The "Naturalness" of Software: A Research Vision

Abram Hindle, **Earl Barr**, Zhendong Su,
Mark Gabel, and Premkumar Devanbu

UCDAVIS

Microsoft® Research


```
public class FunctionCall {
    public static void funct1 () {
      System.out.println ("Inside funct1");
    }
    public static void main (String[] args) {
        int val;
        System.out.println ("Inside main");
        funct1();
        System.out.println ("About to call funct2");
        val = funct2(8);
        System.out.println ("funct2 returned a value of " + val);
        System.out.println ("About to call funct2 again");
        val = funct2(-3);
        System.out.println ("funct2 returned a value of " + val);
   }
    public static int funct2 (int param) {
        System.out.println ("Inside funct2 with param " + param);
        return param * 2;
    }
}
```

English, Tamil, German

Can be rich, powerful, expressive



..but "in nature" are mostly simple, repetitive, boring

Statistical Models

Can be rich, powerful, expressive

Mostly simple, repetitive, boring

Statistical Models

Two Examples

A speech recognizer example

"European Central Fish" ?

Another speech recognizer example

"fish++" ?

Repetition

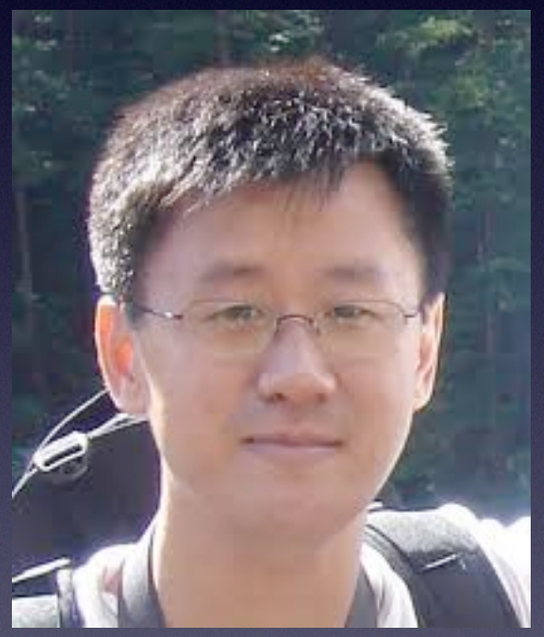
Mathematical Models

Useful Software

Is software really repetitive?

The "Uniqueness" of Code

Mark Gabel



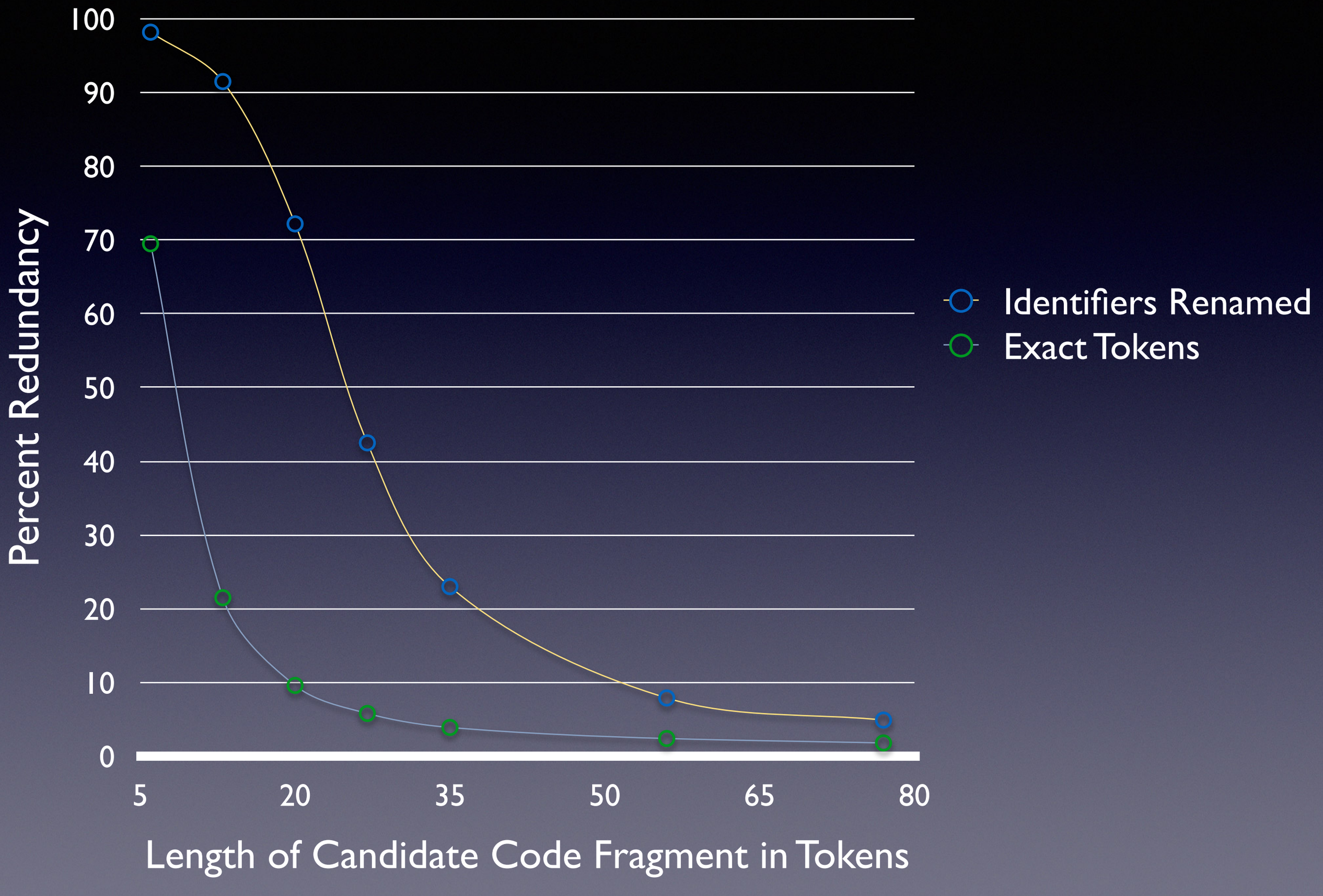
Zhendong Su

A study of the Uniqueness of Source Code, Gabel and Su, *ACM SIGSOFT FSE 2010*

How Redundant is Code?

How much code?	6000 projects (C, C++, Java) 430,000,000 LOC
How long?	Sequences of 6-77 token length
(1) How matched?	Exact Match 1-4 edits
(2) How matched?	Raw Tokens Renamed Identifiers

Non-Uniqueness (Redundancy) in a Large Java Corpus



Software *is* really repetitive.

How can we use this?

How has the “naturalness” (repetitive structure) of natural language been exploited?

Large Corpora

Language Models

Speech Recognition,
Translation, etc.

Language Models

For any utterance U, $0 \leq p(U) \leq 1$

If U_a is often uttered than U_b, $p(U_a) > p(U_b)$

$p(\text{“}EuropeanCentralFish\text{”}) < p(\text{“}EuropeanCentralBank\text{”})$

$p(\texttt{for(i} = \texttt{0; i} < \texttt{10; fish} + +)) < p(\texttt{for(i} = \texttt{0; i} < \texttt{10; i} + +))$

History of Language Models in NLP

- [illegible]
- [illegible]

> "Every time I fire a linguist, the performance of our speech recognizer goes up"
> —Fred Jelenik

➡ Good, high quality language models

➡ Rapid, revolutionary advances

Language Models: a Revolution in NLP

The design and estimation of language models is at the heart of modern NLP

Good

- S
- N
- D
- Document retrieval

But what about code?

and

“code language models”?

Exploiting Code Language Models

Suggest the next token for developers

Complete the current token for developers

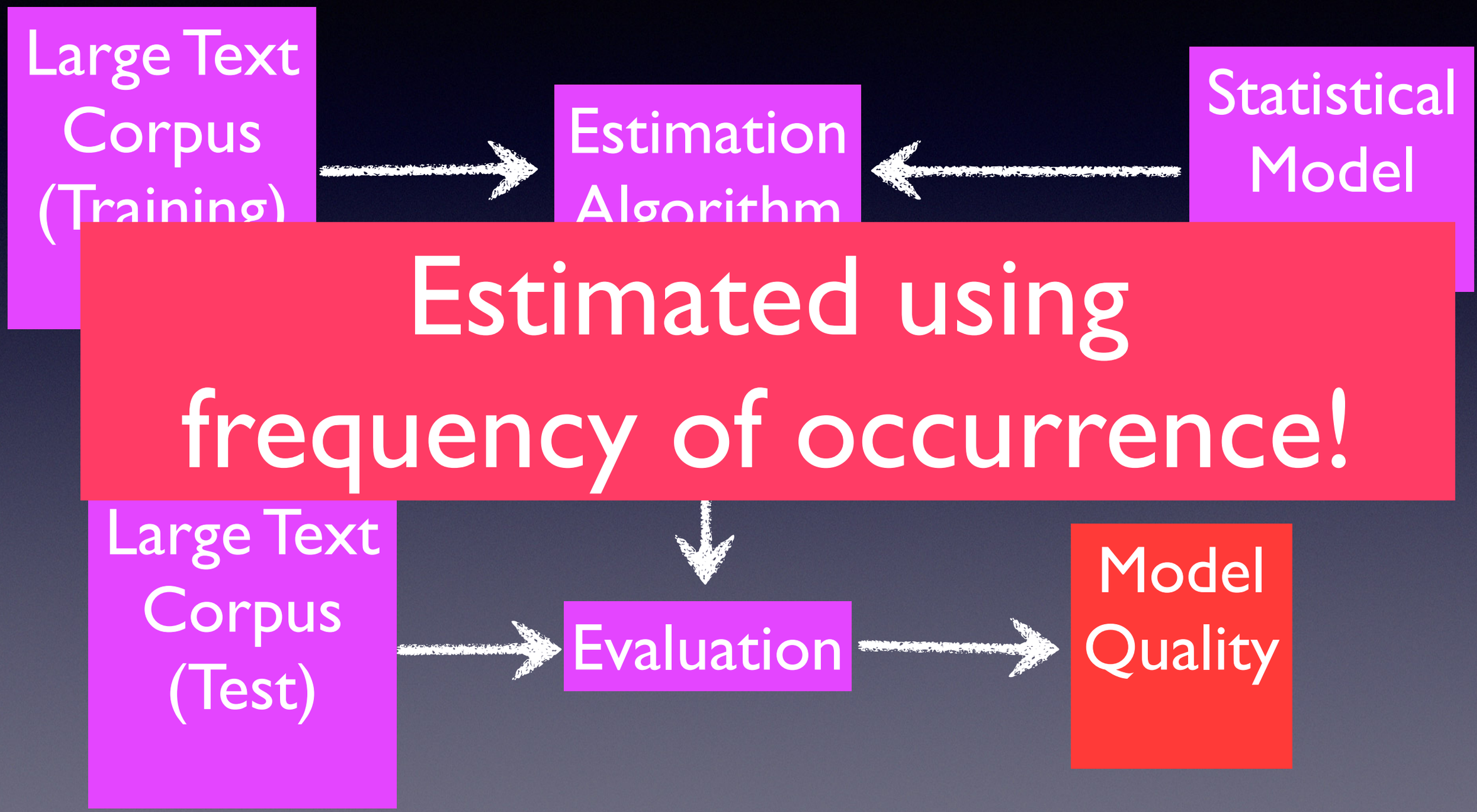
Assistive (speech, gesture) coding

Summarization and retrieval as translation

Fast, “good guess” static analysis

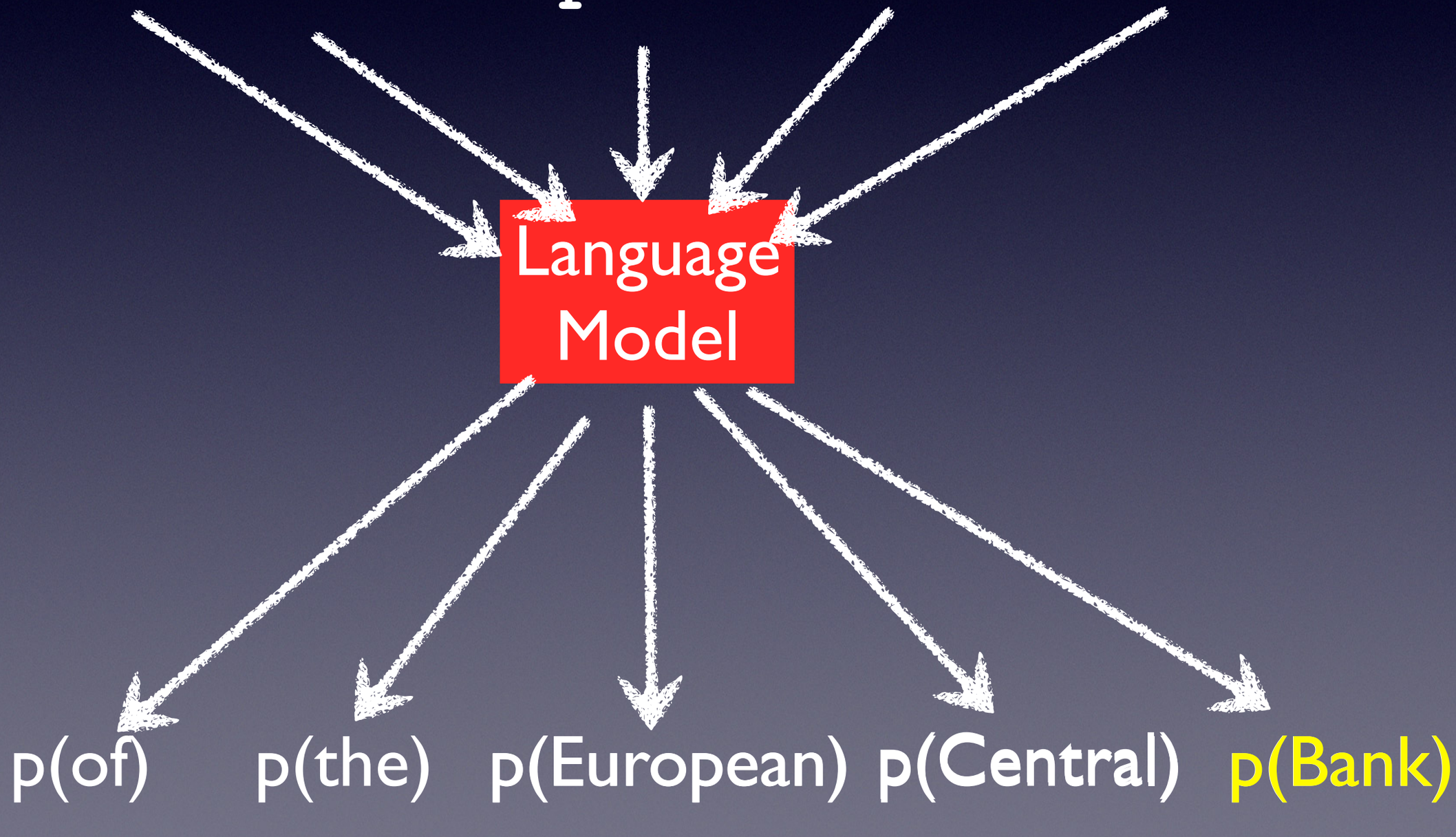
Search-based Software Engineering

Building a Language Model



What a Language Model Does

..of the European Central Bank



Language Models

Vastly more complex

Almost always face data-sparsity

Novel, NLP-specific estimation methods

Evaluating Language Model Quality

The words it encounters are not "too surprising" to it.

- ☑ Frequently encountered language events are assigned higher probability
- ☑ Infrequent language events are assigned lower probability.
- ☑ *....measured using "Cross-Entropy"*

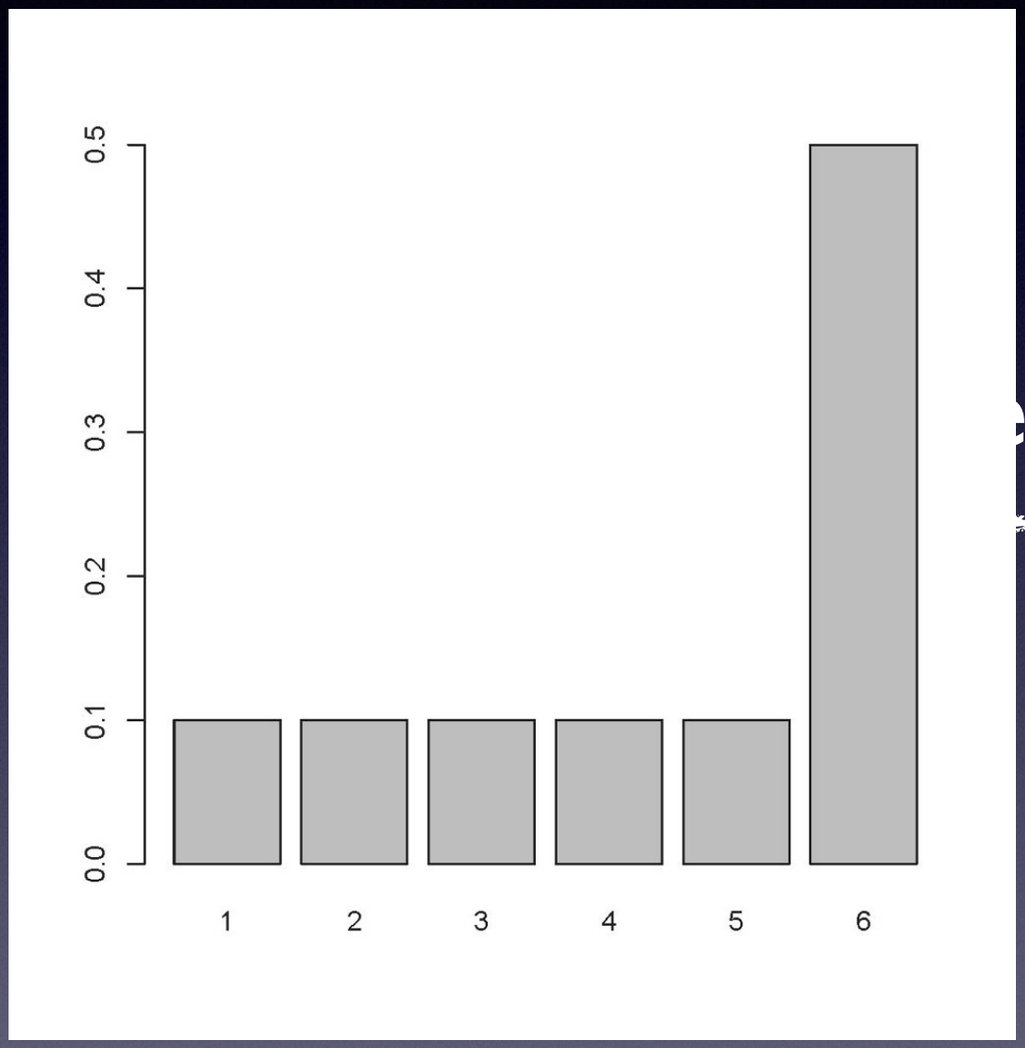
Background
Cross Entropy

Language Model

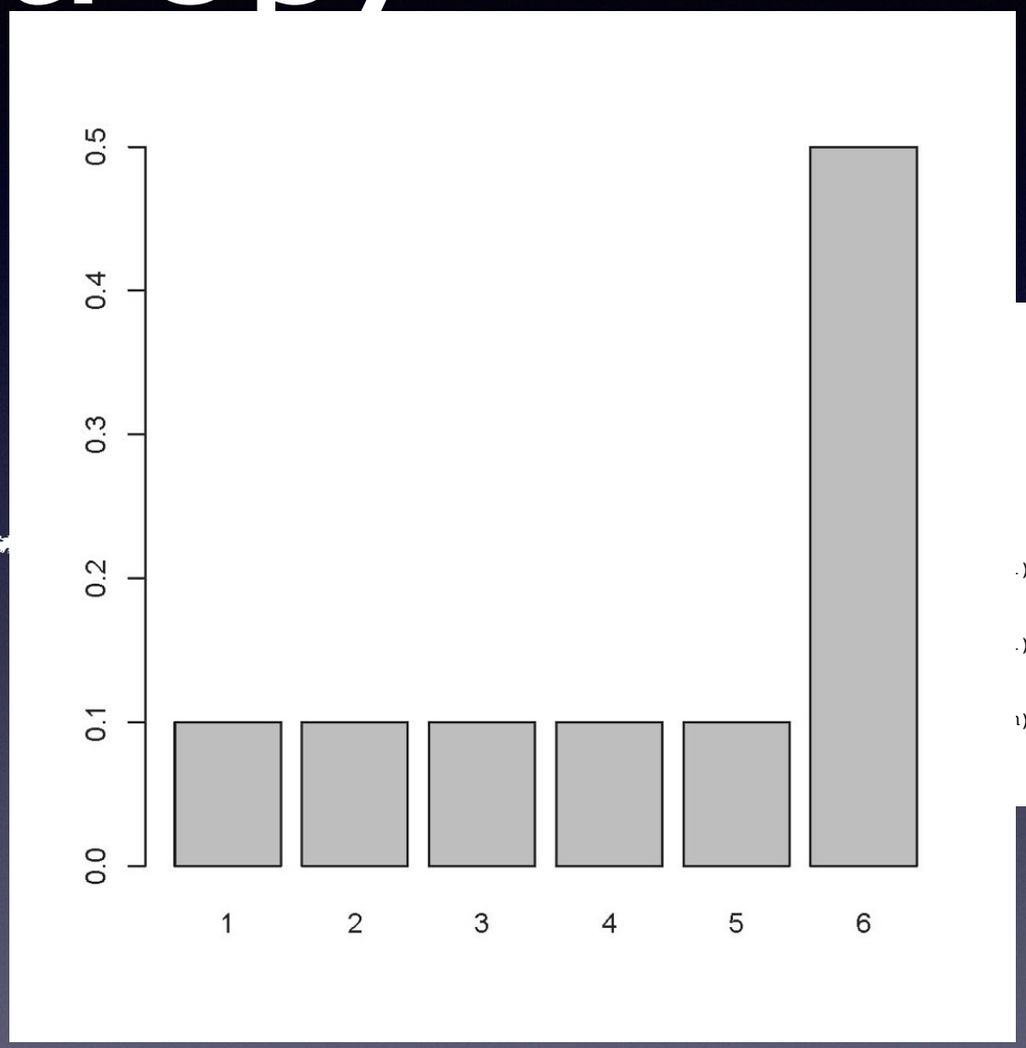
Good Description?

```
public class FunctionCall {
    public static void funct1 () {
        System.out.println ("Inside funct1");
    }
    public static void main (String[] args) {
        int val;
        System.out.println ("Inside main");
        funct1();
        System.out.println ("About to call funct2");
        val = funct2(8);
        System.out.println ("funct2 returned a value of " + val);
        System.out.println ("About to call funct2 again");
        val = funct2(-3);
        System.out.println ("funct2 returned a value of " + val);
   }
    public static int funct2 (int param) {
        System.out.println ("Inside funct2 with param " + param);
        return param * 2;
    }
}
```

Background
Cross Entropy

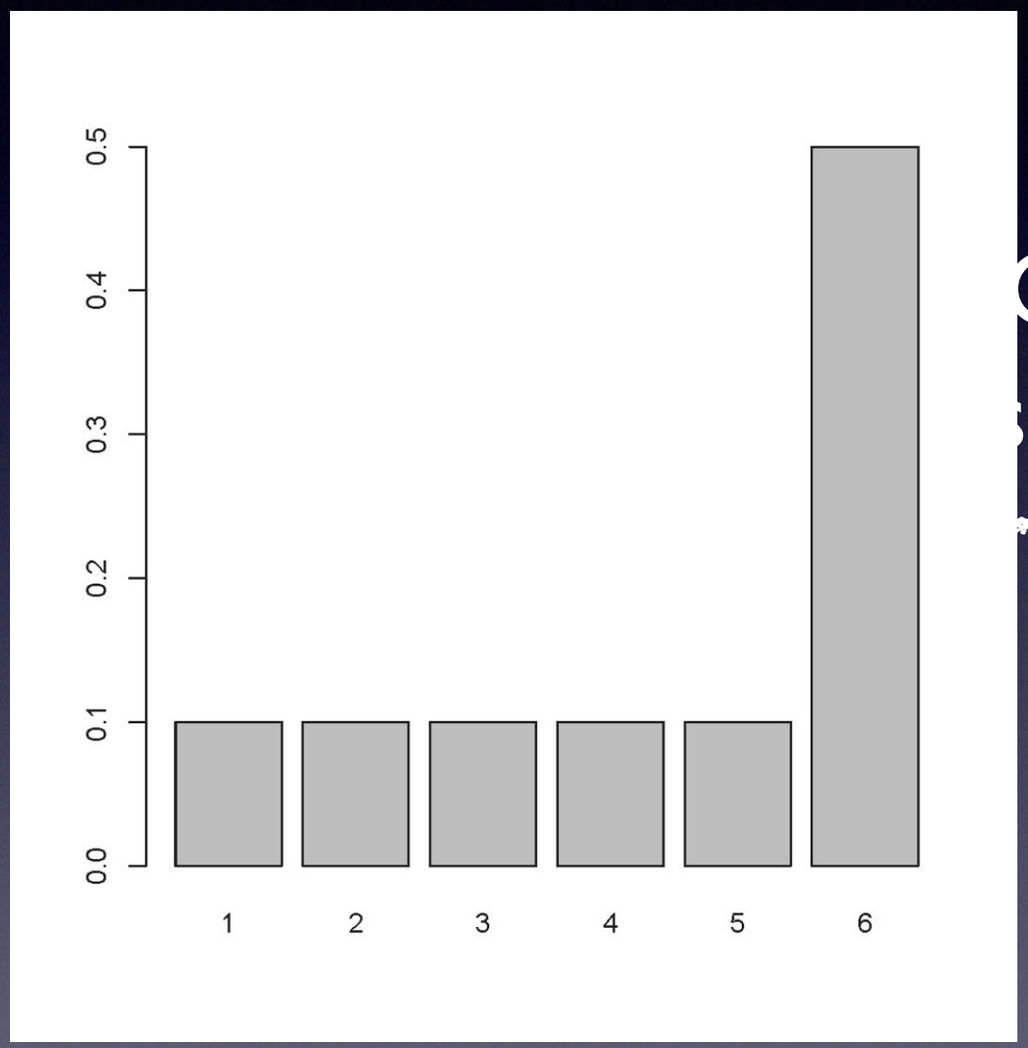


Good
escription

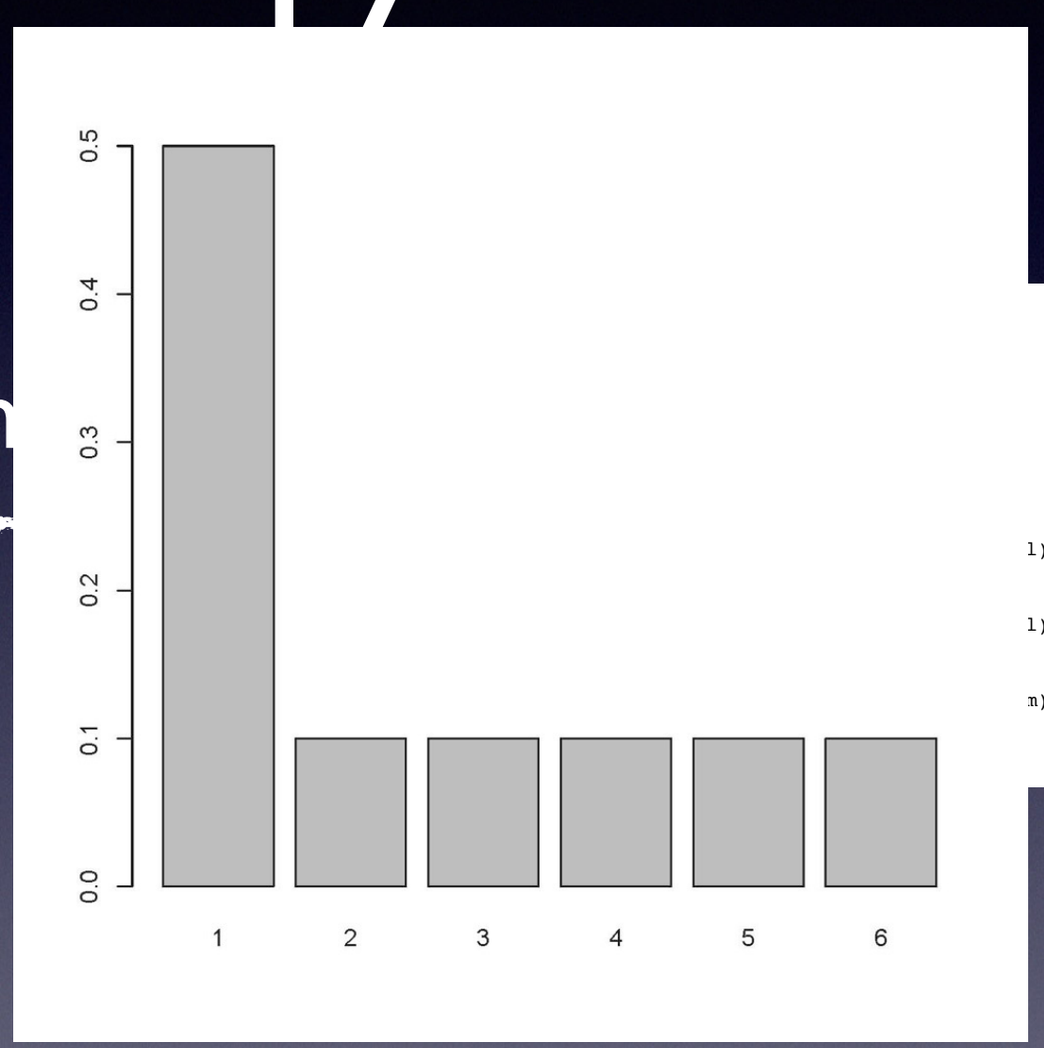


Low Cross Entropy!!

Background
Cross Entropy



Good
scription



High Cross Entropy!!

Measuring Goo
Cross entro

Higher if Model assigns Low-Probability to frequent events

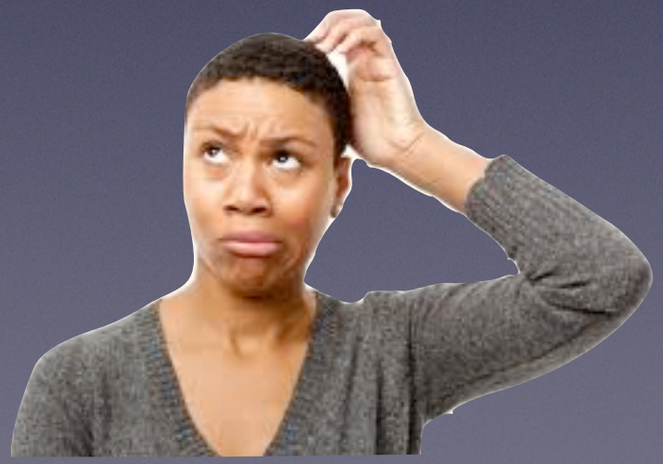
Cross entropy $\frac{1}{n}\sum_{i=1}^{n} -log(p(e_i))$

For a docu with n worc

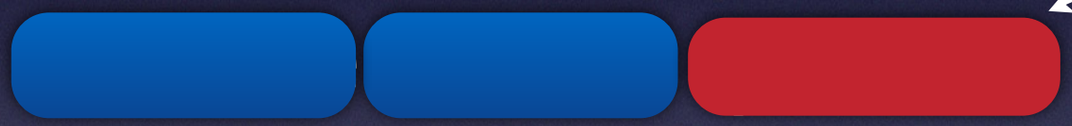
probability assigned by Model to word

Lower if Model assigns High-Probability to frequent events

What language model gives low cross-entropy?



n-gram models

- Intuition: Local Context Helps
- Examples (NL, then code)
  - 
  - 

What is This?

What is This?

More context helps more!

n-gram models of code: Experimental Results

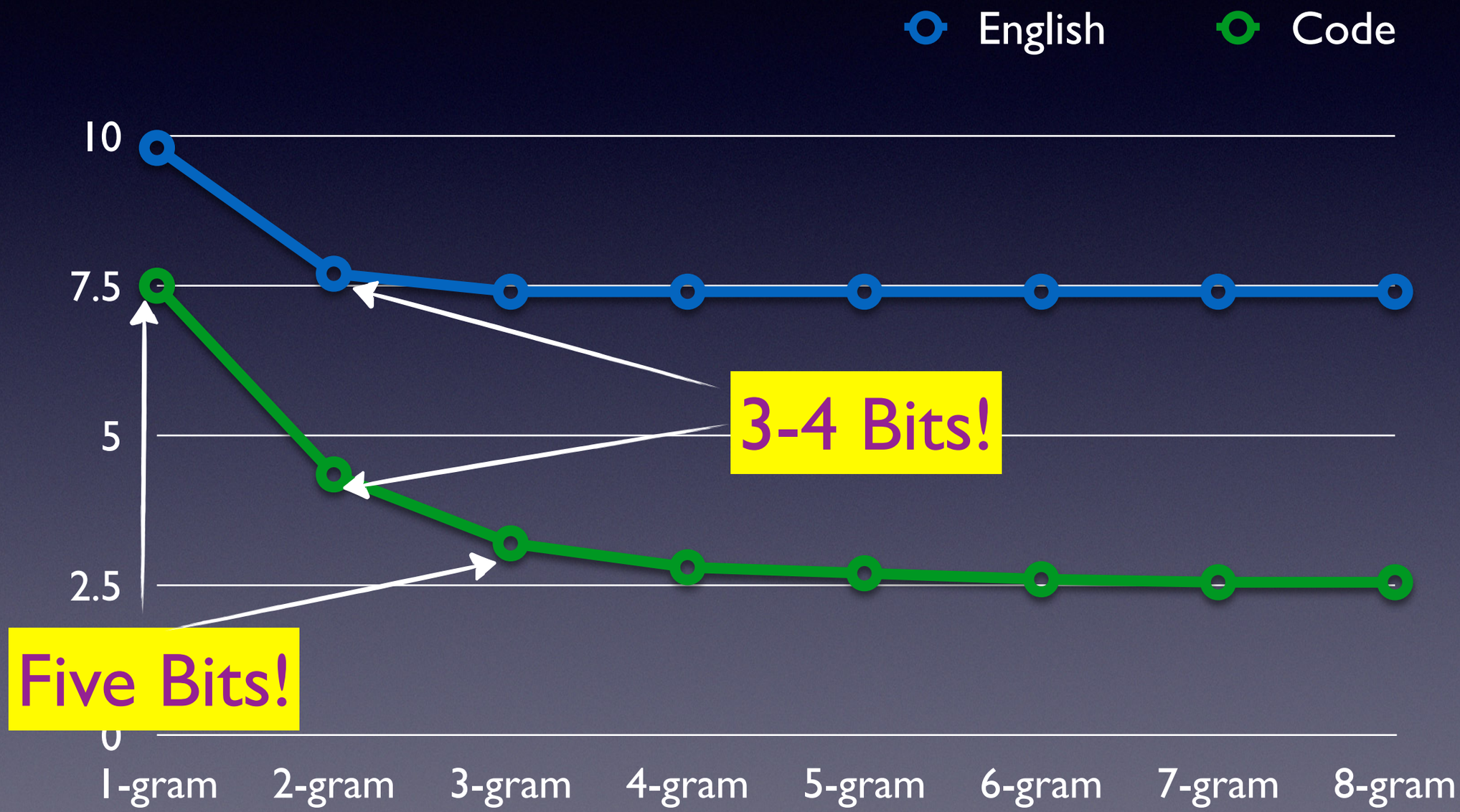
Java Datasets

Java Project	Version	Lines	Tokens	
			Total	Unique
Ant	20110123	254457	919148	27008
Batik	20110118	367293	1384554	30298
Cassandra	20110122	135992	697498	13002
Eclipse-E4	20110426	1543206	6807301	98652
Log4J	20101119	68528	247001	8056
Lucene	20100319	429957	2130349	32676
Maven2	20101118	61622	263831	7637
Maven3	20110122	114527	462397	10839
Xalan-J	20091212	349837	1085022	39383
Xerces	20110111	257572	992623	19542

C Datasets

Ubuntu Domain	Version	Lines	Tokens	
			Total	Unique
Admin (116)	10.10	9092325	41208531	1140555
Doc (22)	10.10	87192	362501	15373
Graphics (21)	10.10	1422514	7453031	188792
Interp. (23)	10.10	1416361	6388351	201538
Mail (15)	10.10	1049136	4408776	137324
Net (86)	10.10	5012473	20666917	541896
Sound (26)	10.10	1698584	29310969	436377
Tex (135)	10.10	1405674	14342943	375845
Text (118)	10.10	1325700	6291804	155177
Web (31)	10.10	1743376	11361332	216474

N-gram Cross Entropy



The Skeptic Asks...

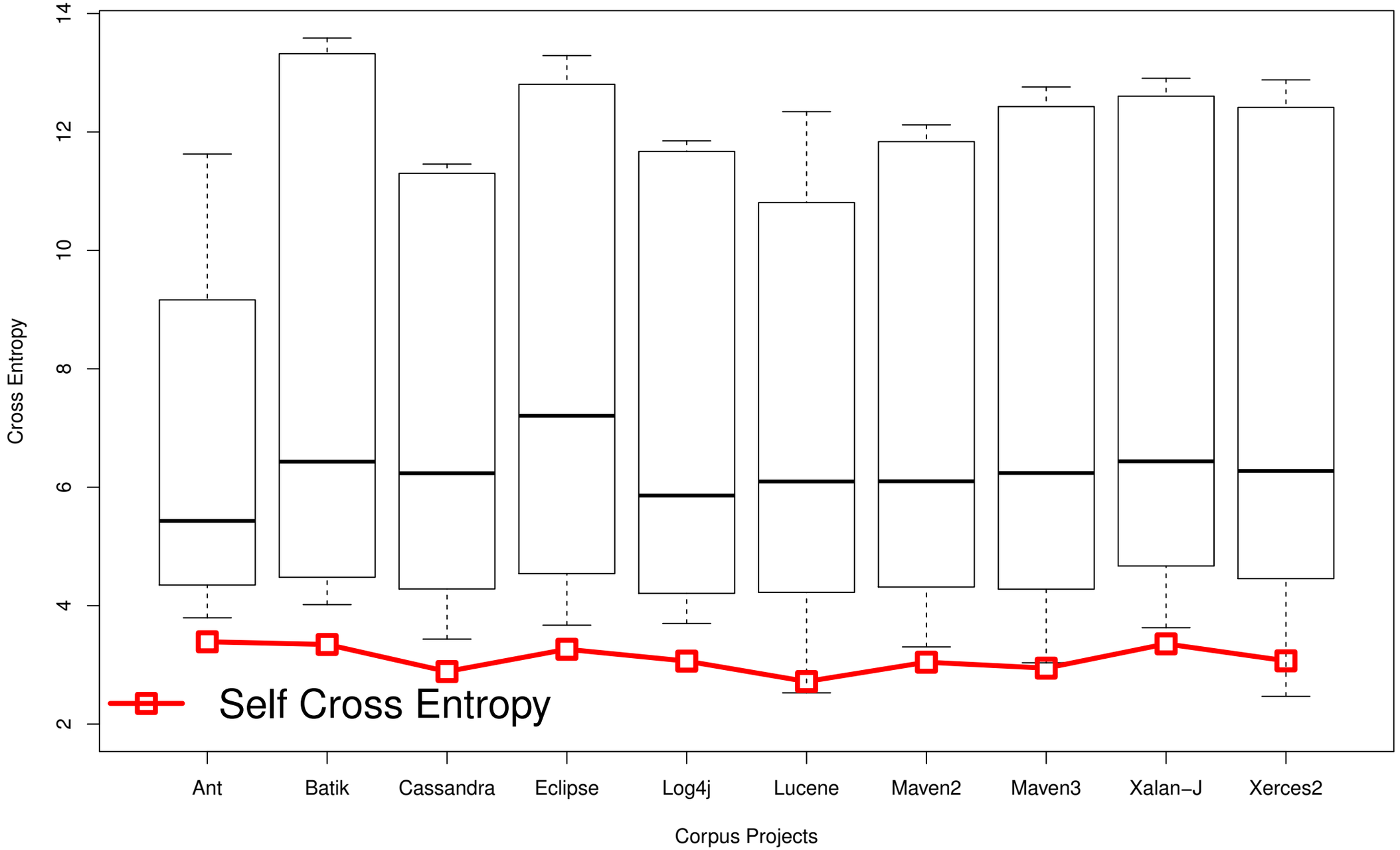
Is it just that C, Java, Python... are simpler than English?

➡ Do cross-project testing!

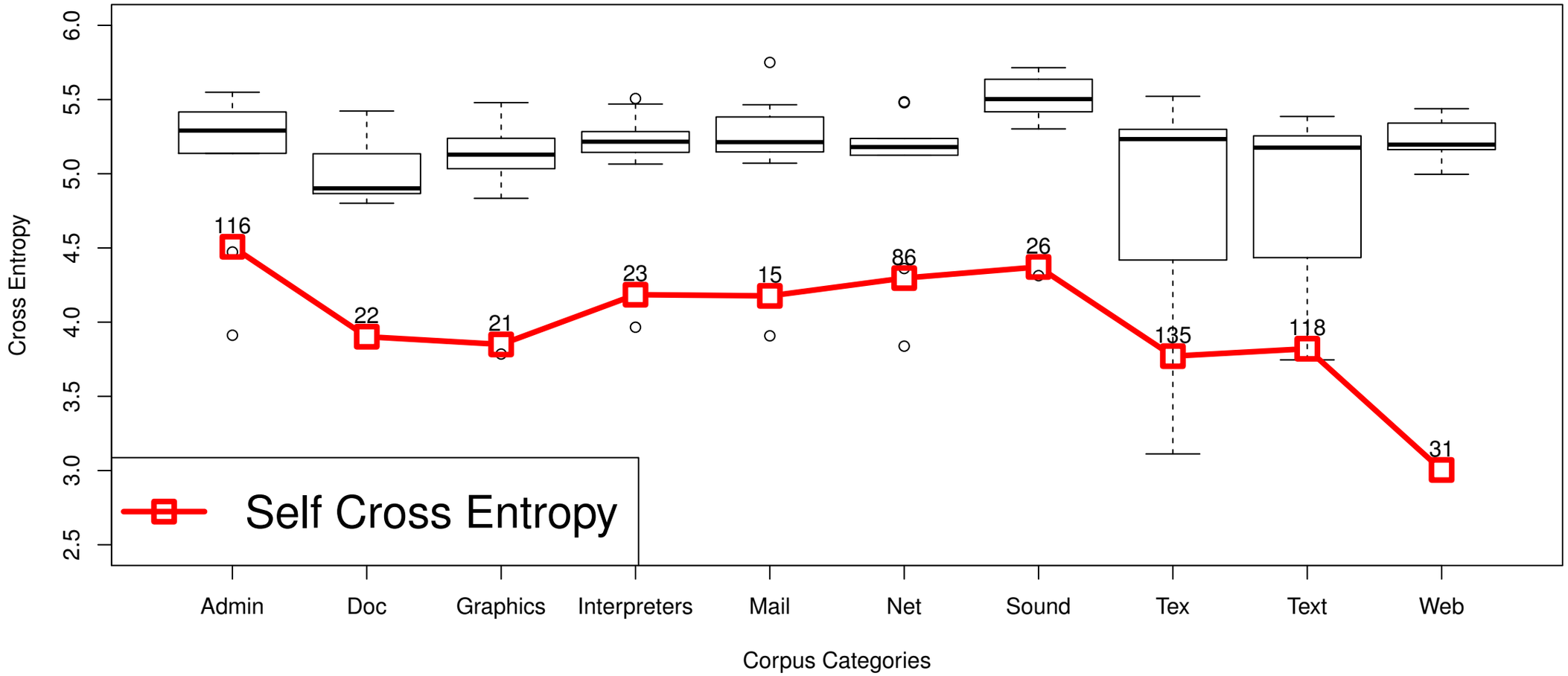
➡ Train on one project, test on the others

➡ If it's all "in the language", entropy should be similar

Train on one project, test on the others.


Cross Entropy
14
12
10
8
6
4
2
Self Cross Entropy
Ant
Batik
Cassandra
Eclipse
Log4j
Lucene
Maven2
Maven3
Xalan−J
Xerces2
Corpus Projects

Train on one Ubuntu application domain,
test on the others.


Cross Entropy
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
116
22
21
23
15
86
26
135
118
31
Self Cross Entropy
Admin
Doc
Graphics
Interpreters
Mail
Net
Sound
Tex
Text
Web
Corpus Categories

The "Naturalness" Vision

Suggest the next token for developers

Complete the current token for developers

Assistive (speech, gesture) coding

Summarization and retrieval as translation

Stupid, statistical, static analysis

Search-based Software Engineering

Uses Type, Scope, Etc !

esting Tokens

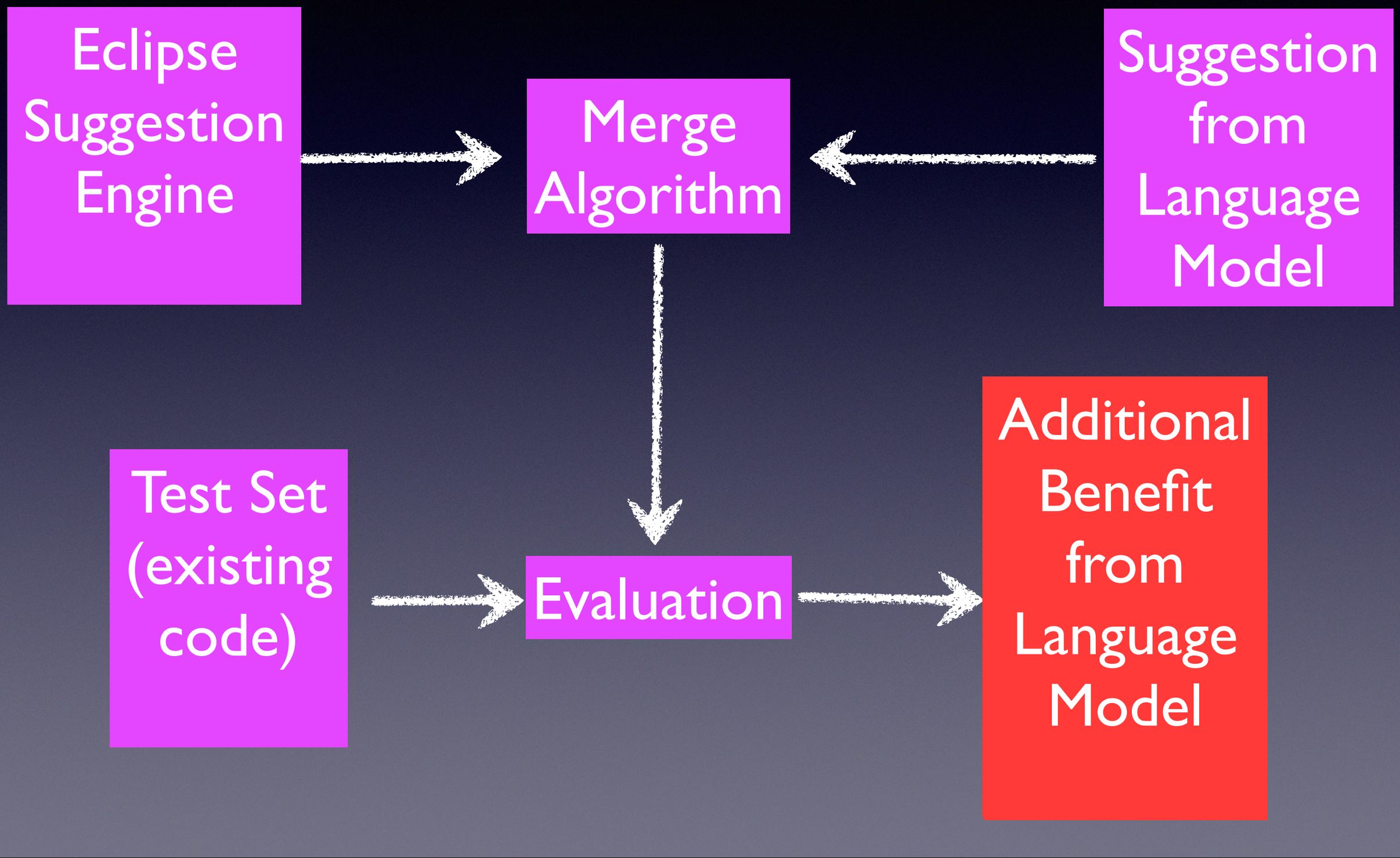
What token *could* appear here?

Visual Studio

ken *has most often* eared he

Use just previous two tokens!

Do n-grams help?



How many more
correct suggestions?

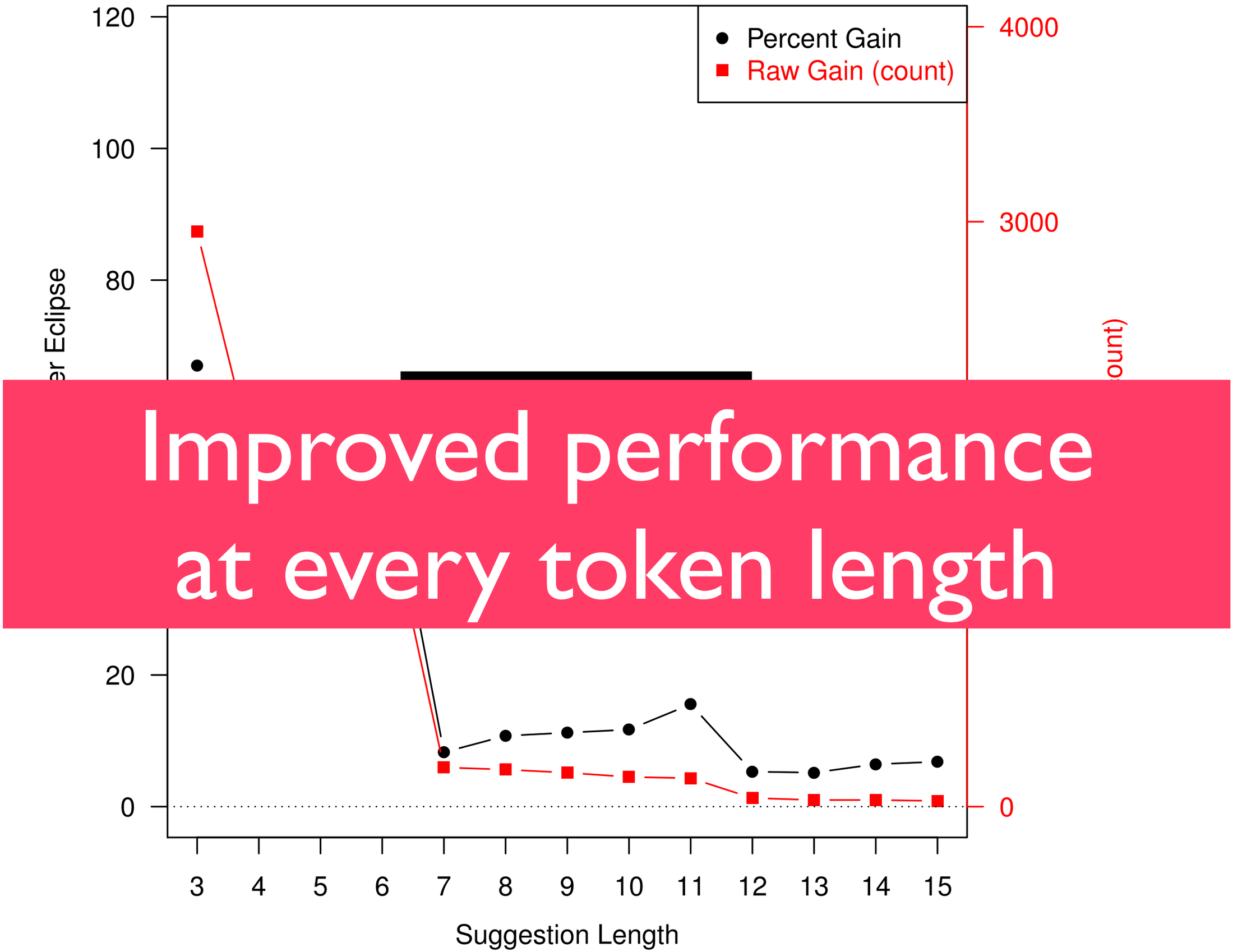
Suggestion I

Suggestion6

Suggestion9

Suggestion10

Language Models
ALWAYS
improve performance



Improved performance at every token length

N-Gram suggestions

always add value to

the native Eclipse suggestion engine,

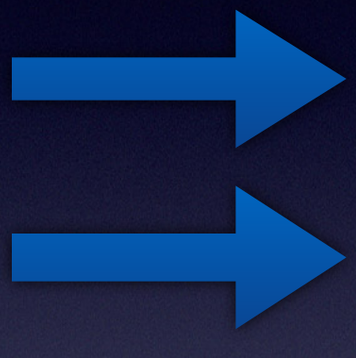
in a *very large* trial.

Can be rich, powerful, expressive

Mostly simple, repetitive, boring

Statistical Models

The "Naturalness" Vision



Suggest the next token for developers

Complete the current token for developers

Assistive (speech, gesture) coding

?????

Summarization and retrieval as translation

?????

Fast, "good guess" static analysis

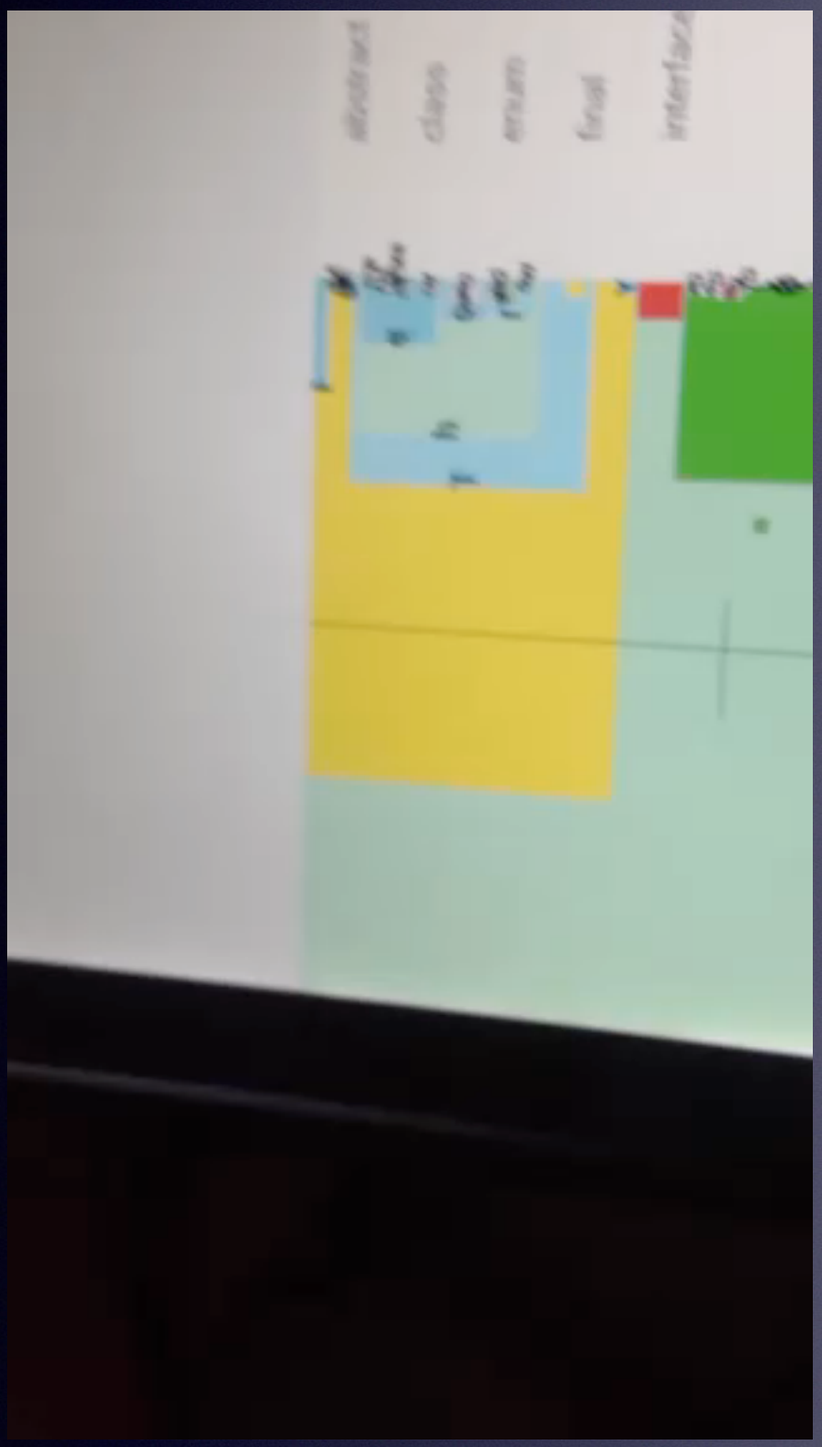
Search-based Software Engineering

Assisted Coding

Rachel Aurand (Graduate Student)

Dasher++

Eclipse


abstract
class
enum
final
interface

The "Naturalness" Vision

Suggest the next token for developers

Complete the current token for developers

Assistive (speech, gesture) coding

Summarization and retrieval as translation

Fast, "good guess" static analysis

Search-based Software Engineering

Noisy Channel Model

"Comment allez vous? "

May be he's saying:
"Do you comment all your code?"

$$p(E \mid F) = \frac{p(F \mid E).p(E)}{p(F)}$$

Most Likely way "it got messed up"

Most Likely English Sentence

$$p(E \mid F) = \frac{p(F \mid E).p(E)}{p(F)}$$

Maximize Numerator over "E" to get best translation

Normalizing Constant

Joint Distribution from Aligned Corpus

English Language Model

$$p(E \mid F) = \frac{p(F \mid E).p(E)}{p(F)}$$

Where do the probability distributions come from?

Normalizing Constant

Noisy Channel Model

```
Toast.makeText(context,
  "hello", 5).show();
```

He's trying to speak English, but it

Maybe his code means
"Make me some toast?"

$$p(E \mid C) = \frac{p(C \mid E).p(E)}{p(C)}$$

Code-English Joint Corpus

"Domain-Specific" English Language Model

$$p(E \mid C) = \frac{p(C \mid E).p(E)}{p(C)}$$

Where do the probability distributions come from?

Normalizing Constant

The "Naturalness" Vision

Suggest or Complete next tokens

Assistive (speech, gesture) coding

Summarization and Retrieval as Translation

Learn and Enforce Coding Conventions

Syntax Errors

Machine Translation for Porting

Fast, "good guess" static analysis

Search-based Software Engineering