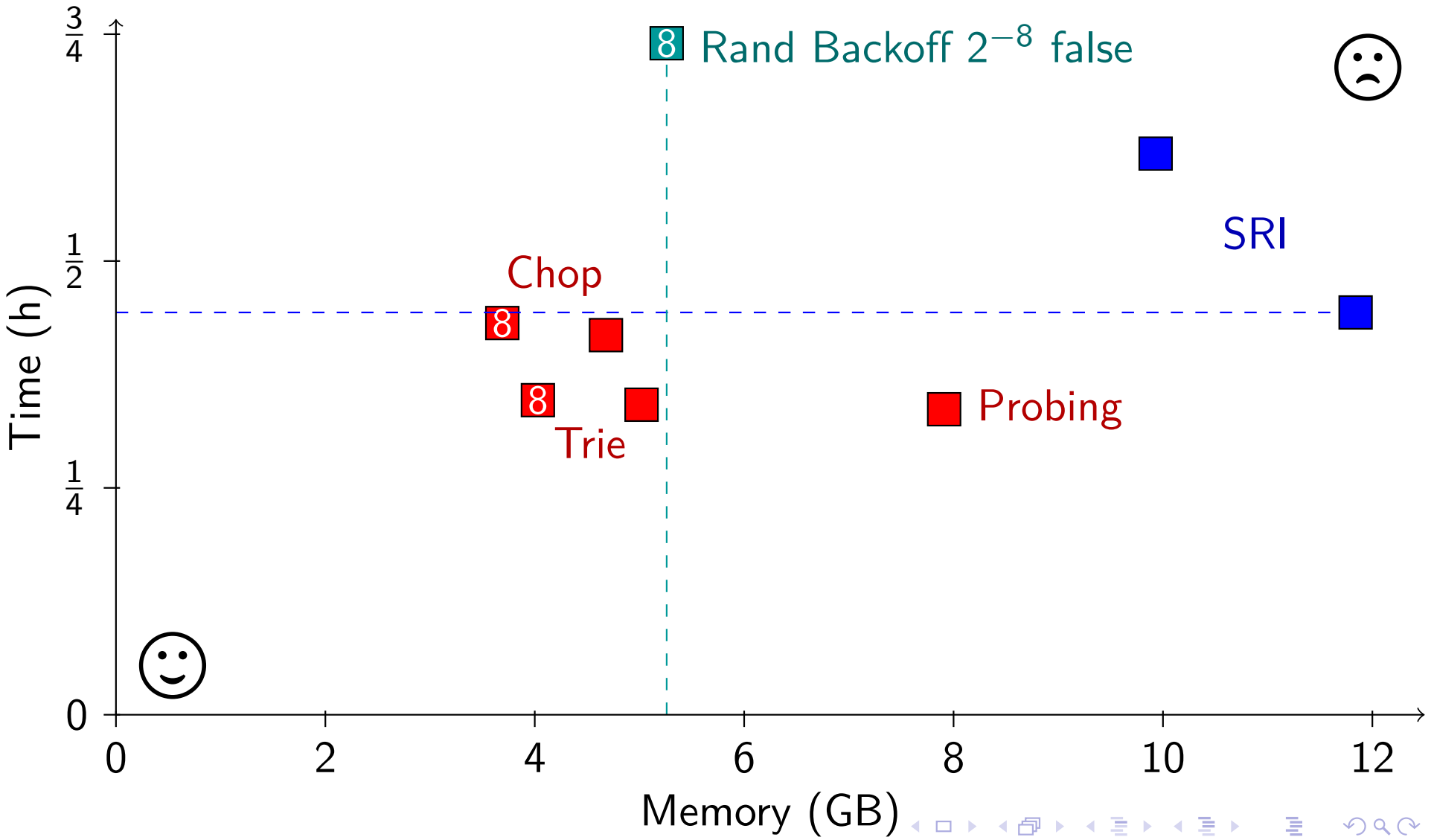
System

WMT 2011 French-English baseline, Europarl+News LM

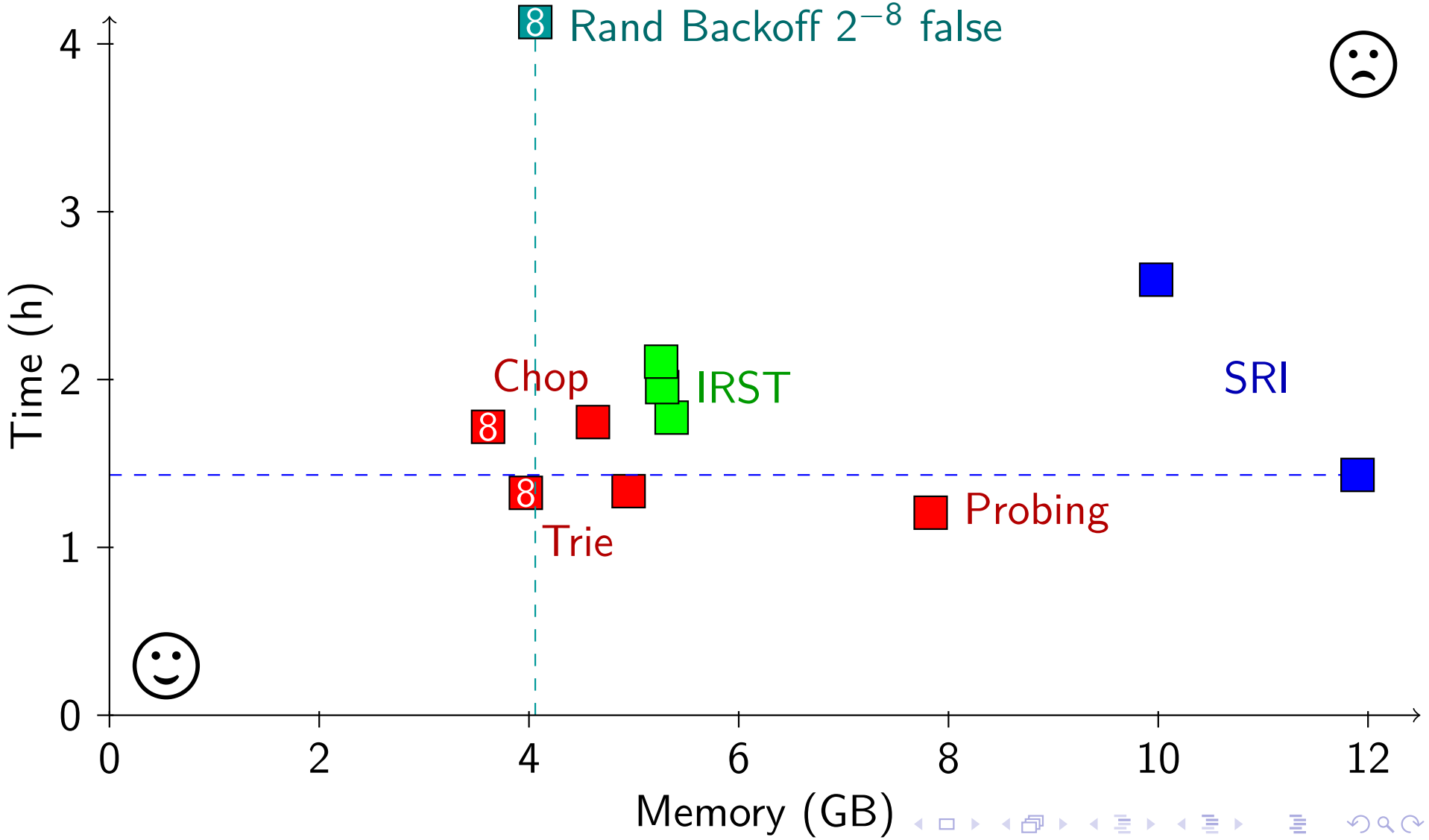
Measurements

Time	Total wall time, including loading
Memory	Total resident memory after decoding

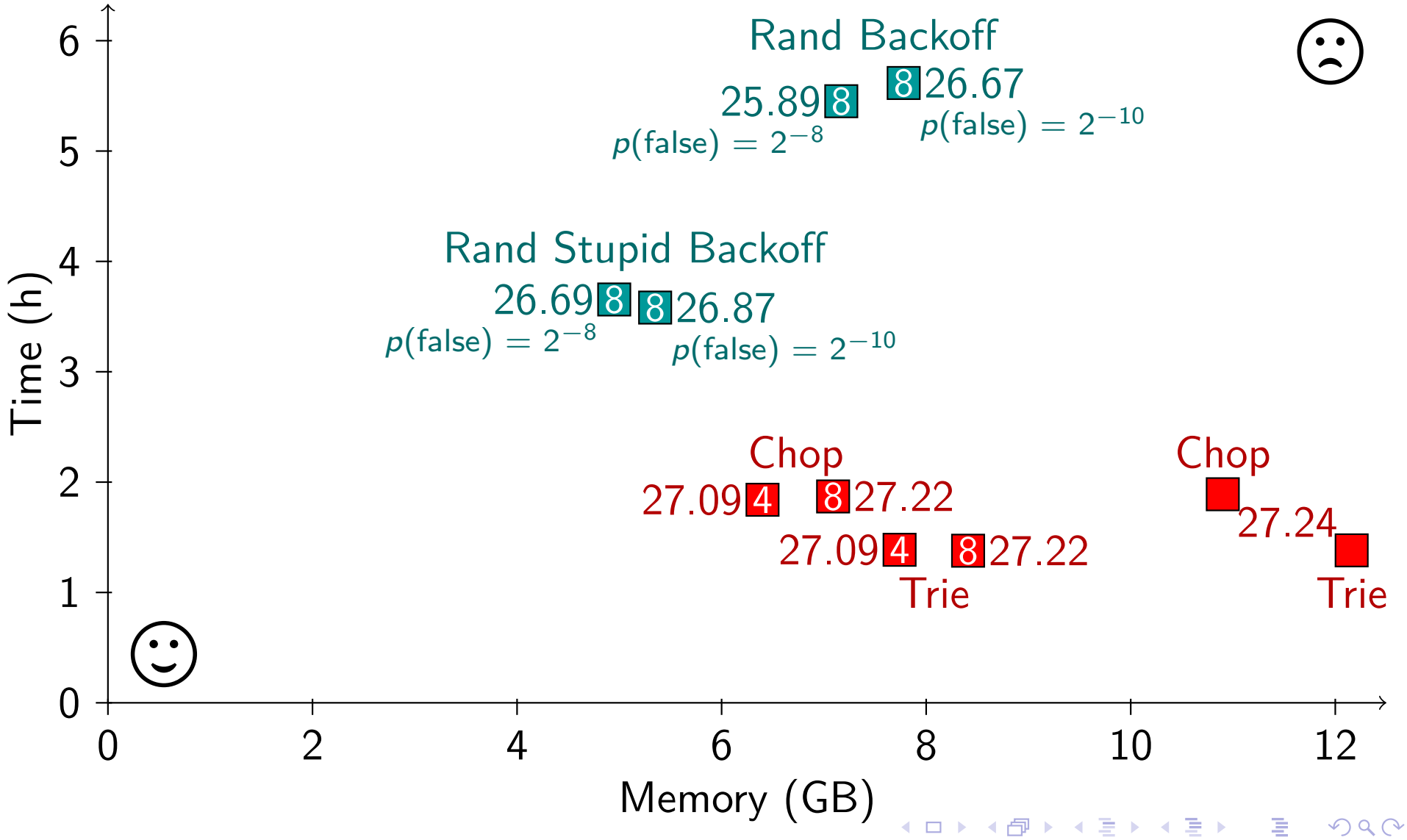
Moses Benchmarks: 8 Threads



Moses Benchmarks: Single Threaded



Comparison to RandLM (Unpruned Model, One Thread)



Conclusion

Maximize speed and accuracy subject to memory.

Probing > Trie > Chop > RandLM Stupid

for both speed and memory.

Distributed with decoders:	Moses	`8 0 5 file`
	cdec	`KLanguageModel`
	Joshua	`use_kenlm=true`

`kheafield.com/code/kenlm/`