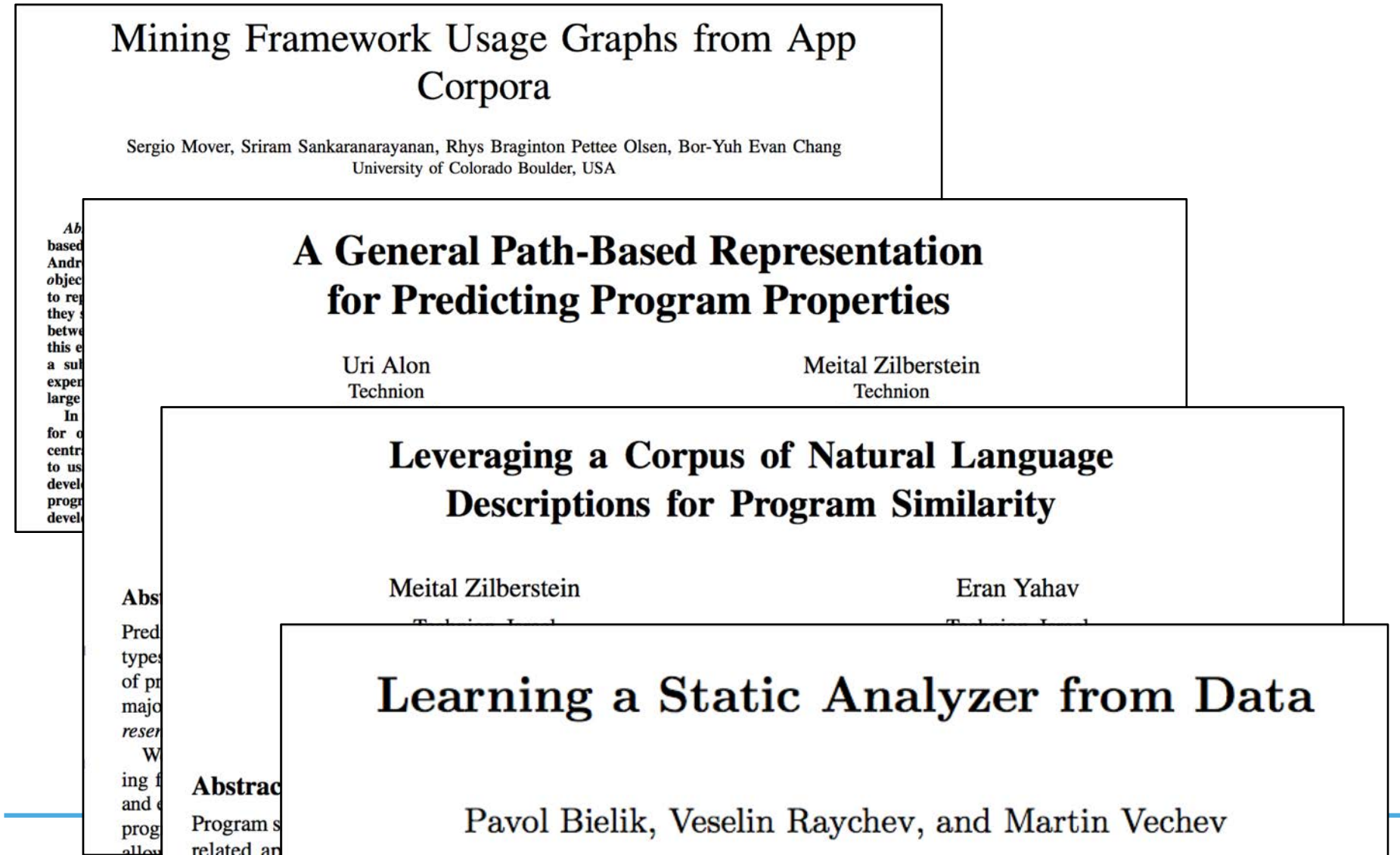
Code + ML:
Will automation take our jobs?

Stephen Magill

CEO, Muse Dev

Principal Scientist, Galois

ML + Code



Mining Framework Usage Graphs from App Corpora

Sergio Mover, Sriram Sankaranarayanan, Rhys Braginton Pettee Olsen, Bor-Yuh Evan Chang
University of Colorado Boulder, USA

A General Path-Based Representation for Predicting Program Properties

Uri Alon
Technion

Meital Zilberstein
Technion

Leveraging a Corpus of Natural Language Descriptions for Program Similarity

Meital Zilberstein

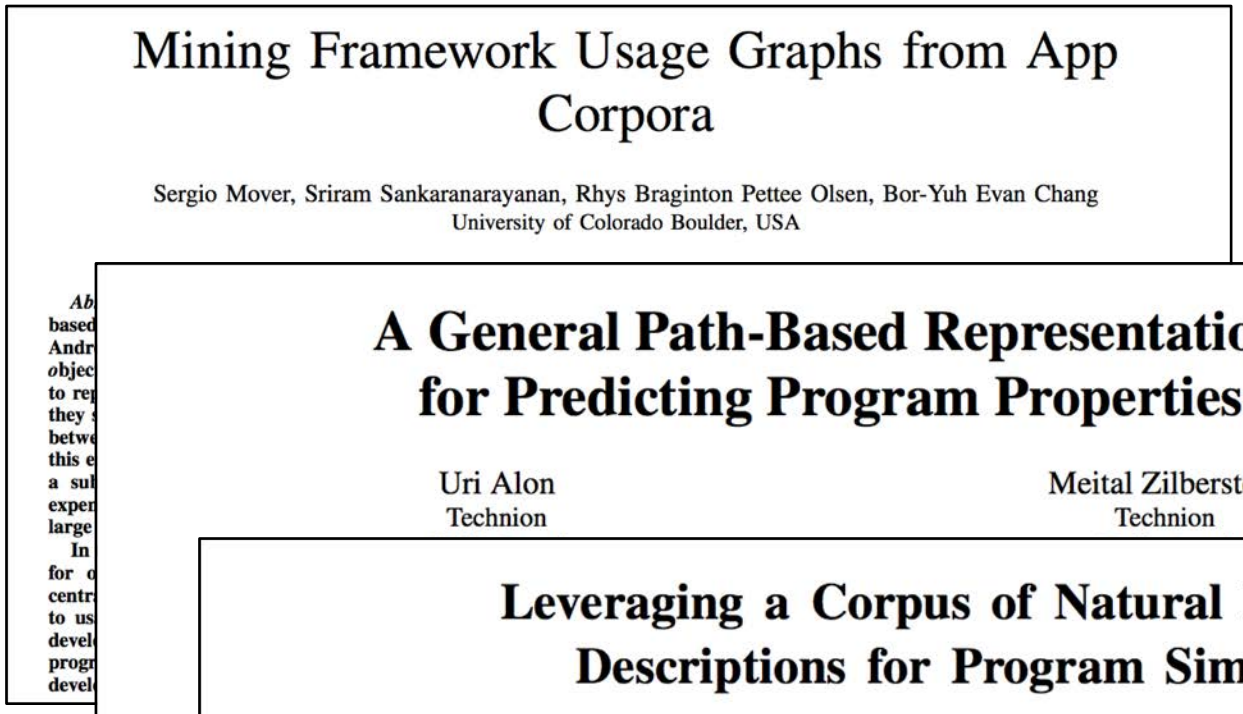
Eran Yahav

Learning a Static Analyzer from Data

Pavol Bielik, Veselin Raychev, and Martin Vechev

muse dev

ML + Code



Mining Framework Usage Graphs from App Corpora

Sergio Mover, Sriram Sankaranarayanan, Rhys Braginton Pettee Olsen, Bor-Yuh Evan Chang
University of Colorado Boulder, USA

A General Path-Based Representation for Predicting Program Properties

Uri Alon
Technion

Meital Zilberstein
Technion

Leveraging a Corpus of Natural Language Descriptions for Program Similarity

Meital Zilberstein

Eran Yahav

Learning a Static Analyzer from Data

Pavol Bielik, Veselin Raychev, and Martin Vechev

prodo.ai

codota

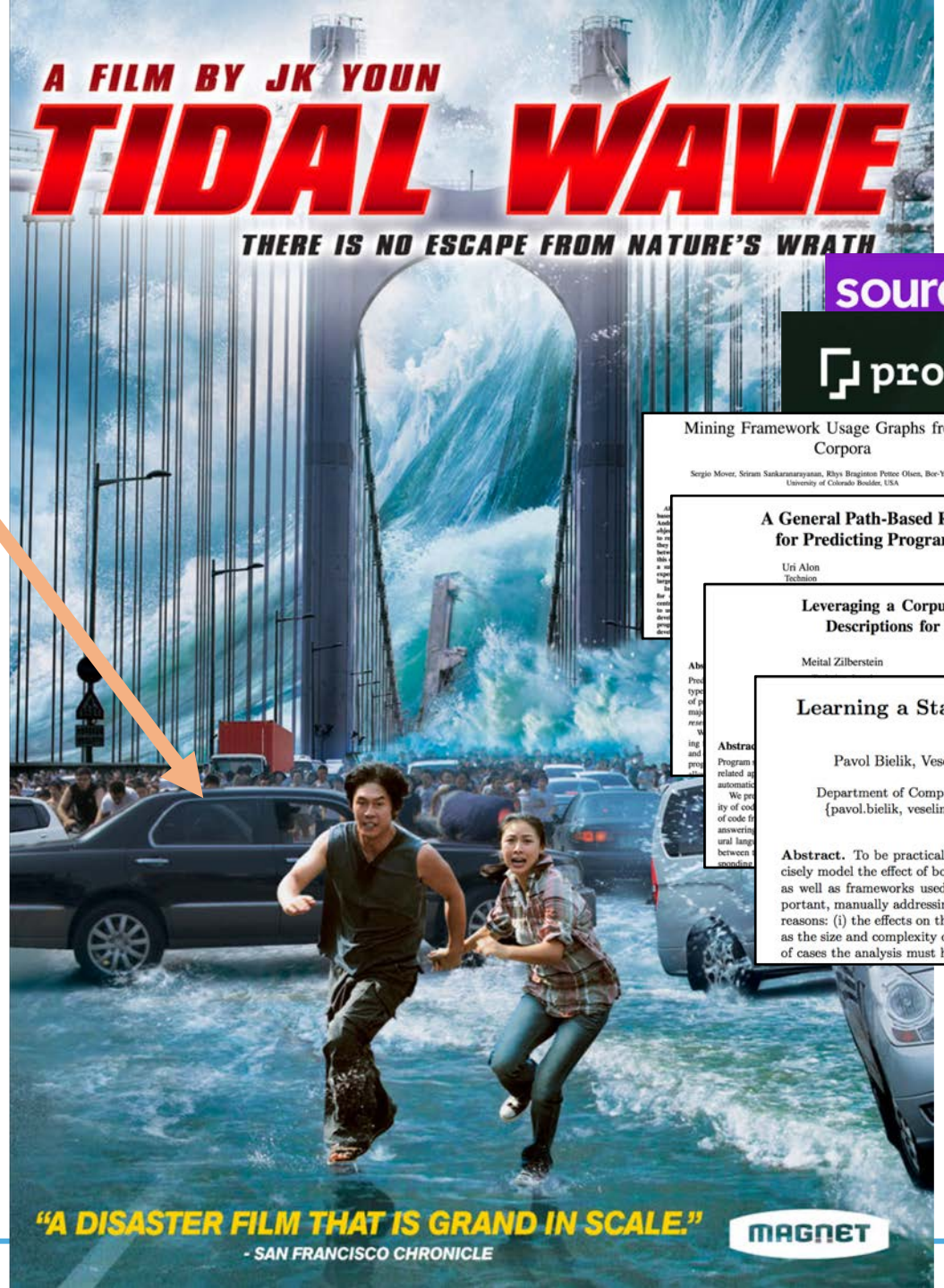
diffblue

DEEPCODE

source{d}

muse dev

Developers?

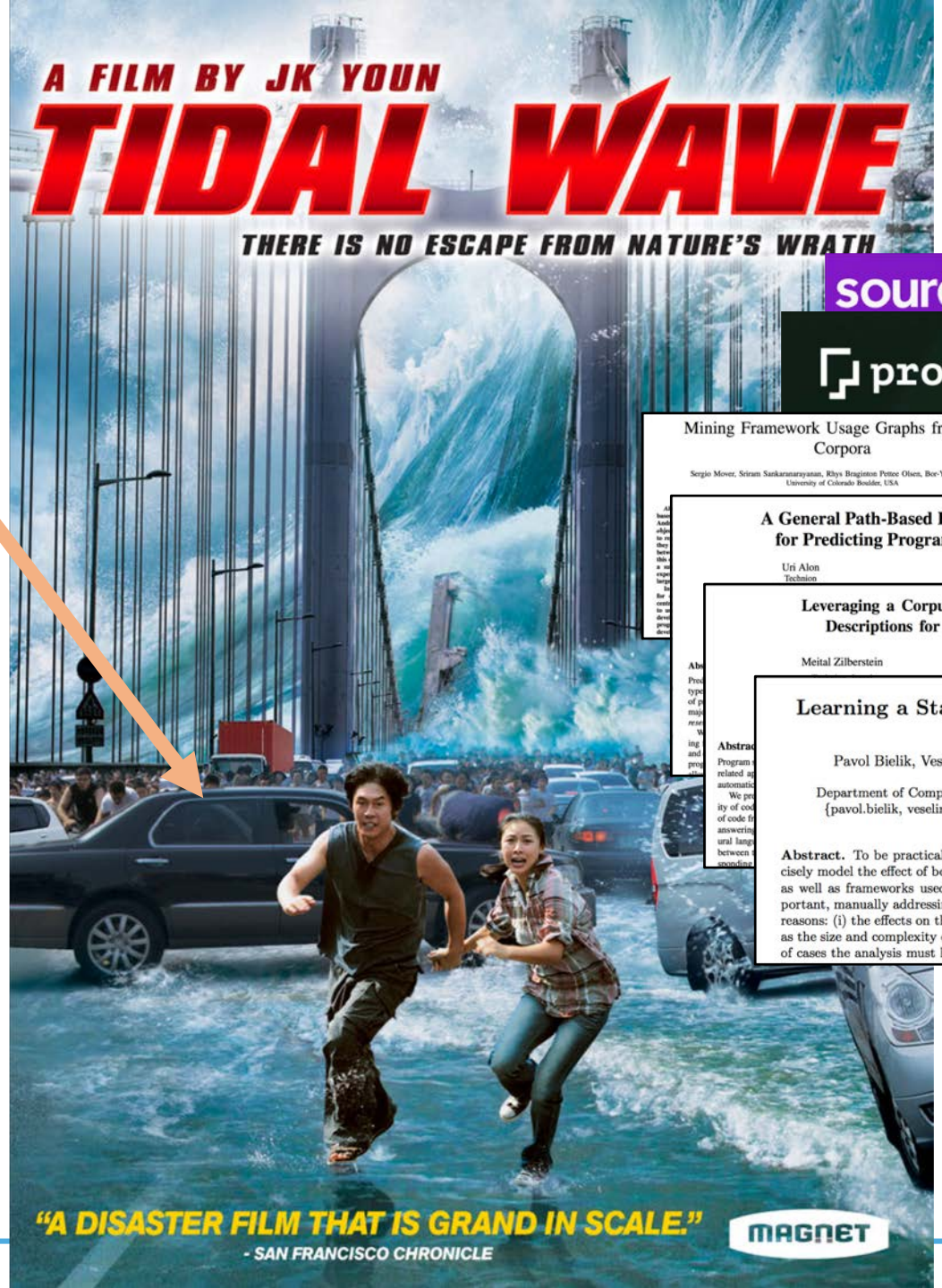

A FILM BY JK YOUN
TIDAL WAVE
THERE IS NO ESCAPE FROM NATURE'S WRATH
"A DISASTER FILM THAT IS GRAND IN SCALE."
- SAN FRANCISCO CHRONICLE
MAGNET

Mining Framework Usage Graphs from App Corpora
A General Path-Based Representation for Predicting Program Properties
Uri Alon
Technion
Meital Zilberstein
Technion
Leveraging a Corpus of Natural Language Descriptions for Program Similarity
Meital Zilberstein
Eran Yahav
Learning a Static Analyzer from Data
Pavol Bielik, Veselin Raychev, and Martin Vechev
Department of Computer Science, ETH Zürich, Switzerland
{pavol.bielik, veselin.raychev, martin.vechev}@inf.ethz.ch
Abstract. To be practically useful, modern static analyzers must precisely model the effect of both, statements in the programming language as well as frameworks used by the program under analysis. While important, manually addressing these challenges is difficult for at least two reasons: (i) the effects on the overall analysis can be non-trivial, and (ii) as the size and complexity of modern libraries increase, so is the number of cases the analysis must handle.

sourced
codota
prodo.ai
P CODE
ffblue

Developers?

… or developers?



sourced

codota

prodo.ai

P CODE

ffblue

Mining Framework Usage Graphs from App Corpora

Sergio Mover, Sriram Sankaranarayanan, Rhys Braginton Pettee Olsen, Bor-Yuh Evan Chang
University of Colorado Boulder, USA

A General Path-Based Representation for Predicting Program Properties

Uri Alon
Technion

Meital Zilberstein
Technion

Leveraging a Corpus of Natural Language Descriptions for Program Similarity

Meital Zilberstein

Eran Yahav

Learning a Static Analyzer from Data

Pavol Bielik, Veselin Raychev, and Martin Vechev

Department of Computer Science, ETH Zürich, Switzerland
{pavol.bielik, veselin.raychev, martin.vechev}@inf.ethz.ch

Abstract. To be practically useful, modern static analyzers must precisely model the effect of both, statements in the programming language as well as frameworks used by the program under analysis. While important, manually addressing these challenges is difficult for at least two reasons: (i) the effects on the overall analysis can be non-trivial, and (ii) as the size and complexity of modern libraries increase, so is the number of cases the analysis must handle.

muse dev

ML + Code

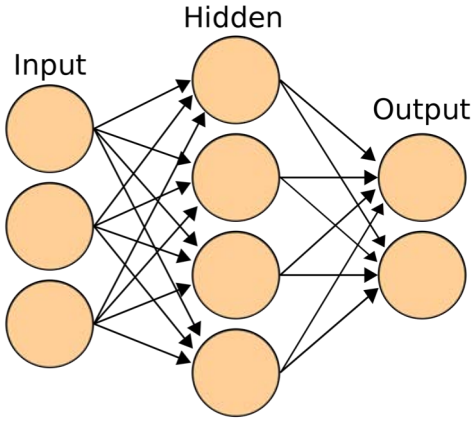
Code is written using programming *languages*…

…Look at natural language processing techniques

ML + Code

Code is written using programming *languages*…

…Look at natural language processing techniques

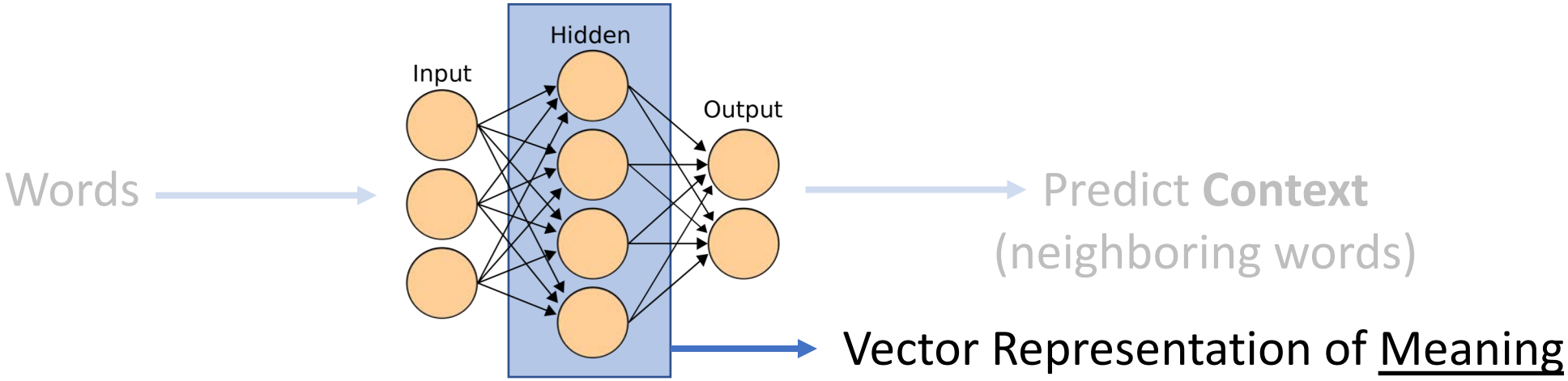
Words → 

→ Predict **Context** (neighboring words)

ML + Code

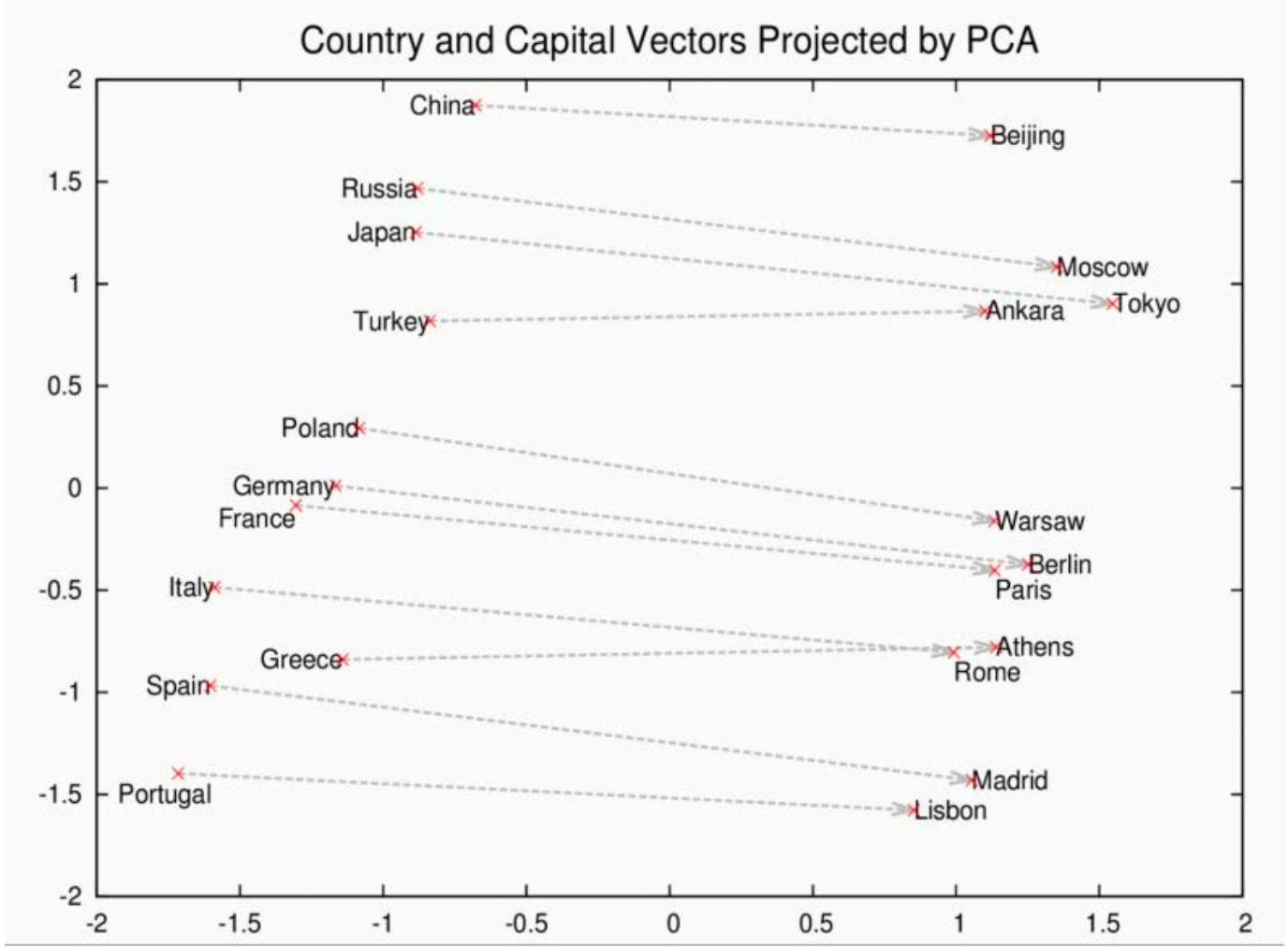
Code is written using programming *languages*…

…Look at natural language processing techniques



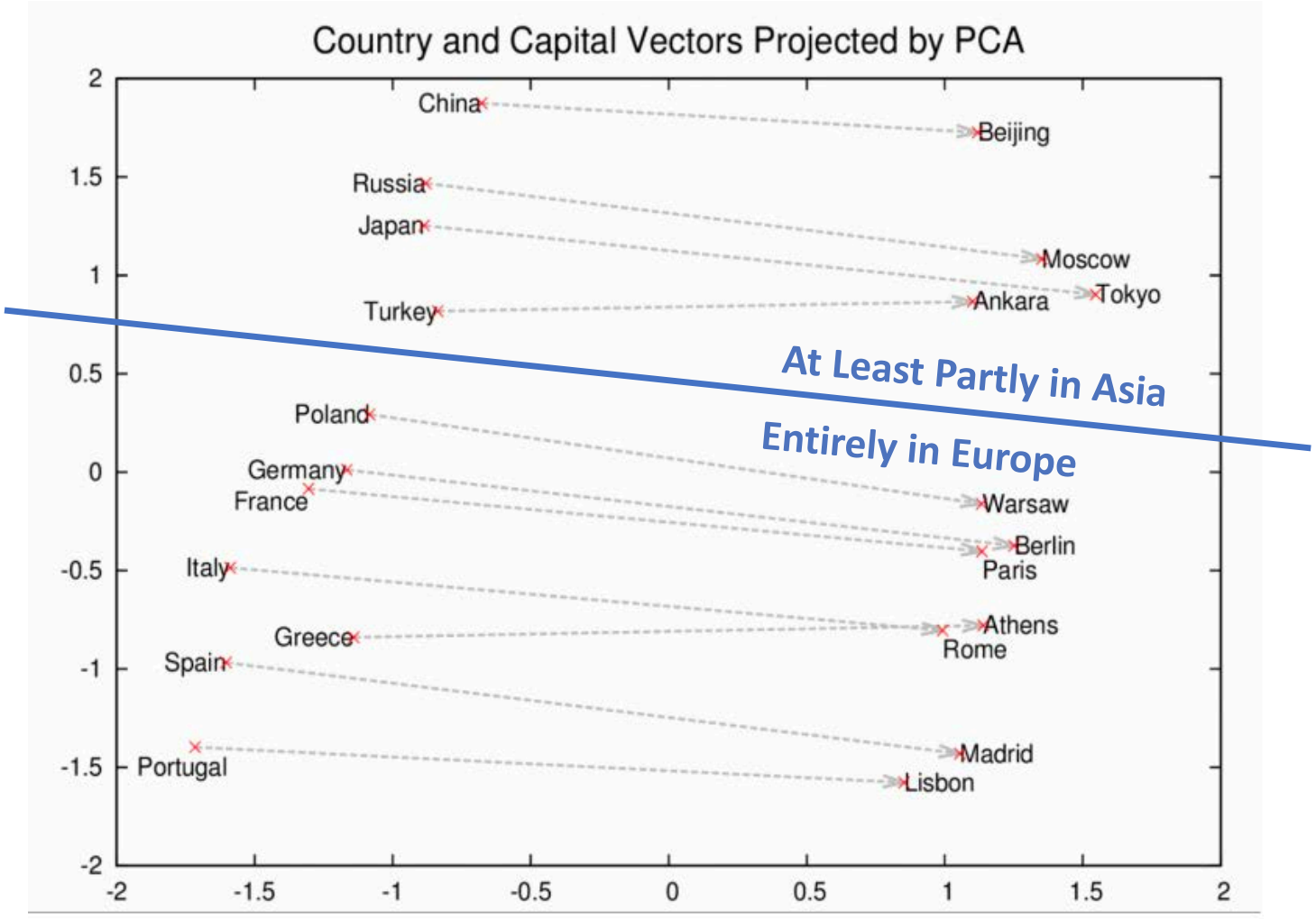
word2vec

by Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean



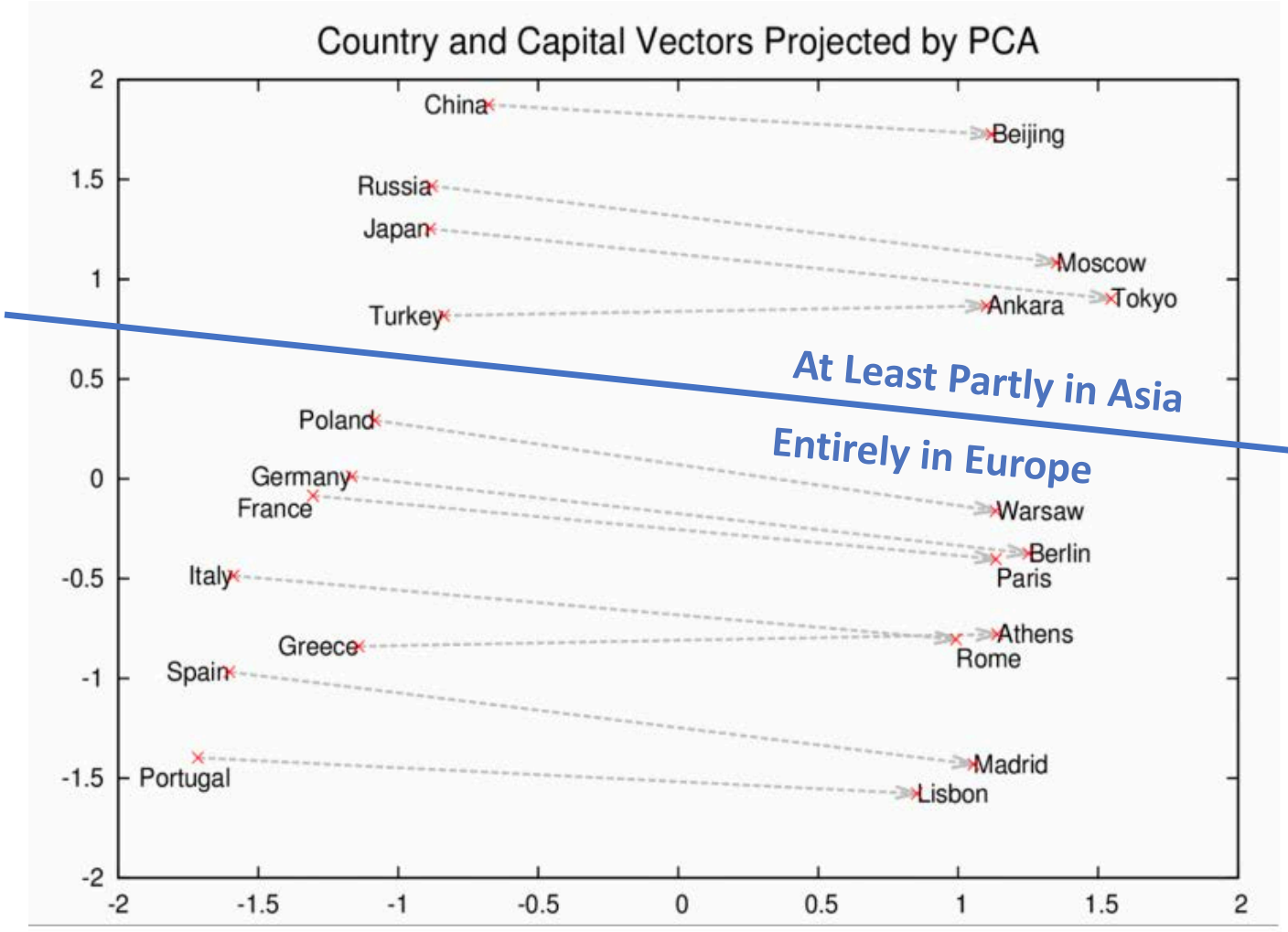
word2vec

by Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean



word2vec

by Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean

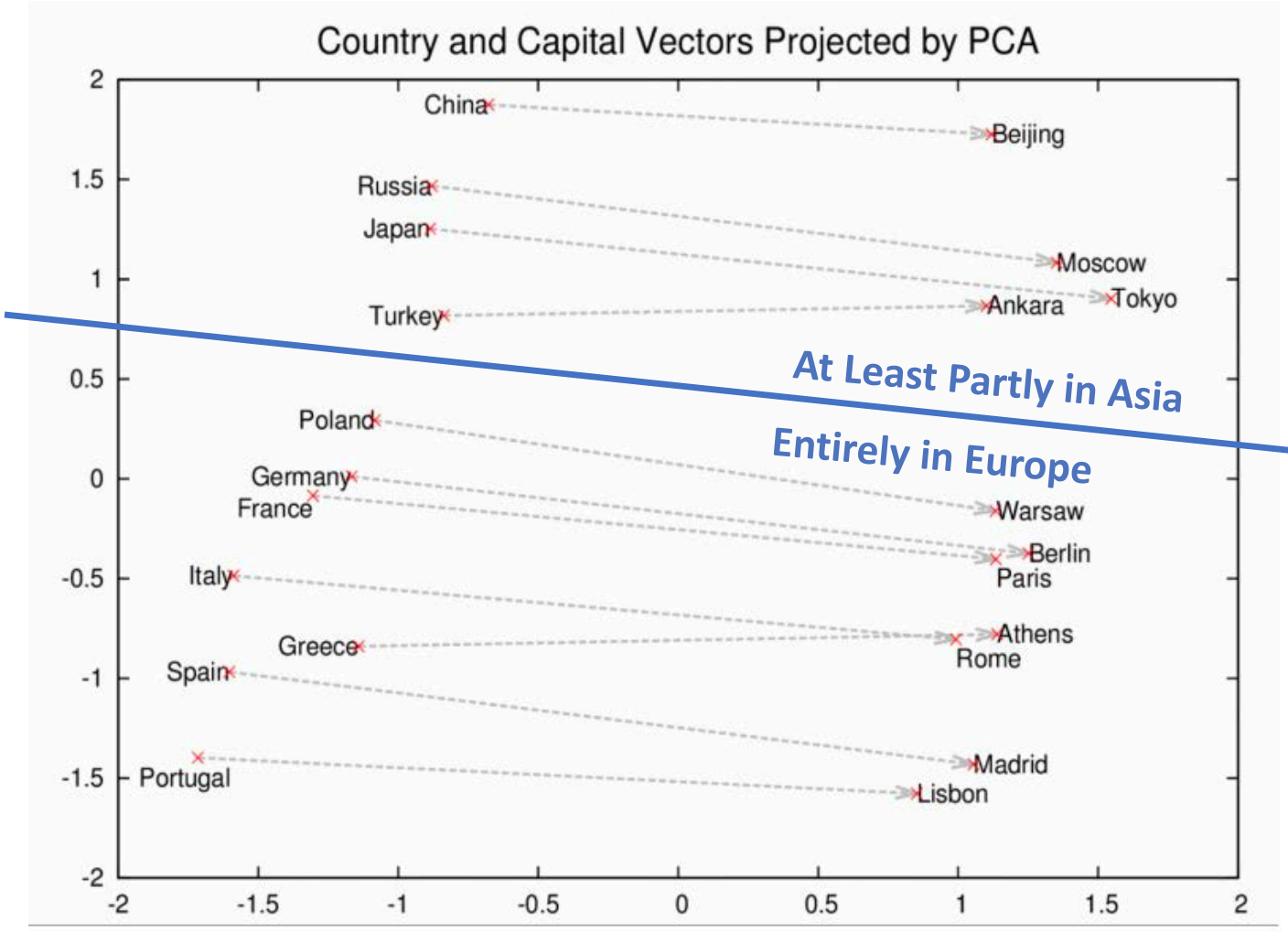


Distance Captures Similarity

Russia is closer to China than to Italy

word2vec

by Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean



Distance Captures Similarity

Russia is closer to China than to Italy

Math Creates Analogies

Russia – Moscow + Paris = France

(Russia:Moscow :: Paris:France)

code2vec

by Uri Alon, Meital Zilberstein, Omer Levy, Eran Yahav

Distance Captures Similarity

`count` is similar to `getCount`

Math Creates Analogies

`equals` + `toLower` = `equalsIgnoreCase`

code2vec

by Uri Alon, Meital Zilberstein, Omer Levy, Eran Yahav

Distance Captures Similarity

`count` is similar to `getCount`

Math Works Out

`equals` + `toLower` = `equalsIgnoreCase`

Applications

Deobfuscation
Adding Code Comments
Code Completion
Code Similarity

ML + Code = ??

ML Task

Classification

or

ML + Code Task

“Code Smell” Detection

safe or suspicious?

ML + Code = ??

ML Task

Classification

or

Automated Translation

That is an ugly cat -> Das ist eine hässliche katze

ML + Code Task

"Code Smell" Detection

safe or suspicious?

Automated Language Porting

System.out.println("Hello!");

-> print("Hello!")

ML + Code = ??

ML Task

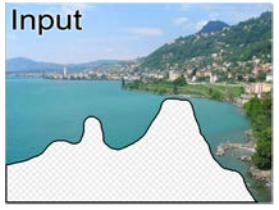
Classification

or

Automated Translation

That is an ugly cat -> Das ist eine hässliche katze

Image Completion



->

ML + Code Task

"Code Smell" Detection

safe or suspicious?

Automated Language Porting

System.out.println("Hello!");

-> print("Hello!")

Smarter Code Completion

```
#ifdef IPG_DEBUG
static void ipg_dump_rfdlist(struct net_device *dev)
{
        struct ipg_nic_private *sp = netdev_priv(dev);
```

Other Tasks

- Focusing attention during code review.
- Automatically generating “glue code.”
- Checking API usage.
- Predicting performance problems.
- Translating English descriptions to code.

The Result

- Developers: Focus on the fun, creative parts
- Tools: Focus on the formulaic parts
- Result: Scalable, quality code with less annoyance

Try It!

- TensorFlow: https://www.tensorflow.org/
- Open Images Dataset: https://storage.googleapis.com/openimages/web/download.html
- Deep Learning Implementations: https://github.com/tdeboissiere/DeepLearningImplementations
- Word2Vec: https://code.google.com/archive/p/word2vec/
- Code2Vec: https://github.com/tech-srl/code2vec