

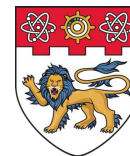
Translating Source Code into Natural Language with AI

Dr. Misha Filippov
Chief Scientist at Quod AI

Dr. Misha Filippov
Chief Scientist & Co-founder at
Quod AI




- ▶ **Mathematical physicist**
- ▶ PhD in Physics, **NTU** (Nanyang Technological University)
- ▶ Honorary research fellow, **UCL** (University College London)
- ▶ **IEEE/ICCC** award winning **NLP** paper
- ▶ Built AI models for **NASA** & **Monetary Authority of SG**



**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

**ÉCOLE
POLYTECHNIQUE**
UNIVERSITÉ PARIS-SACLAY

A photograph of two software engineers working in a modern office. They are seated at desks with multiple computer monitors displaying code. The scene is dimly lit with blue ambient lighting. A semi-transparent dark blue box is overlaid on the center of the image, containing white text.

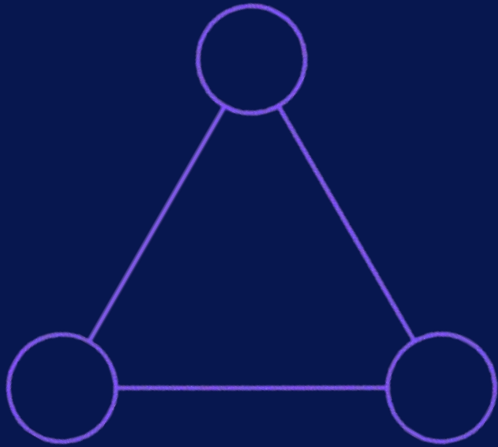
How much time do
software engineers
code?

5 to 28% of time is spent on **coding**

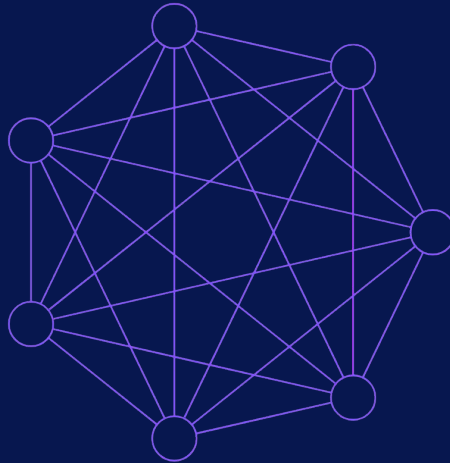
1. Source: [Characterizing Software Engineering Work with Personas Based on Knowledge Worker Actions](#), Microsoft & UNC (2017)

2. Source: [The Developer Coefficient](#), Stripe (2018)

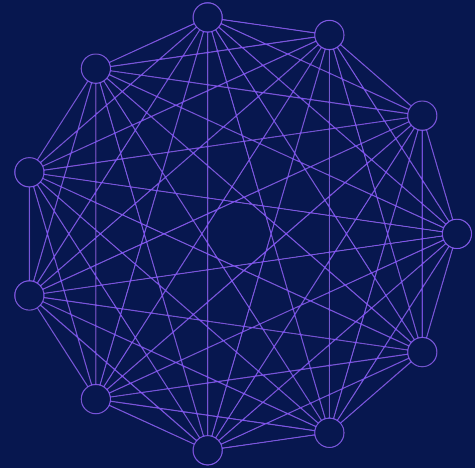
“Communication overheads increase as the number of people increases” - Brook's law ¹



3 PEOPLE
3 lines



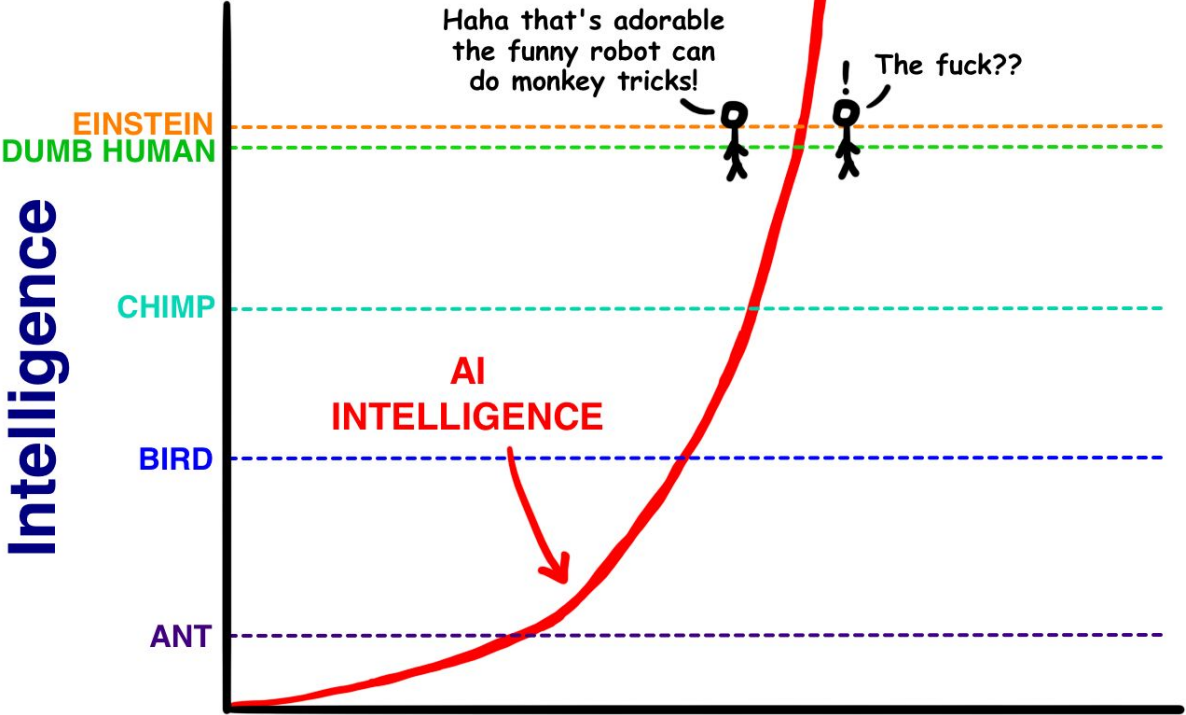
7 PEOPLE
21 lines



11 PEOPLE
55 lines



Reality



Time

Code completion^[1]

[1] [Aroma: Code Recommendation via Structural Code Search](#), [Code Completion with Neural Attention and Pointer Networks](#)

Code completion_[1]

Code optimization_[2]

[1] [Aroma: Code Recommendation via Structural Code Search, Code Completion with Neural Attention and Pointer Networks](#)

[2] [Learning to superoptimize programs: The Case for Learned Index Structures](#)

Code completion_[1]

Code optimization_[2]

Automated code review_[3]

[1] [Aroma: Code Recommendation via Structural Code Search](#), [Code Completion with Neural Attention and Pointer Networks](#)

[2] [Learning to superoptimize programs](#), [The Case for Learned Index Structures](#)

[3] [Intelligent Code Reviews Using Deep Learning](#)

Code completion_[1]

Code optimization_[2]

Automated code review_[3]

Bug detection and automated bug fixing_[4]

[1] [Aroma: Code Recommendation via Structural Code Search](#), [Code Completion with Neural Attention and Pointer Networks](#)

[2] [Learning to superoptimize programs](#), [The Case for Learned Index Structures](#)

[3] [Intelligent Code Reviews Using Deep Learning](#)

[4] [Learning How to Mutate Source Code from Bug-Fixes](#), [DeepBugs: A Learning Approach to Name-based Bug Detection](#)

Code completion_[1]

Code optimization_[2]

Automated code review_[3]

Bug detection and automated bug fixing_[4]

Documentation generation

[1] [Aroma: Code Recommendation via Structural Code Search](#), [Code Completion with Neural Attention and Pointer Networks](#)

[2] [Learning to superoptimize programs](#), [The Case for Learned Index Structures](#)

[3] [Intelligent Code Reviews Using Deep Learning](#)

[4] [Learning How to Mutate Source Code from Bug-Fixes](#), [DeepBugs: A Learning Approach to Name-based Bug Detection](#)

>> { } + [] === [] + { }

```
>> { } + [ ] === [ ] + { }  
< false
```

```
>> { } + [ ] === [ ] + { }
```

```
< false
```

```
>> [ ] + { } === { } + [ ]
```

```
>> { } + [ ] === [ ] + { }
```

```
< false
```

```
>> [ ] + { } === { } + [ ]
```

```
< true
```


What is code?

Code is:

- a sequence of bytes
- a sequence of characters
- a sequence of tokens
- a tree

Code != text

Good:

Regularity comes from the language

definition

Language syntax

Language semantics

Bad:

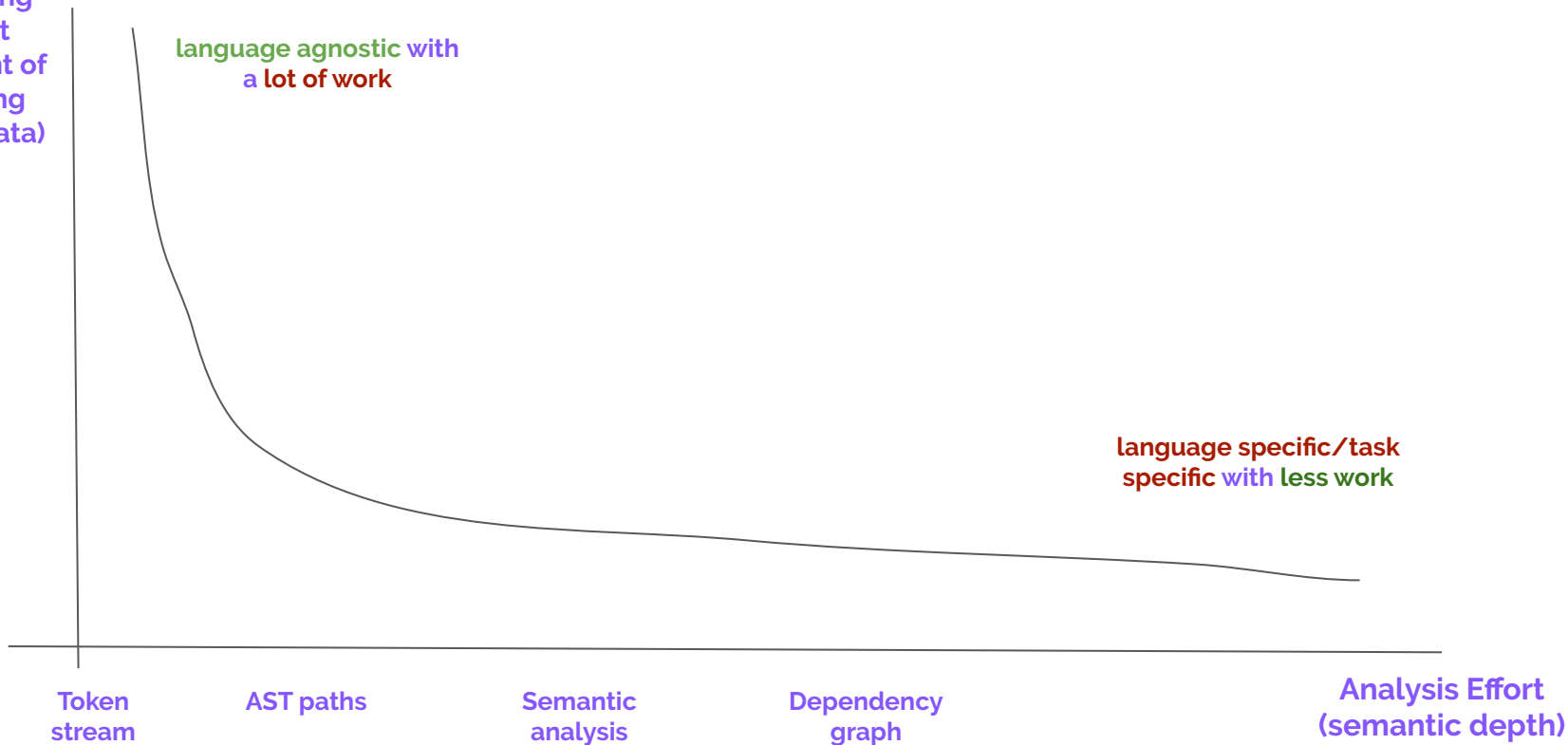
- Re-learning
- Expressing language knowledge
 - Requires program analysis expertise
 - Some aspects have to be repeated per language

How to find the good simplification level?

Learning Effort
(amount of training time, data)

language agnostic with a lot of work

language specific/task specific with less work



Unsupervised approach

Tensorflow, NumPy, PyTorch, Keras

Tree based LSTM + syntax aware tokenization

300M lines of code from GitHub

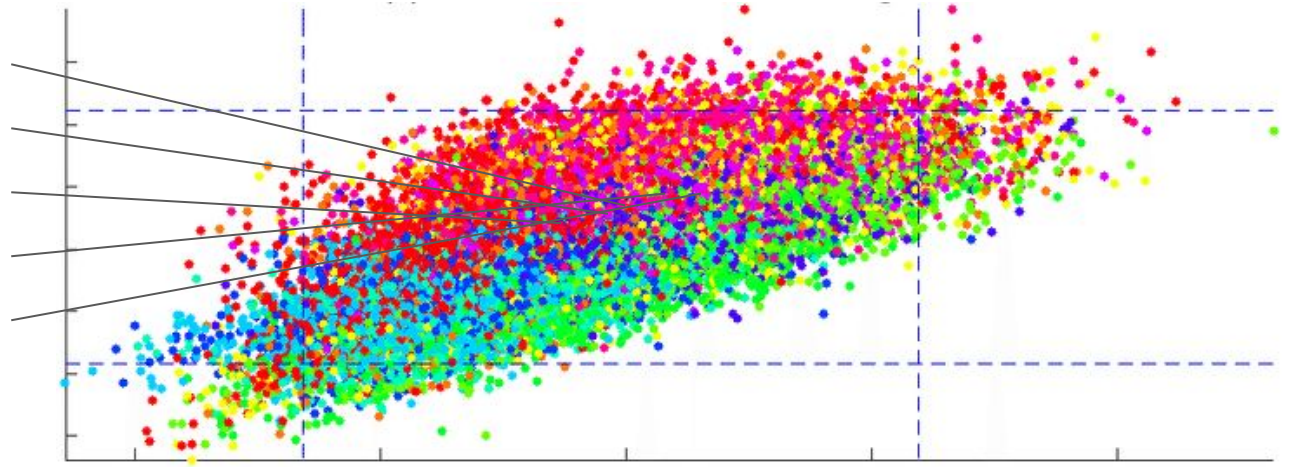
2 hours of training on 16 CPUs + 2 GPUs



Quod AI

The first AI platform that explains source code in plain English

- > CreateArticle
- > create_article
- > MakeArticle
- > CreateBlogPost
- > post_init



Evaluation

BLEU & ROUGE

F1 score over sub-tokens, case-insensitive

Evaluation

BLEU & ROUGE

F1 score over sub-tokens, case-insensitive

False Positives vs False Negatives

Evaluation

BLEU & ROUGE

F1 score over sub-tokens, case-insensitive

False Positives vs False Negatives

Relevance score (satisfaction metric)

Problems with unsupervised approach

“Get free space on internal memory”

```
File path = Environment.getDataDirectory();
StatFs stat = new StatFs(path.getPath());
long blockSize = stat.getBlockSize();
long availableBlocks = stat.getAvailableBlocks();
return Formatter.formatFileSize(this, availableBlocks * blockSize);
```

Problems with unsupervised approach

“Get free space on internal memory”

```
File path = Environment.getDataDirectory();
StatFs stat = new StatFs(path.getPath());
long blockSize = stat.getBlockSize();
long availableBlocks = stat.getAvailableBlocks();
return Formatter.formatFileSize(this, availableBlocks * blockSize);
```

Labeling freedom should be based on a given taxonomy

Supervised approach. Details + demo

Tensorflow, NumPy,
PyTorch, fastText, Keras

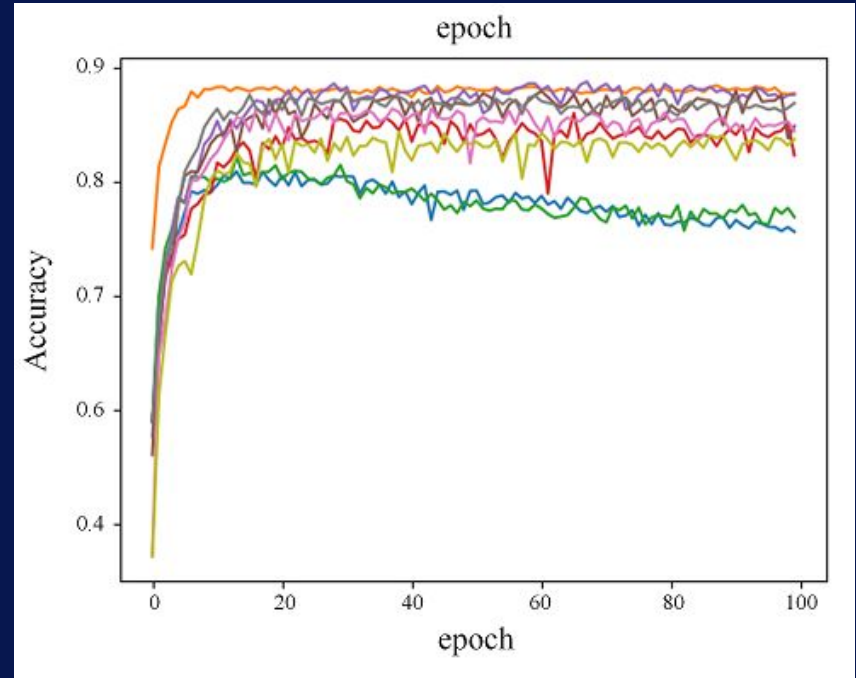
RNNs + CNNs + LSTMs +
attention

Supervised approach. Details + demo

Tensorflow, NumPy,
PyTorch, fastText, Keras

RNNs + CNNs + LSTMs +
attention

First stable results after 10
first data points



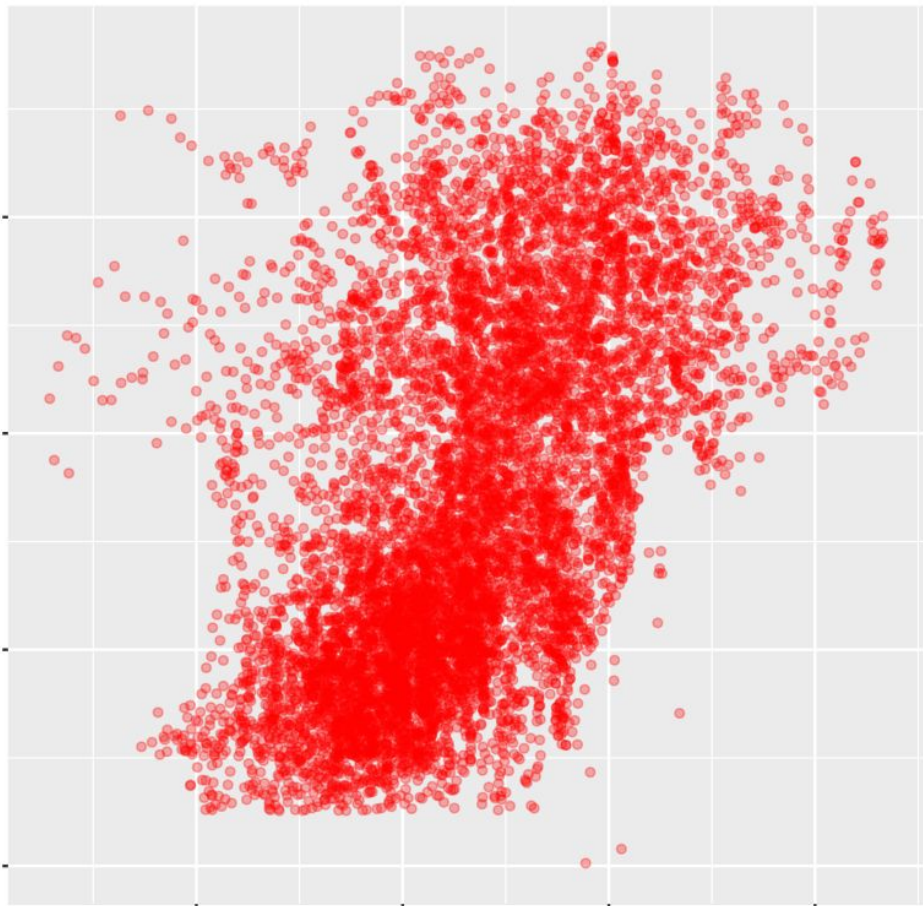
Supervised approach. Details + demo

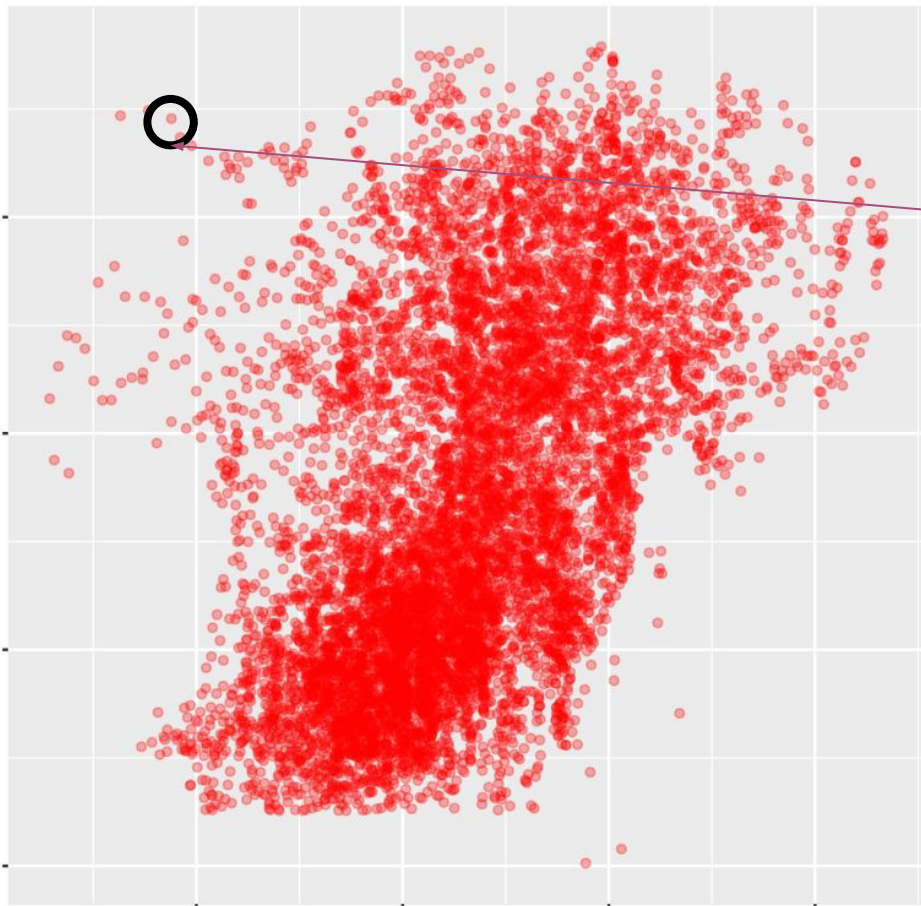
What are security related components?

Supervised approach. Details + demo

What are security related components?

```
session_set_cookie_params(0, '/', '', true, true); session_start();
```





```
React.__SECRET_DOM_DO_NOT_USE_OR_YOU  
_WILL_BE_FIRED = ReactDOM;
```



“We’re on a mission to fundamentally **rethink** & **retool** developer toolchain.”

Erik Meijer, Director of Engineering at Facebook.

Facebook started integrating AI into developer tools in 2017.

- Probabilistic programming techniques
- Suite of AI-driven developer tools
- Query semantic information
- Auto patching
- Neural code search





Thank you
..and we are hiring!

Dr. Misha Filippov
misha@quod.ai