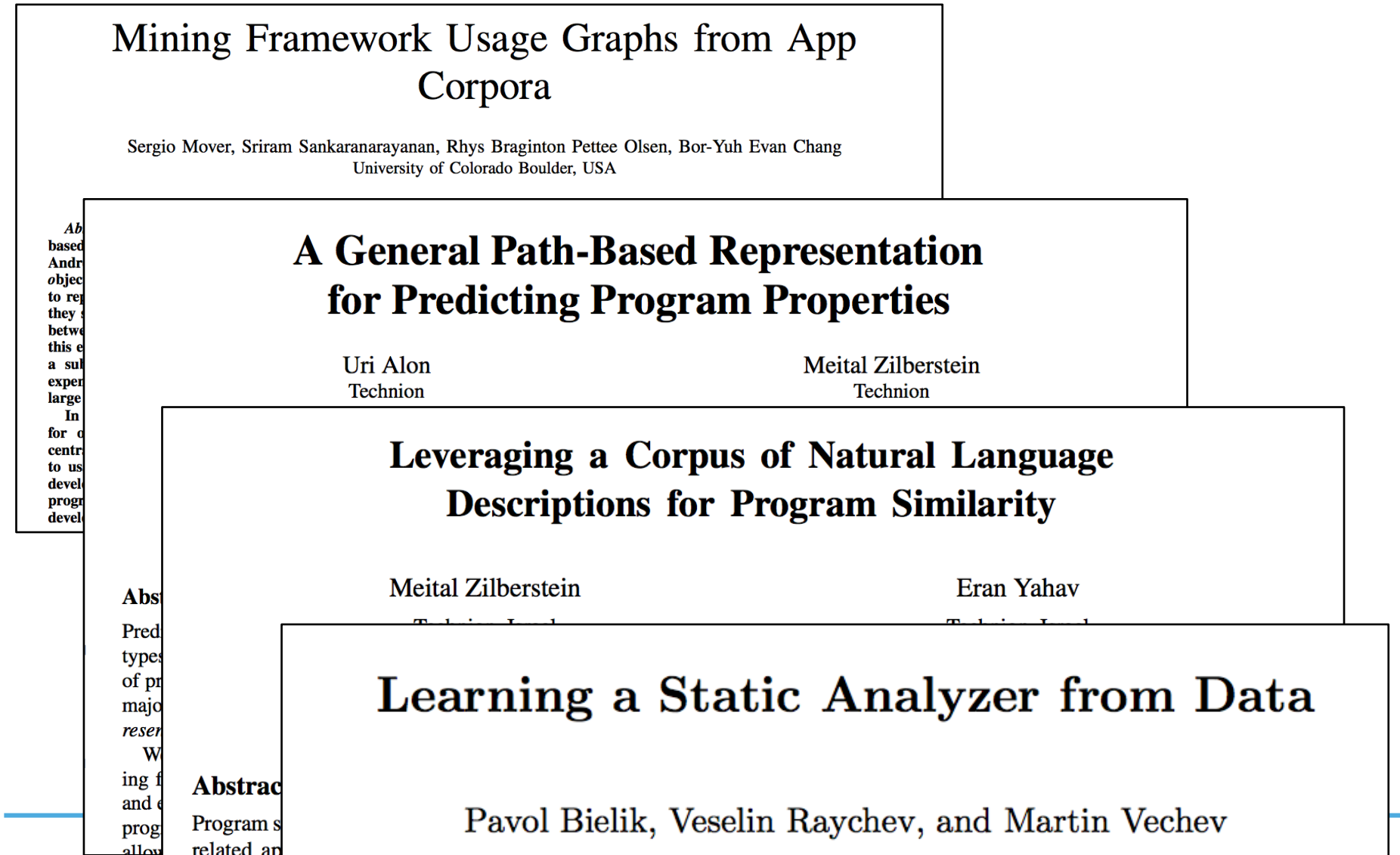
Code + ML:
Will automation take our jobs?

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

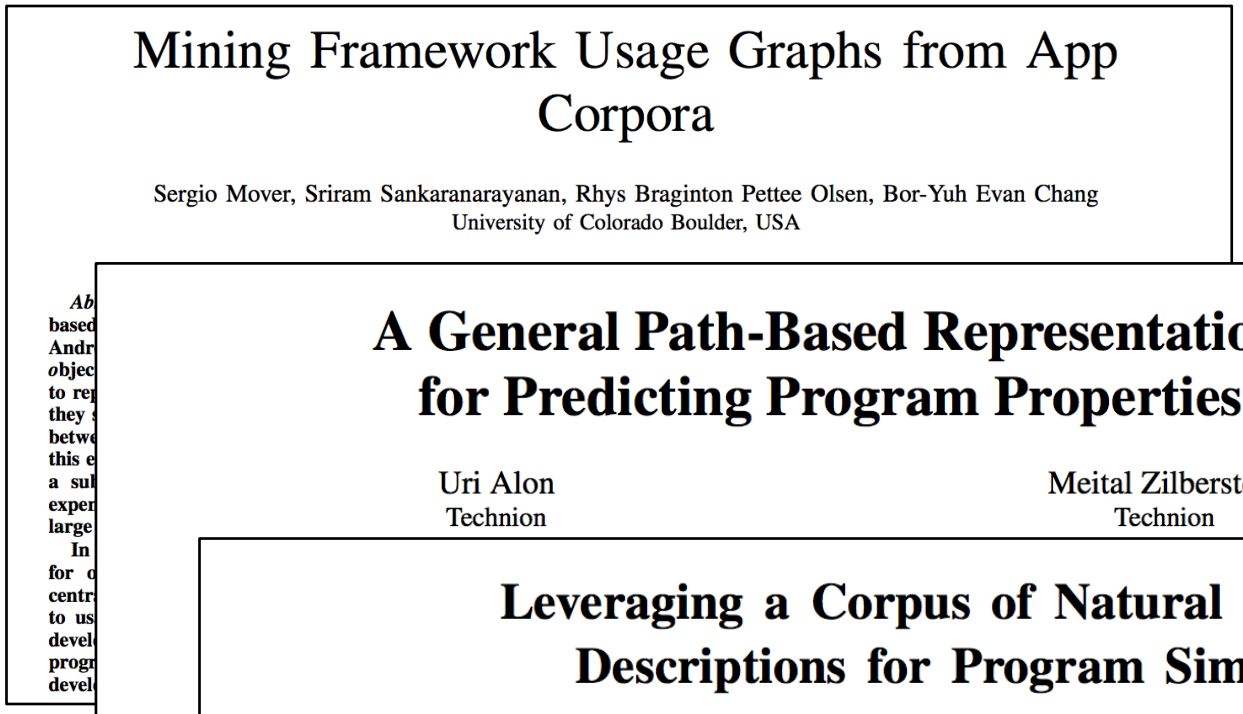
Eran Yahav

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

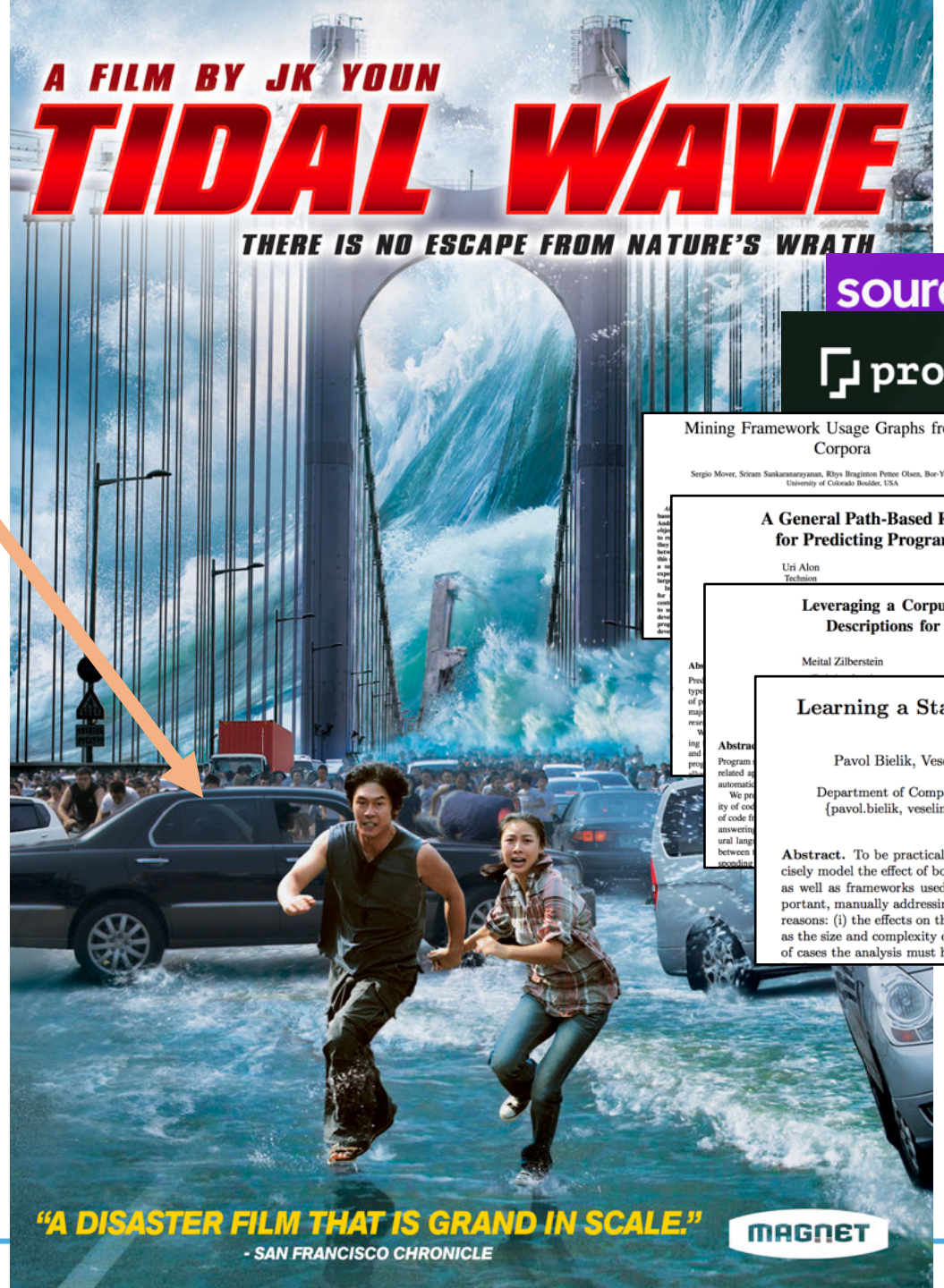
diffblue

DEEPCODE

source{d}

muse dev

Developers?

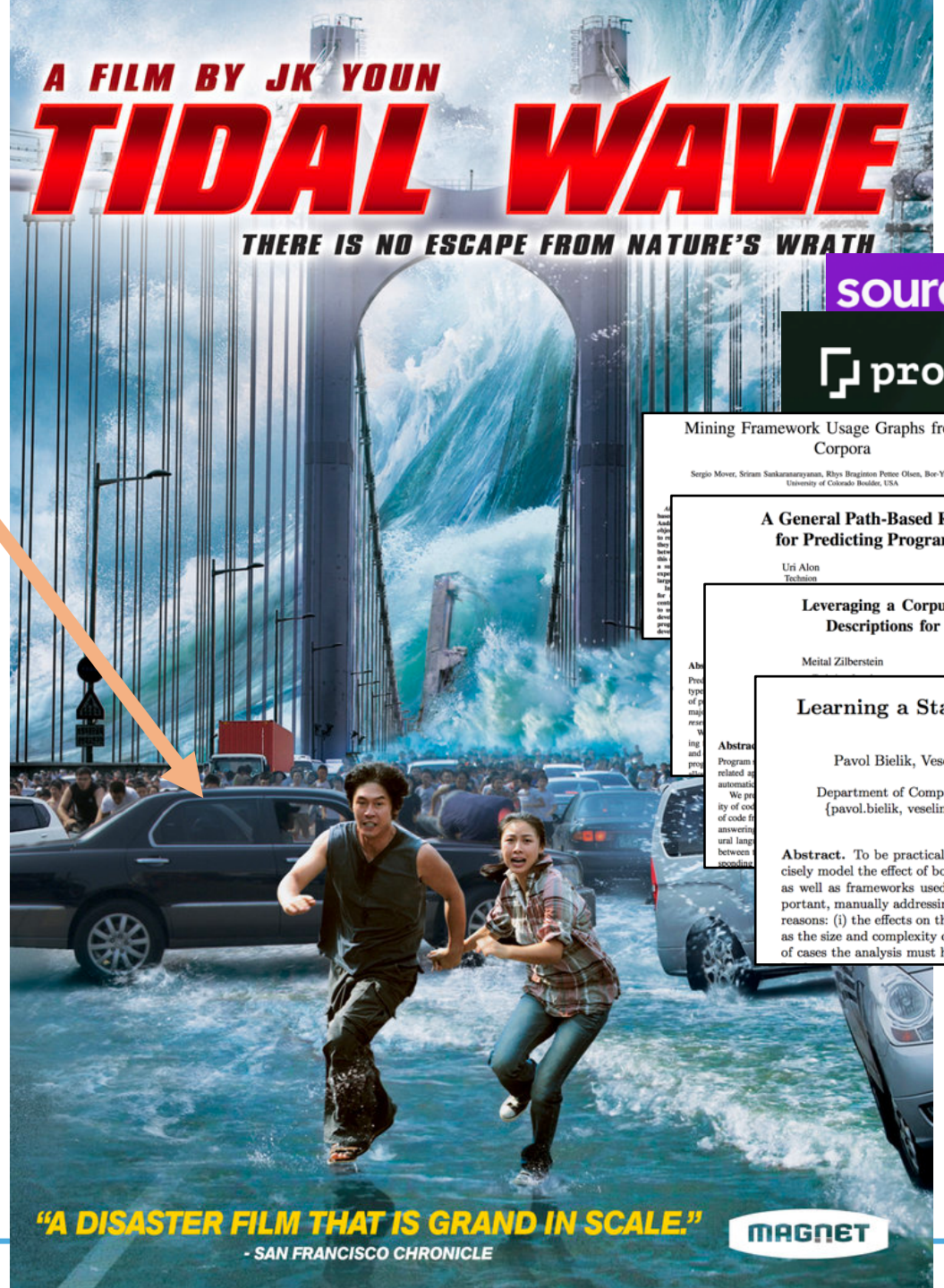


sourced
codota
prodo.ai
PCODE
ffblue

muse dev

Developers?

… or developers?



sourced

codota

prodo.ai

PCODE

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

How Did We Get Here?

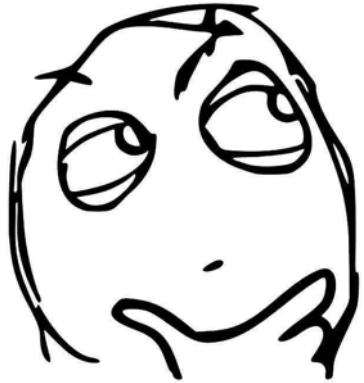
1 hour version: Easy!

well…

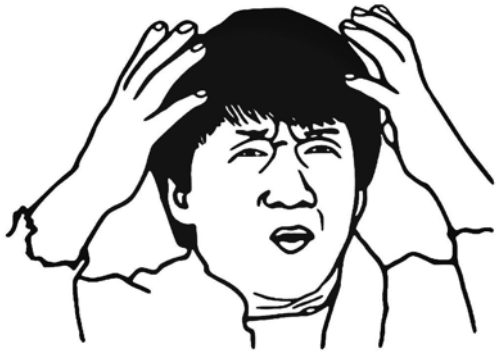
Down The Rabbit Hole

Topics

- What does ML applied to code enable?
- What is ML / AI / NN?
- Deep dive on one cutting-edge technique.
- Quick mention of other techniques.
- Lots of links



balanced with



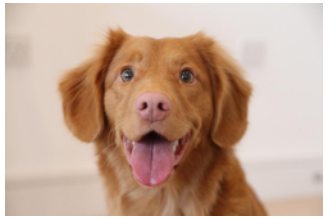
_____ : Images :: _____ : Code

Classification : Images :: _____ : Code

ML Task

Classification

or

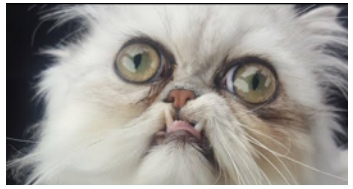


Classification : Images :: _____ : Code

ML Task

Classification

or



Normal Cat

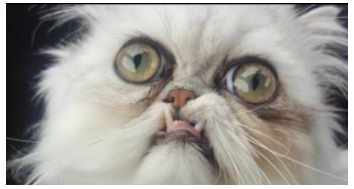
Memeable Cat

Classification : Images :: _____ : Code

ML Task

Classification

or



ML + Code Task

Code Categorization

Binary:

- safe or suspicious?
- high or low quality?
- readable or impenetrable?

Multi-valued:

- “purpose” of function
- Search for similar functions

Translation : English :: _____ : Code

ML Task

Automated Translation

That is a strange cat	->	Das ist eine seltsame katze

Translation : English :: _____ : Code

ML Task

Automated Translation

That is a strange cat -> Das ist eine seltsame katze

ML + Code Task

Automated Language Porting

```
System.out.println("Hello!");
```
->
```
print("Hello!")
```

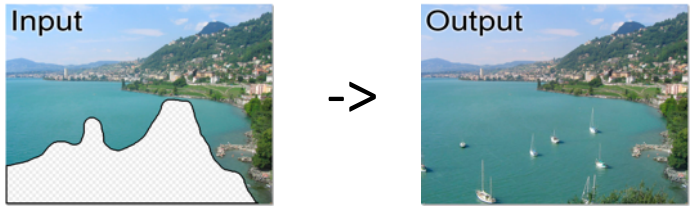
API Translation

```
BufferedReader br = new BufferedReader(new FileReader(file));
st = br.readLine();
```
->
```
Scanner sc = new Scanner(new File(file));
st = sc.nextLine();
```

Completion : Images :: _____ : Code

ML Task

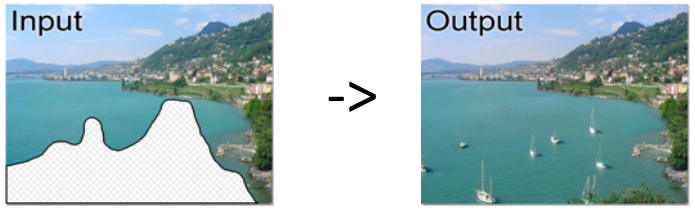
Image Completion



Completion : Images :: _____ : Code

ML Task

Image Completion



->

ML + Code Task

Smarter Code Completion

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

Das, Subhasis. "Contextual Code Completion Using Machine Learning." (2015).

```
import java.io.*;
import java.util.*;
public class TestIO {
   void read(File file) {
      /// call:readLine type:FileReader type:BufferedReader
   }
}
```

->

```
FileReader fr1;
BufferedReader br1;
String s1;
fr1 = new FileReader(file);
br1 = new BufferedReader(fr1);
while ((s1 = br1.readLine()) != null) {}
br1.close();
```

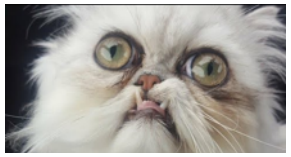
Murali, Vijayaraghavan, et al. "Neural sketch learning for conditional program generation." *arXiv preprint arXiv:1703.05698* (2017).

ML + Code = ??

ML Task

Classification

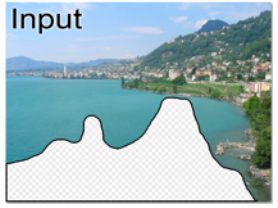
or



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!")`

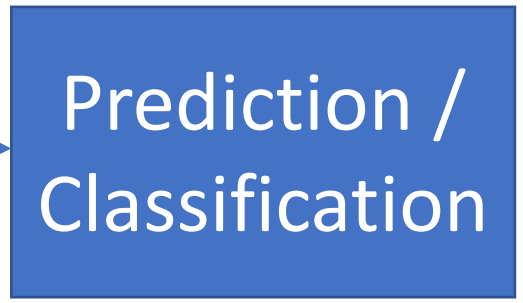
Smarter Code Completion

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

What is Machine Learning?

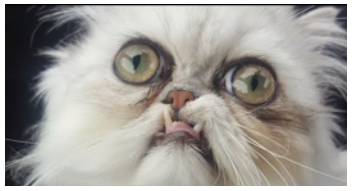
Deep Learning $\subset$ ANNs $\subset$ ML $\subset$ AI

Artificial Intelligence



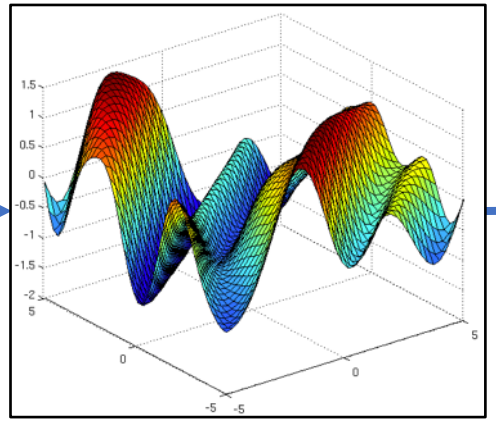
Normal

Memeable



Machine Learning

Data

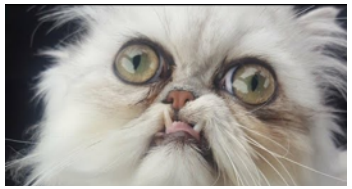


http://katbailey.github.io/post/gaussian-processes-for-dummies/

Prediction / Classification

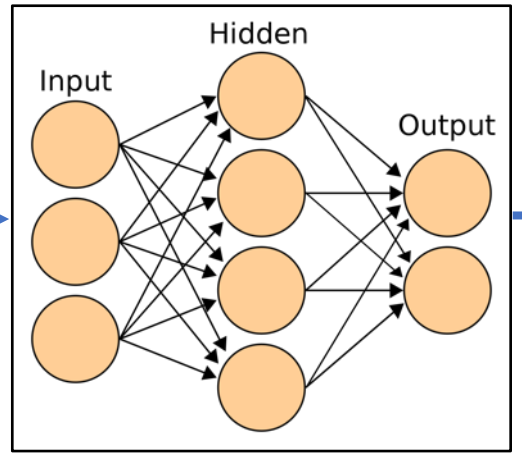
Normal

Memeable



Artificial Neural Networks

Data

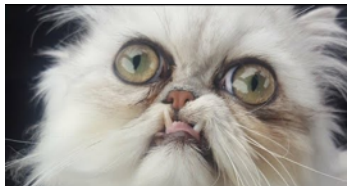


Prediction / Classification

https://www.analyticsvidhya.com/blog/2014/10/ann-work-simplified/

Normal

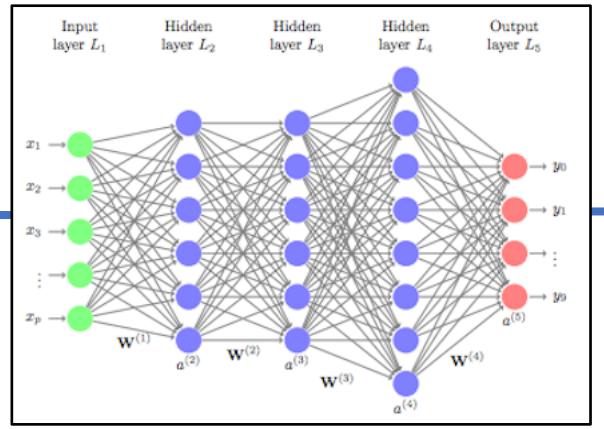
Memeable



Deep Learning

Data

Lots (Millions) of Images

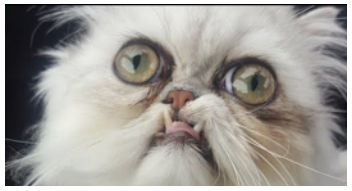


http://uc-r.github.io/feedforward_DNN

Prediction / Classification

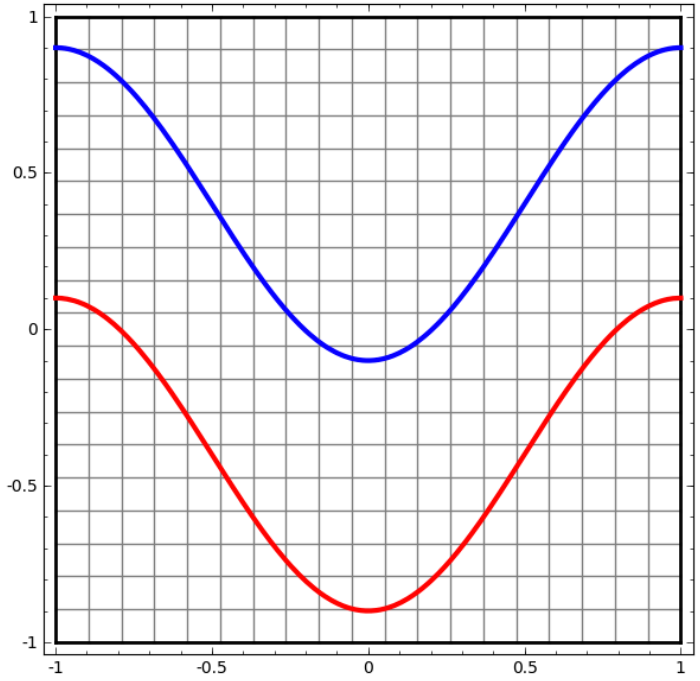
Normal

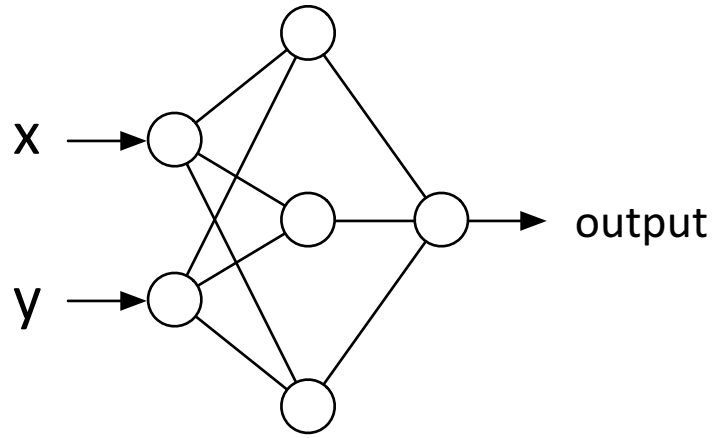
Memeable



How Do Neural Networks Work?

Images From: http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

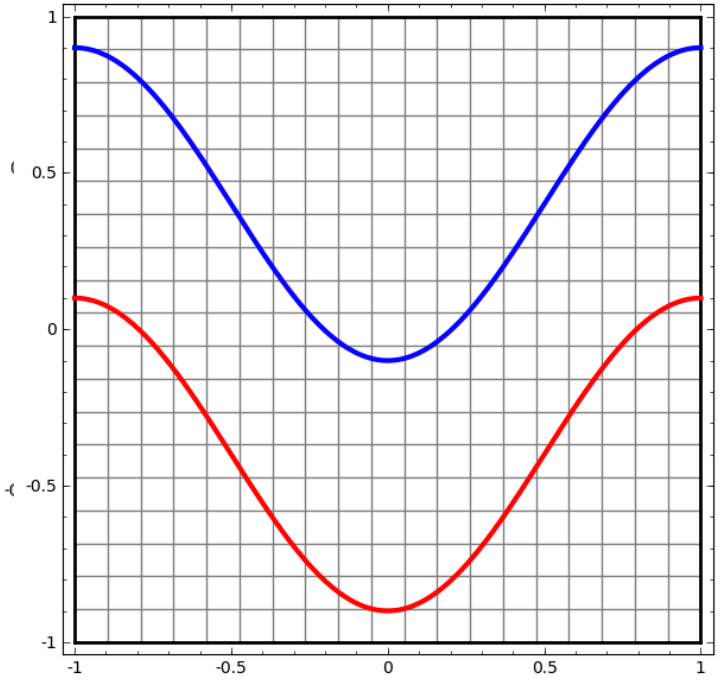


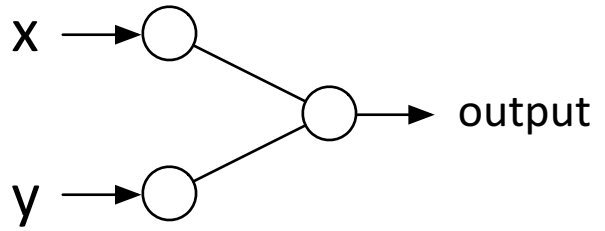


Red if output < 0, blue otherwise

How Do Neural Networks Work?

Images From: http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/



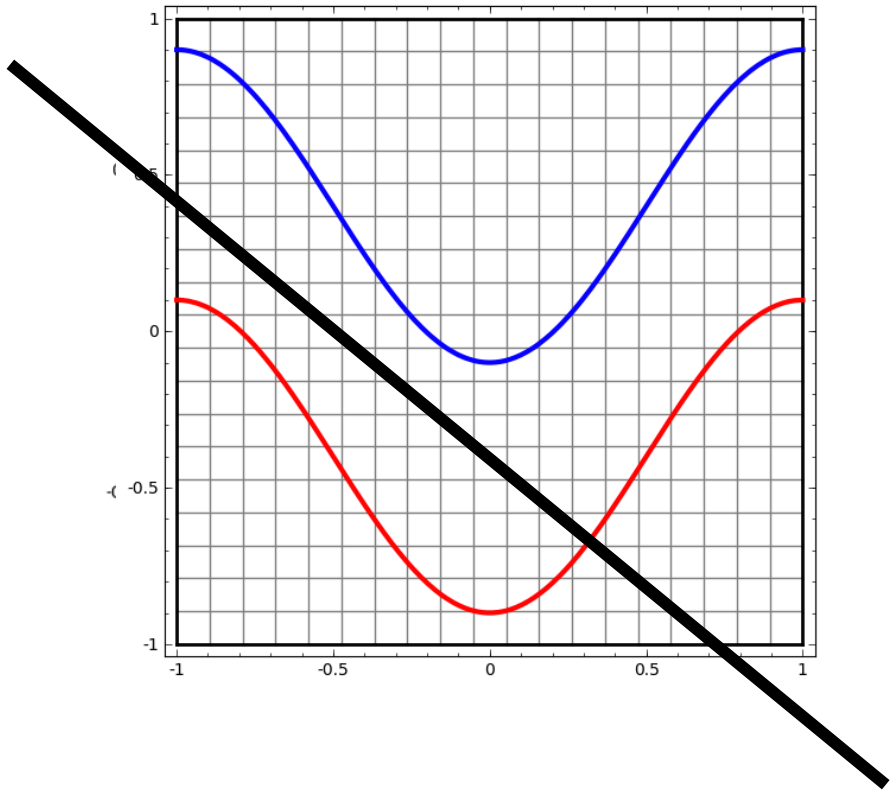


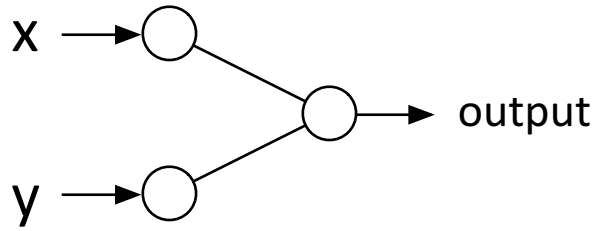
$$\text{output} = w_0 \cdot x + w_1 \cdot y$$

Red if output < 0, blue otherwise

How Do Neural Networks Work?

Images From: http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/



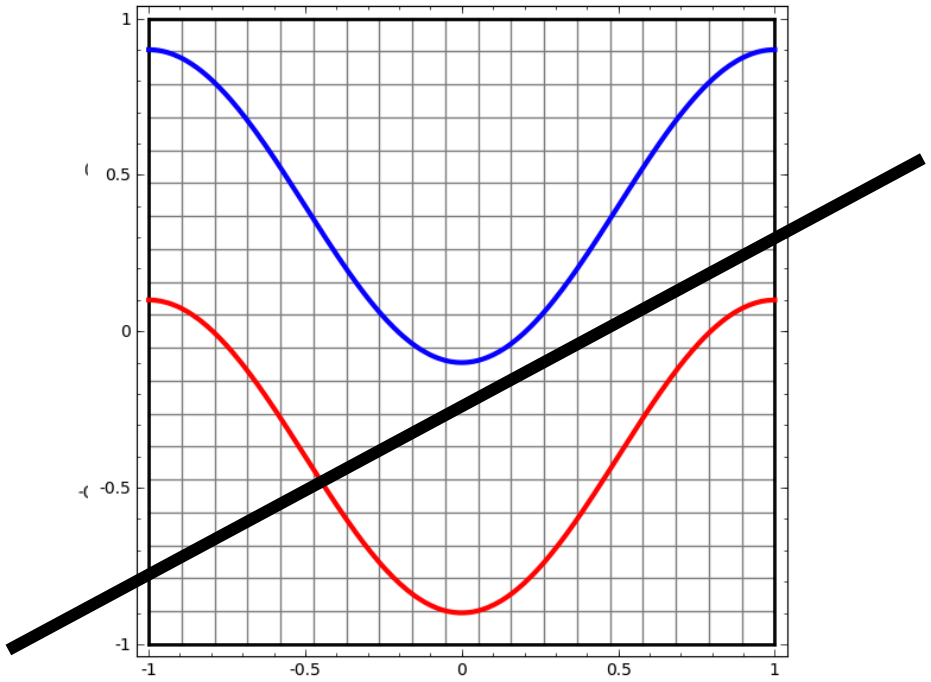


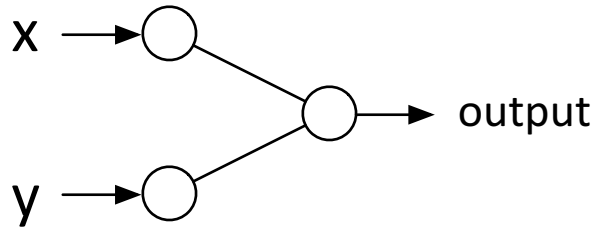
$$\text{output} = w_0 \cdot x + w_1 \cdot y$$

Red if output < 0, blue otherwise

How Do Neural Networks Work?

Images From: http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/



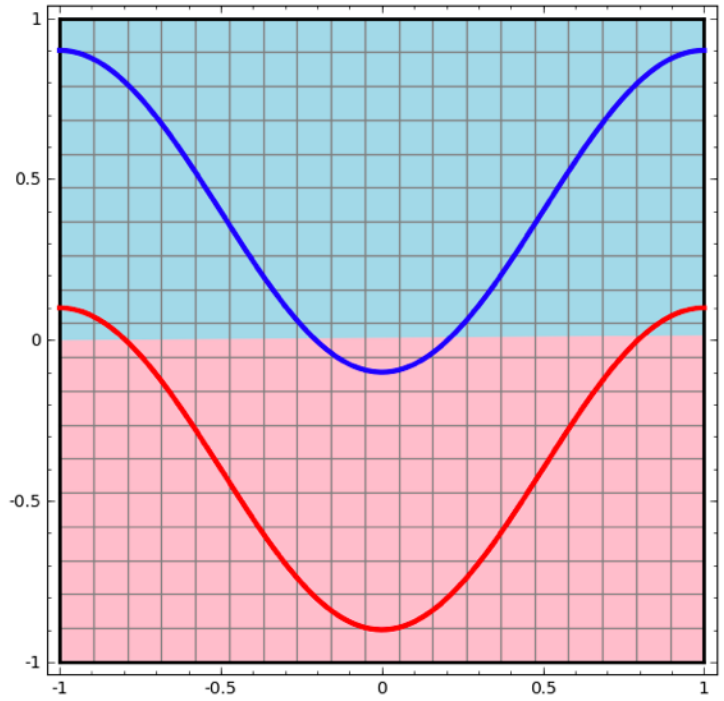


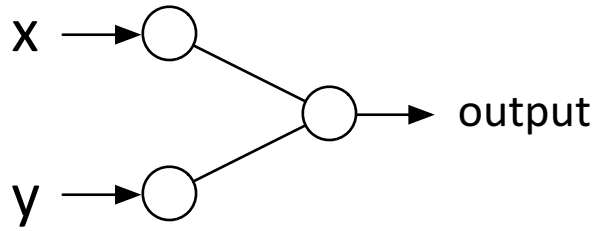
$$\text{output} = w_0 \cdot x + w_1 \cdot y$$

Red if output < 0, blue otherwise

How Do Neural Networks Work?

Images From: http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/



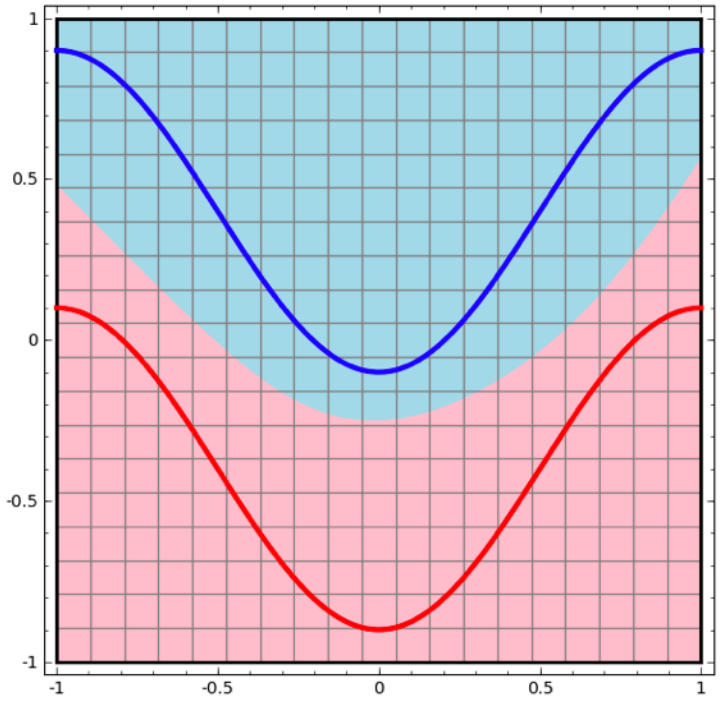


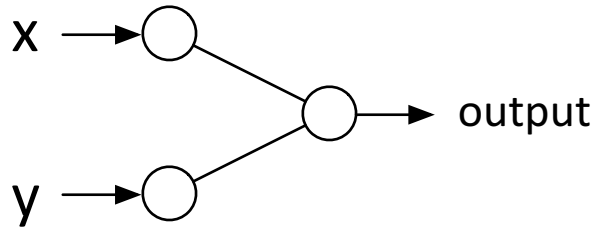
$$\text{output} = w_0 \cdot x + w_1 \cdot y$$

Red if output < 0, blue otherwise

How Do Neural Networks Work?

Images From: http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/



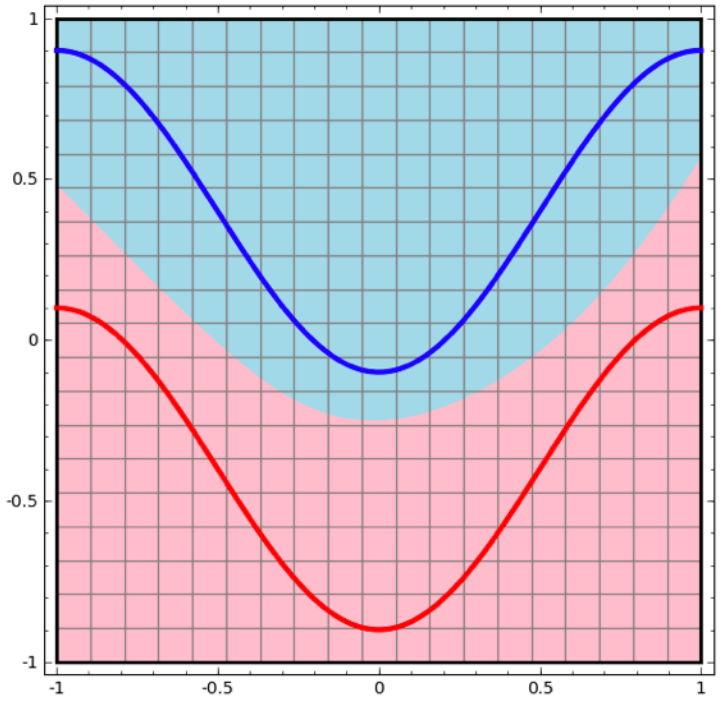


$$\text{output} = f(x, y)$$

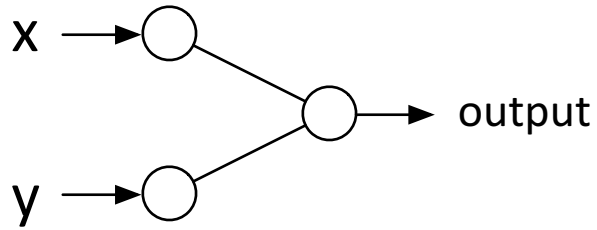
Red if output < 0, blue otherwise

How Do Neural Networks Work?

Images From: http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/



What is f?

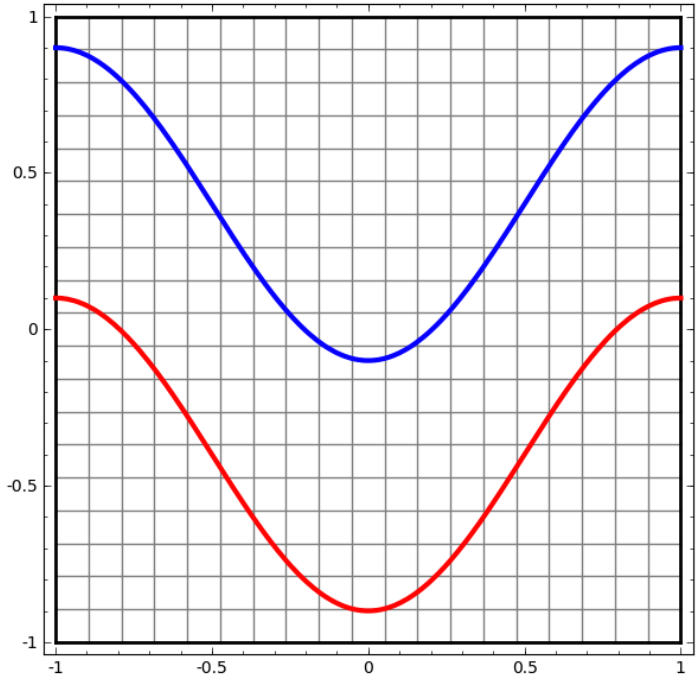


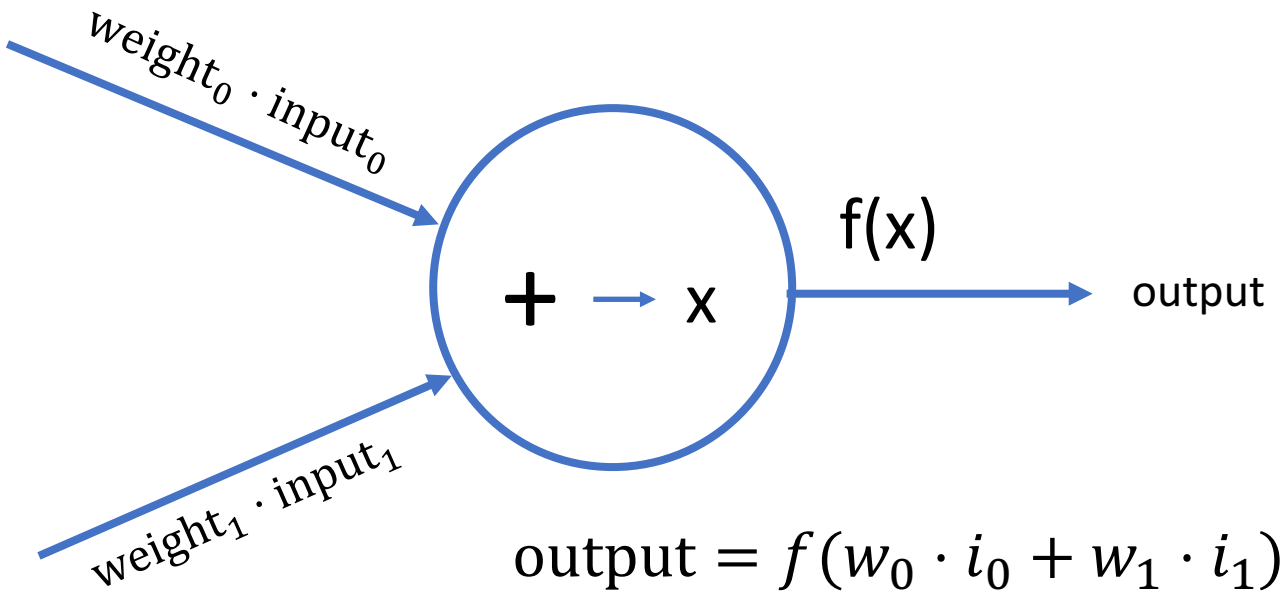
$$\text{output} = f(x, y)$$

Red if output < 0, blue otherwise

How Do Neural Networks Work?

Images From: http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

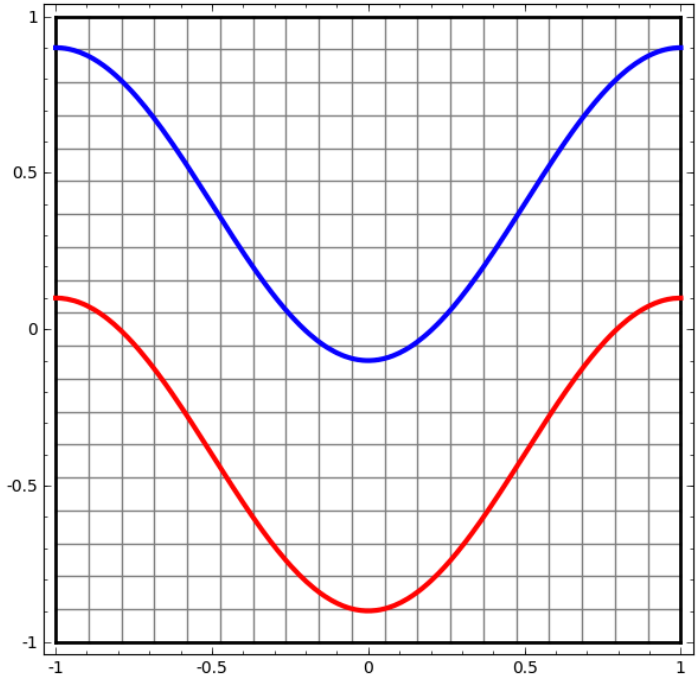


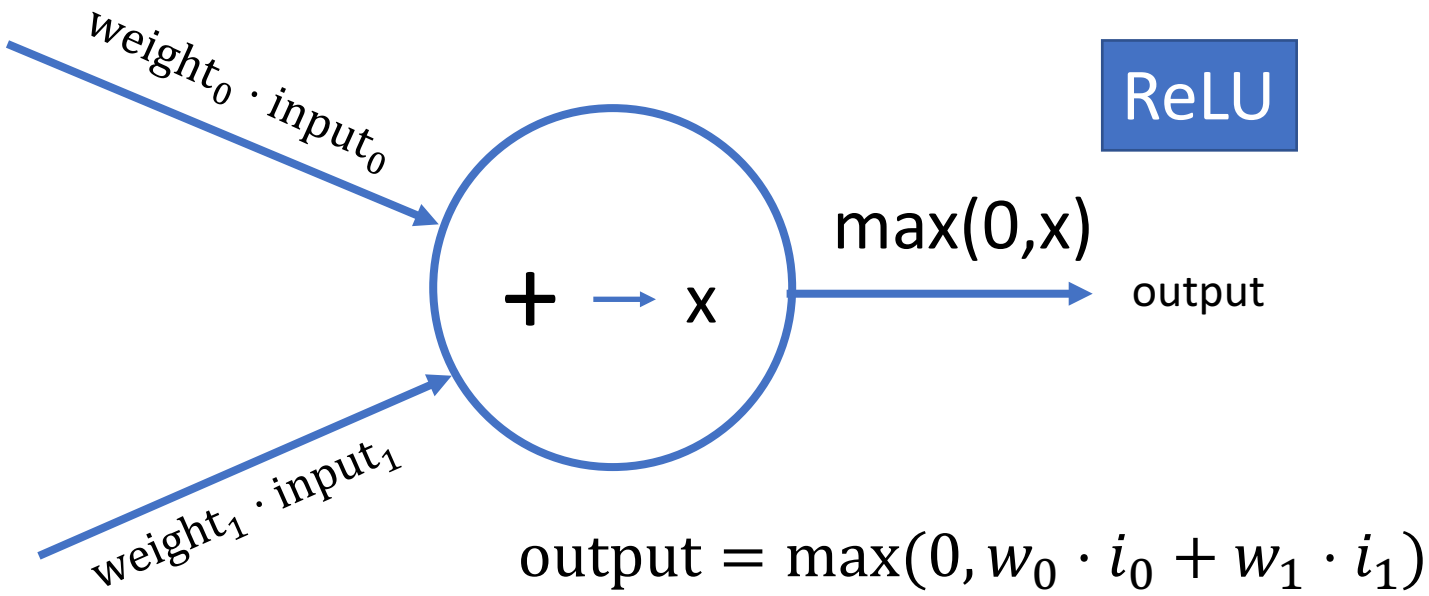


$$\text{output} = f(w_0 \cdot i_0 + w_1 \cdot i_1)$$

How Do Neural Networks Work?

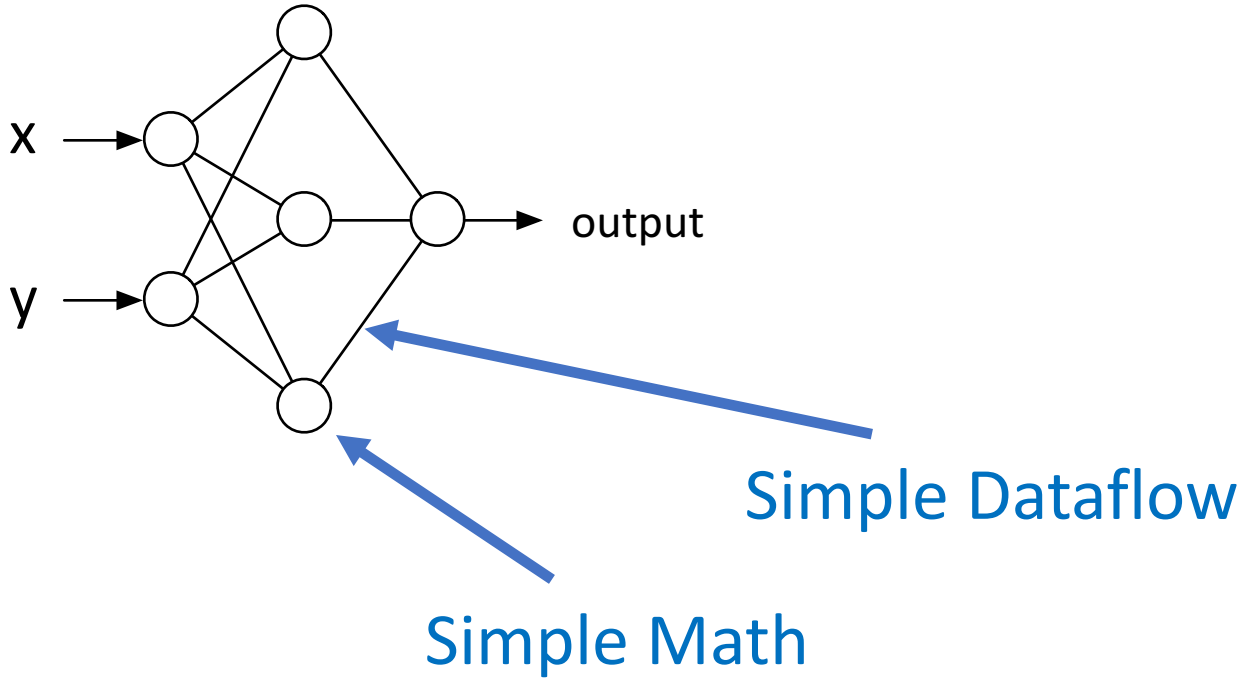
Images From: http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/





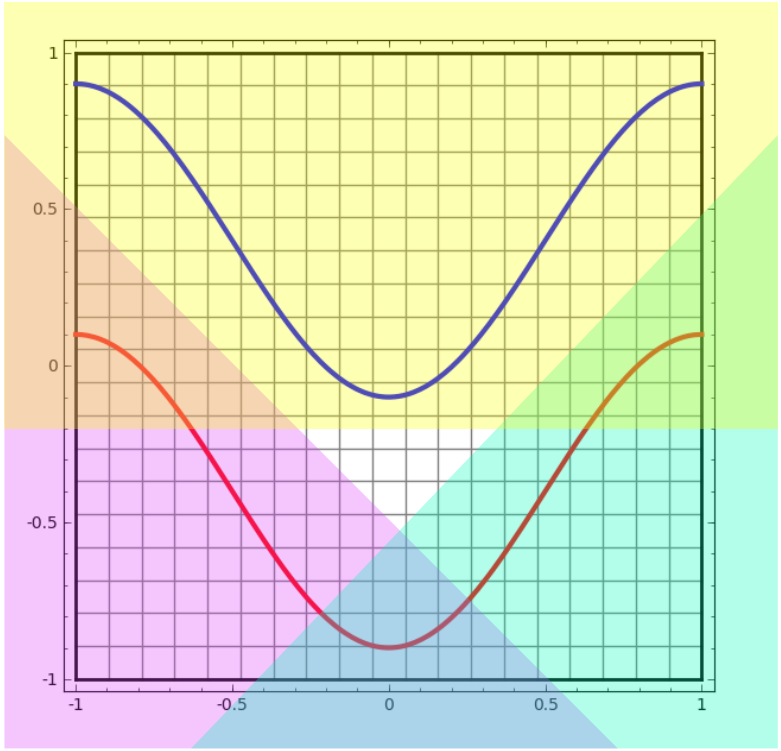
$$\text{output} = \max(0, w_0 \cdot i_0 + w_1 \cdot i_1)$$

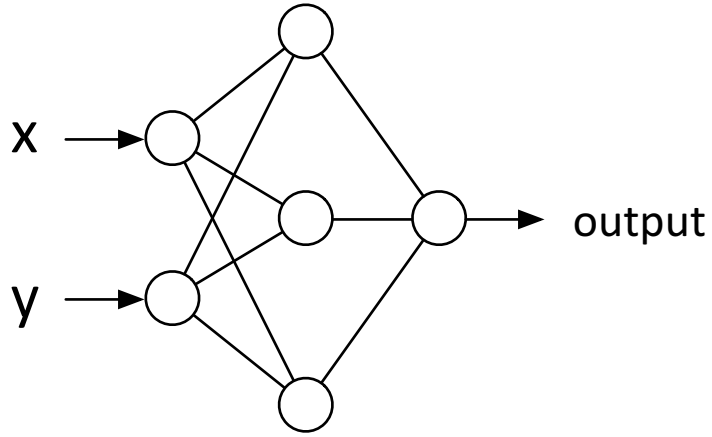
Computationally – Dead Simple



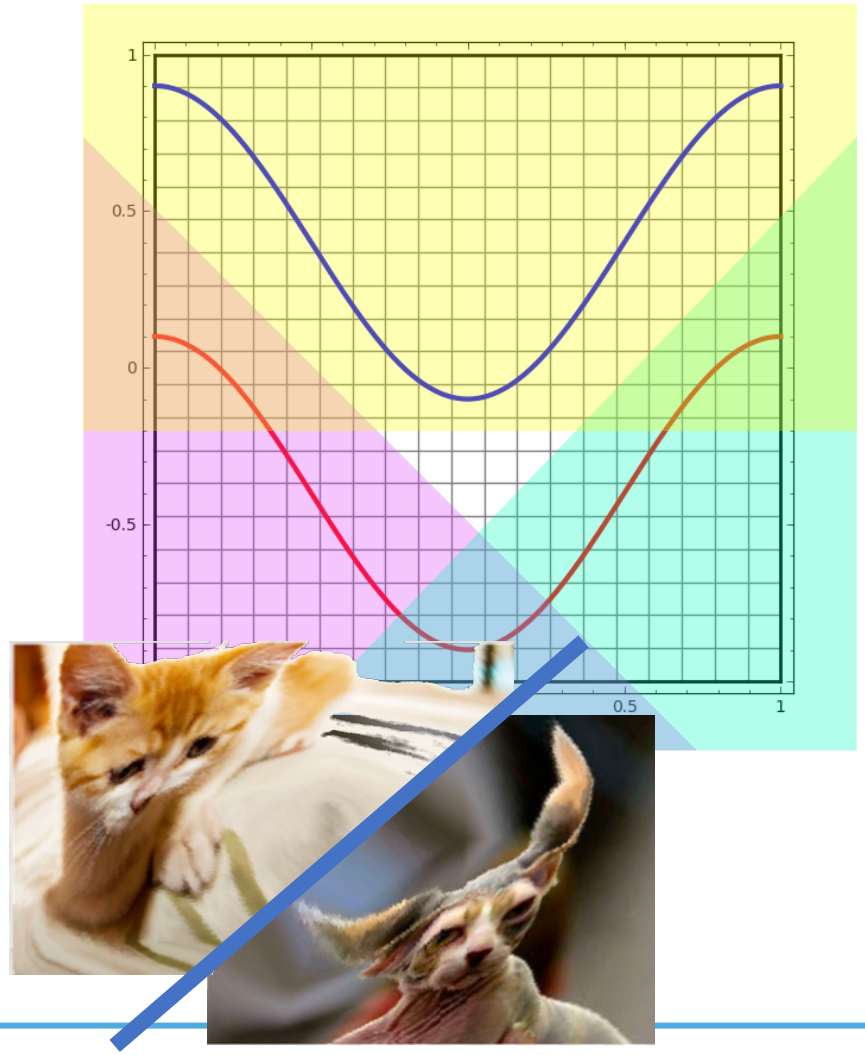
Demo Time!

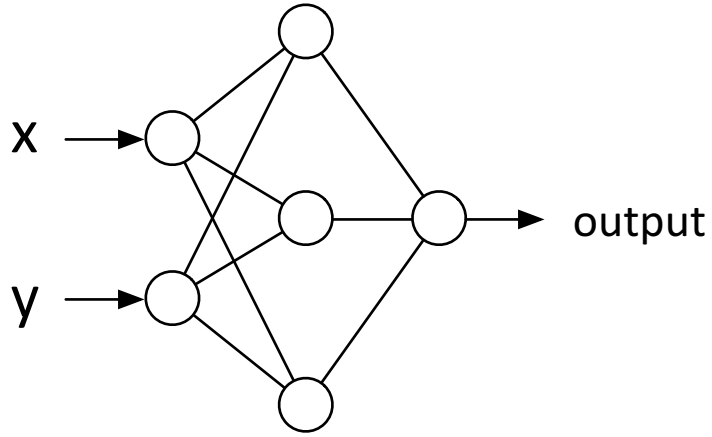
Learning Boundaries



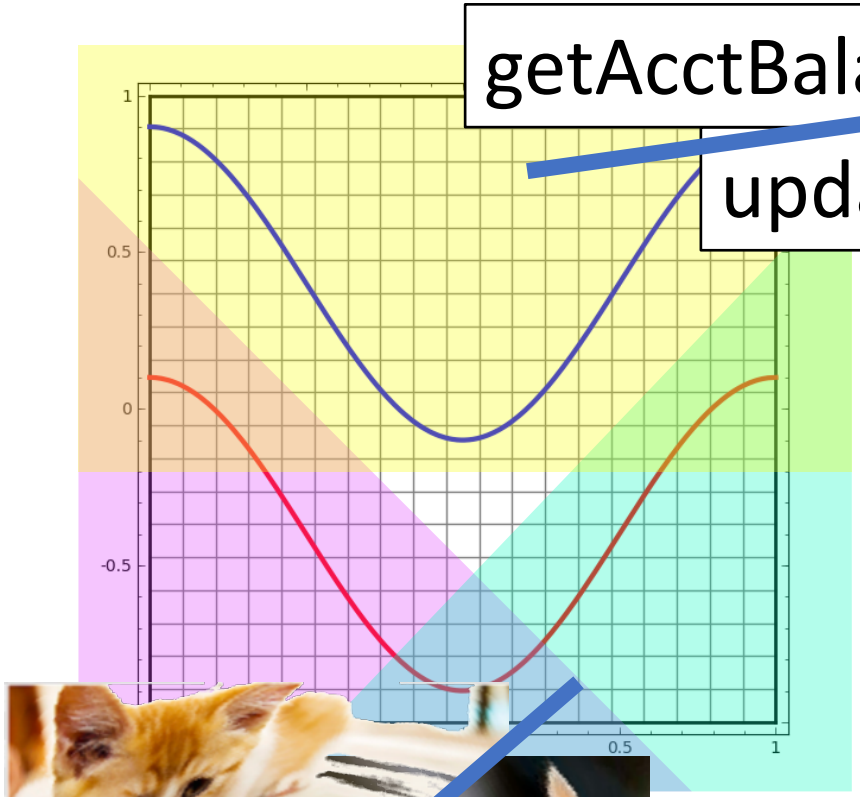


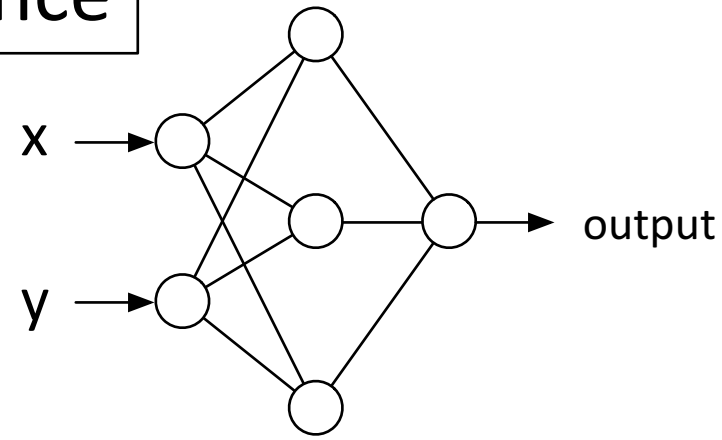
Learning Boundaries





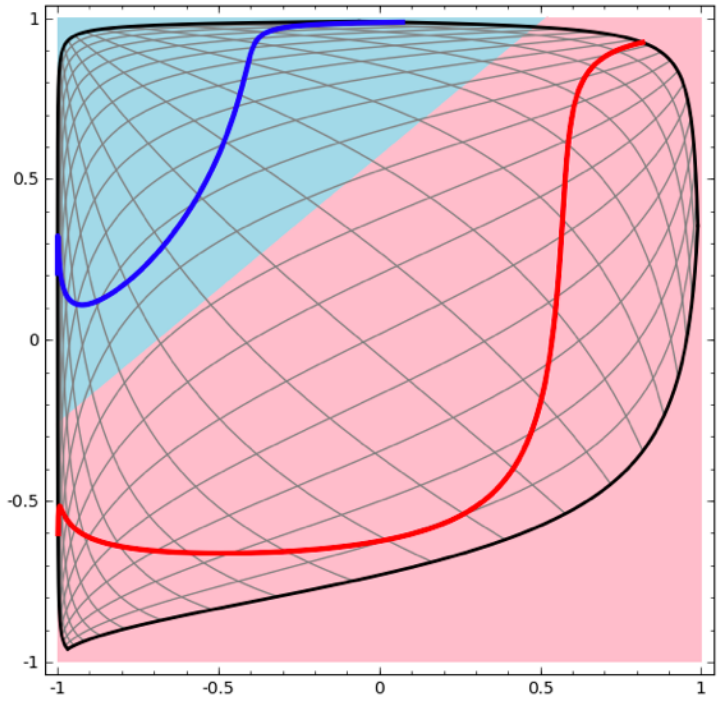
Learning Boundaries

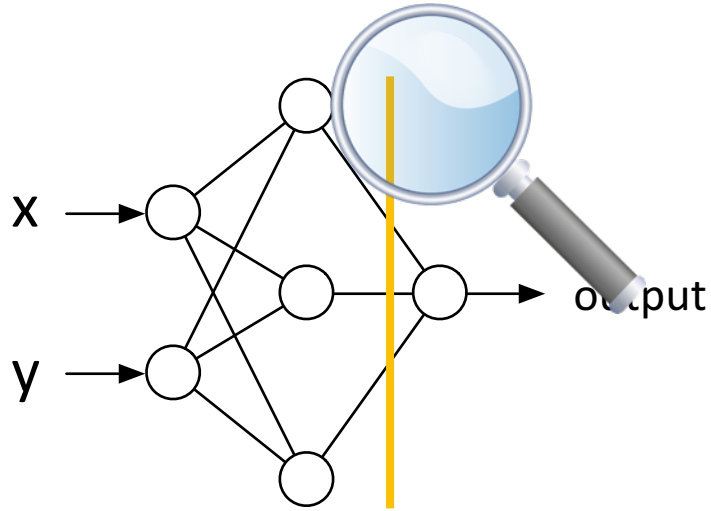




How Do Neural Networks Work?

Images From: http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/



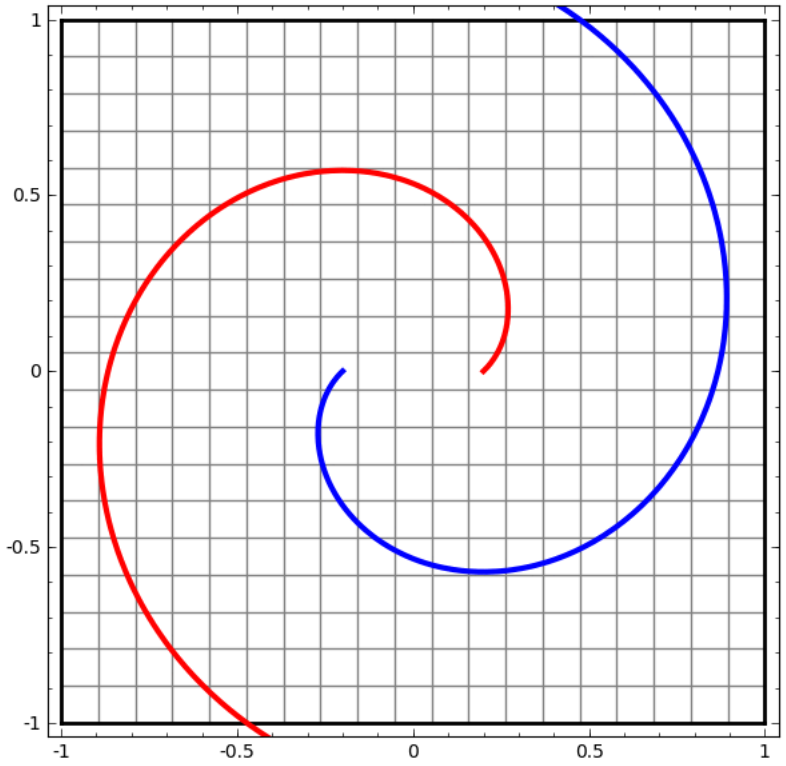


$$\text{output} = w_0 \cdot i_0 + w_1 \cdot i_1 + w_2 \cdot i_2$$

Red if output < 0, blue otherwise

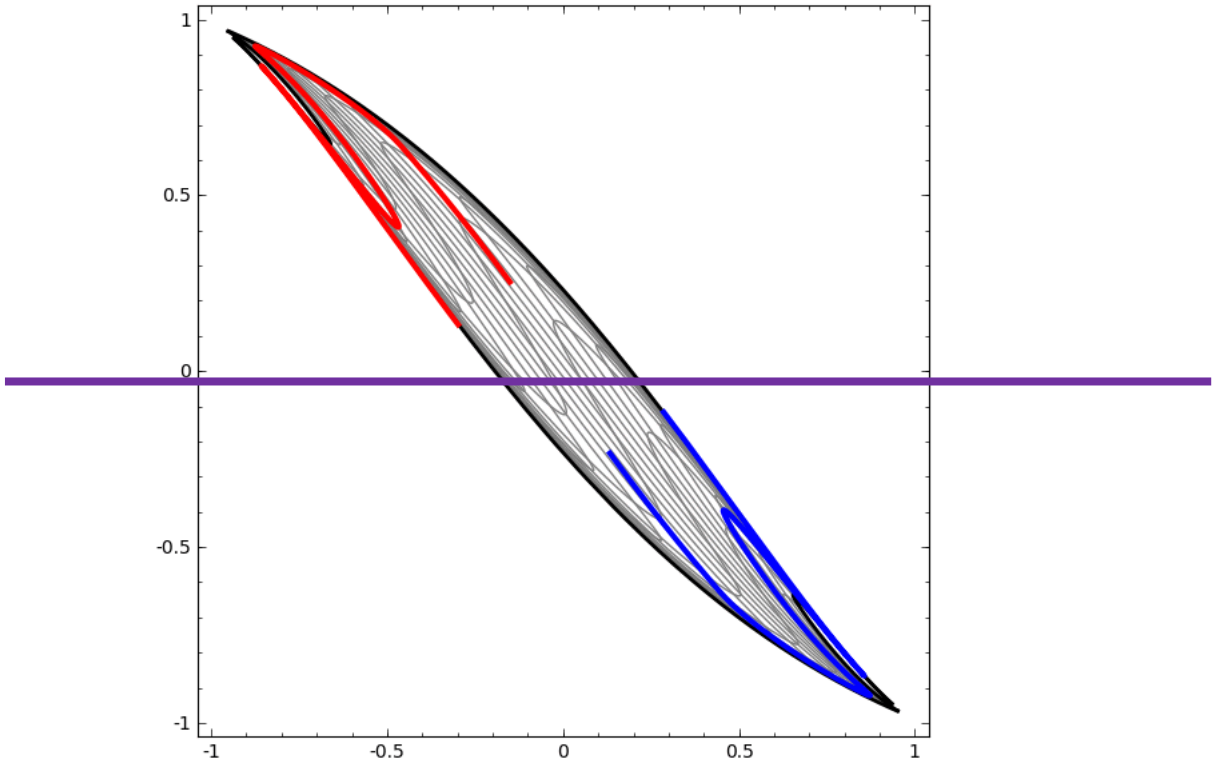
How Do Neural Networks Work?

Images From: http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/



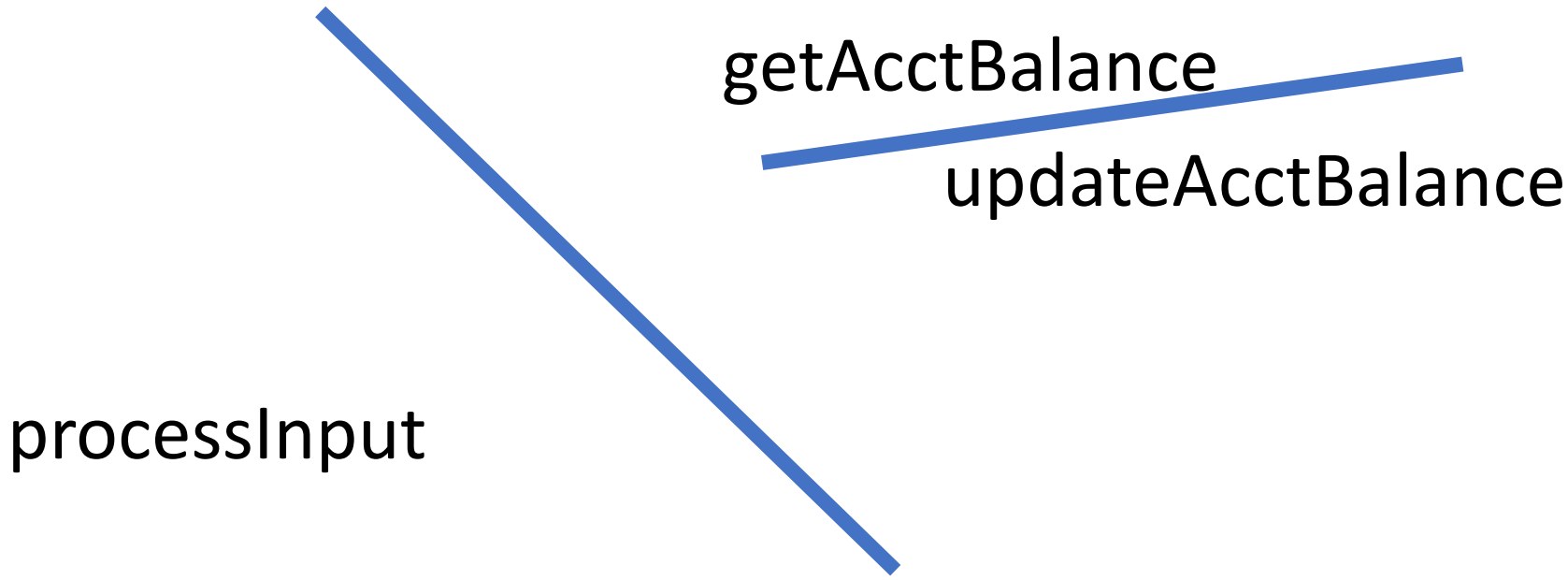
How Do Neural Networks Work?

Images From: http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/



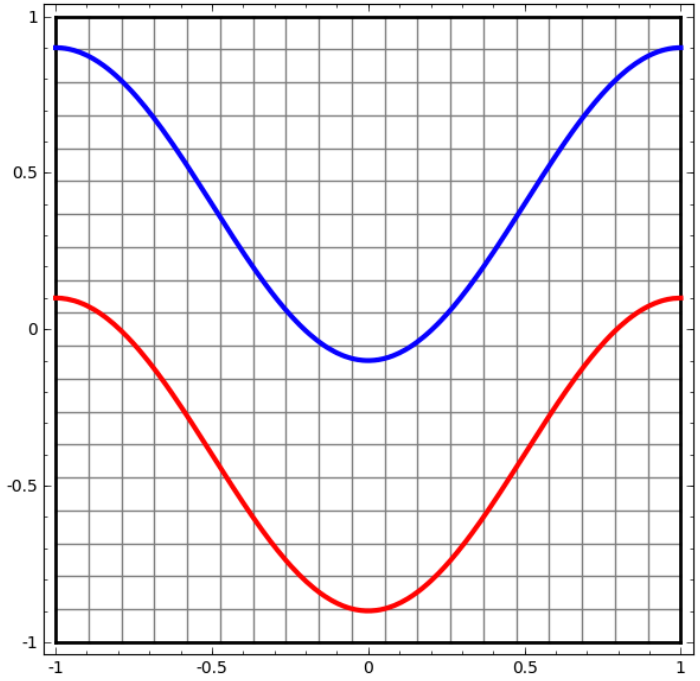
How Do Neural Networks Work?

Images From: http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

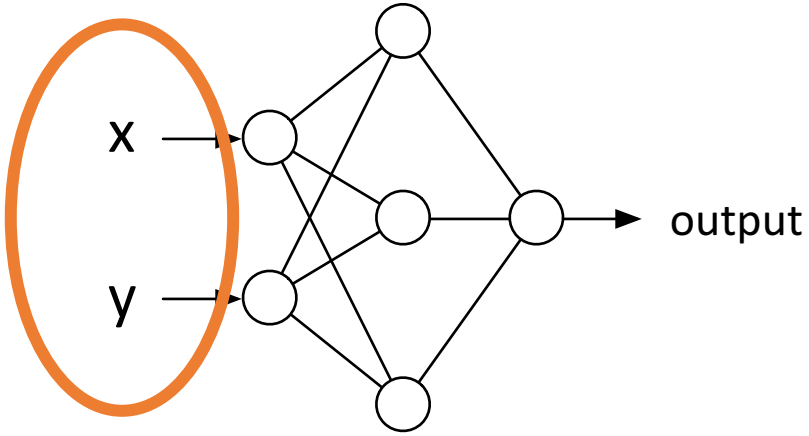


Neural Network Input

Images From: http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

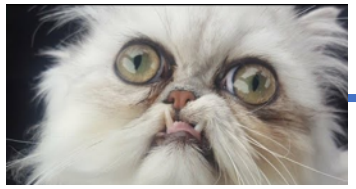


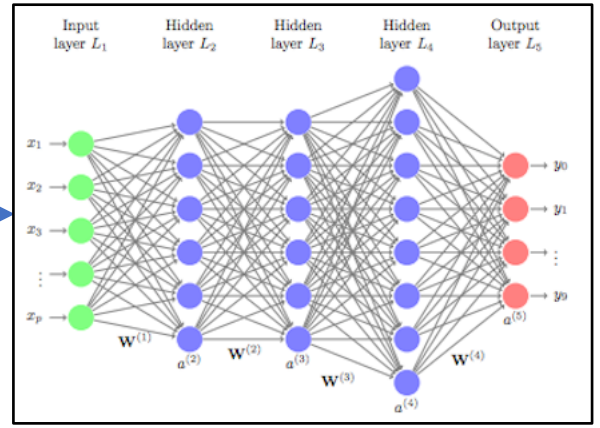
What goes here for images?



Red if output > 0, blue otherwise

Neural Network Input



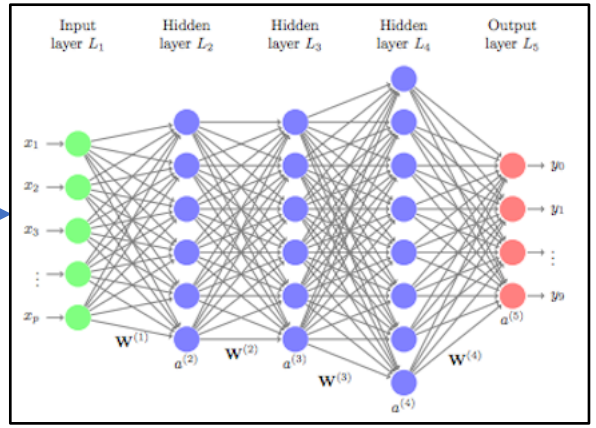


How to connect data to NN inputs?

Neural Network Input for **Code**

```
while (iter.hasNext()) {
  elem = iter.next();
  if(elem = v)
```

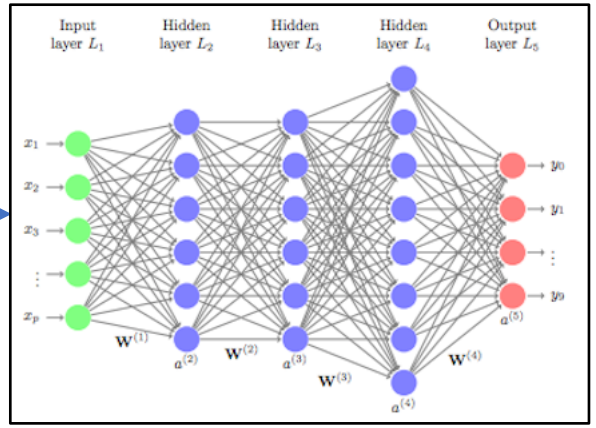
Character stream /
Token stream



Neural Network Input for **Code**

```
while (iter.hasNext()) {
  elem = iter.next();
  if(elem = v)
```

Abstract Syntax Tree



Neural Network Input for **Code**

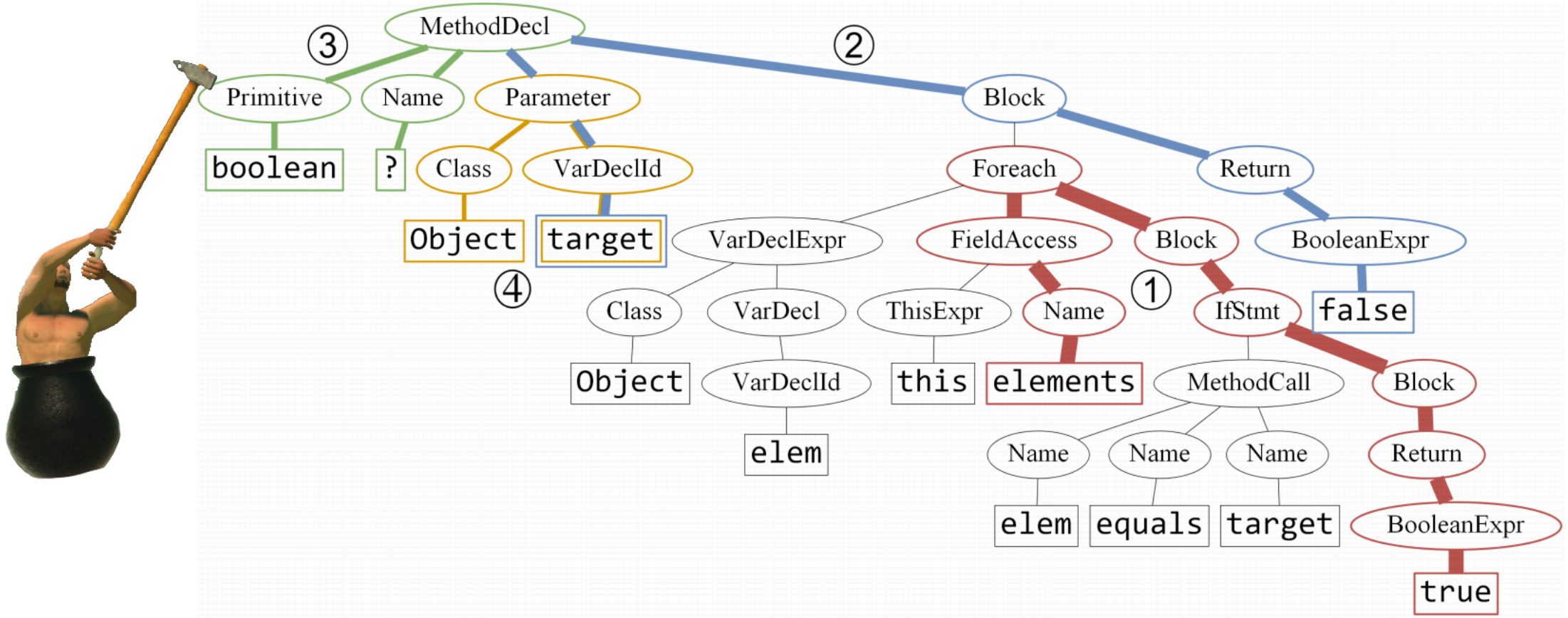
code2vec: Learning Distributed Representations of Code

```
boolean contains(Object target) {
    for (Object elem: this.elements) {
        if (elem.equals(target)) {
            return true;
        }
    }
    return false;
}
```

Alon, Uri, et al. "code2vec: Learning distributed representations of code." *Proceedings of the ACM on Programming Languages* 3.POPL (2019): 40.

Neural Network Input for **Code**

code2vec: Learning Distributed Representations of Code

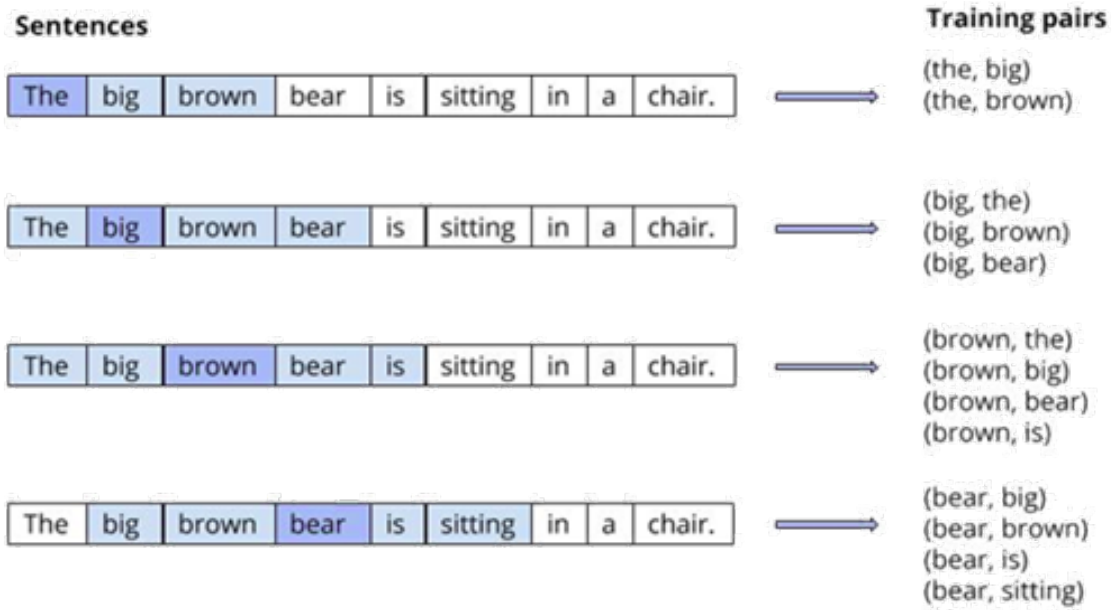


(elements, Name↑FieldAccess↑Foreach↓Block↓IfStmt↓Block↓Return↓BooleanExpr, true)

muse dev

Neural Network Output

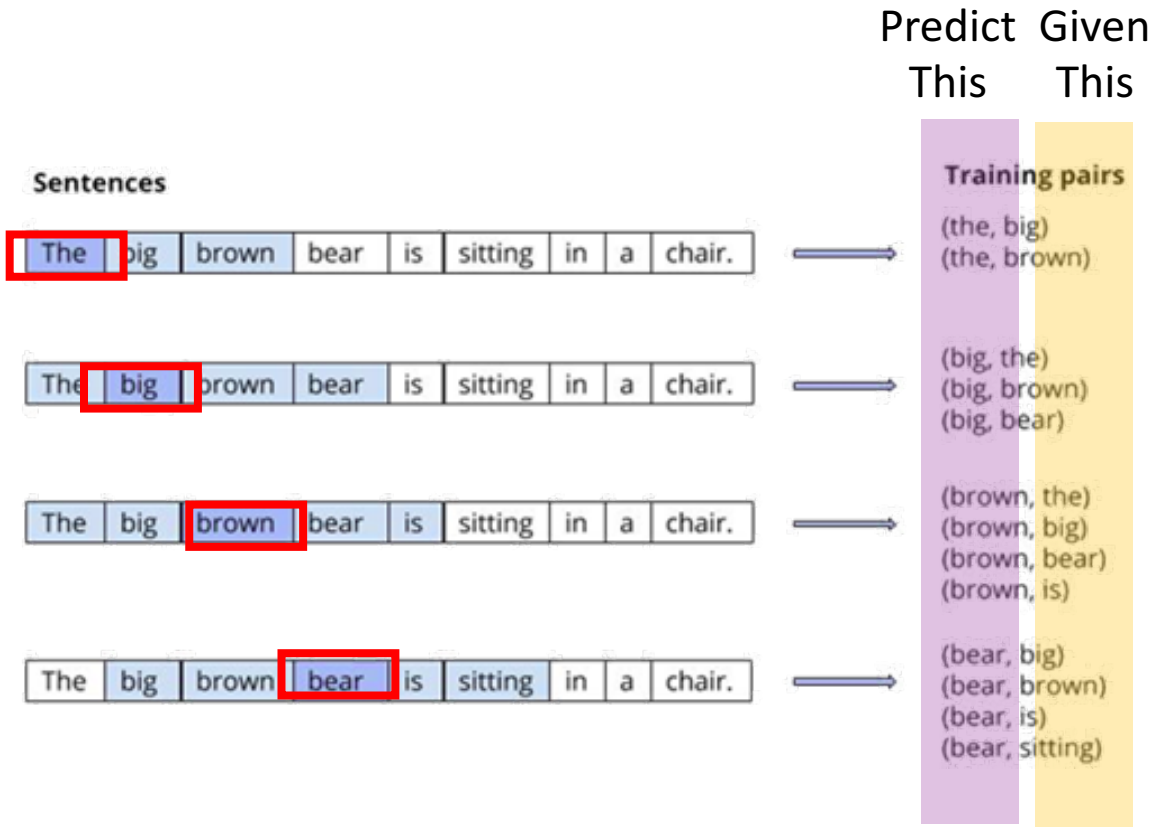
word2vec by Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean



https://www.smartcat.io/blog/2017/word2vec-the-world-of-word-vectors/

Neural Network Output

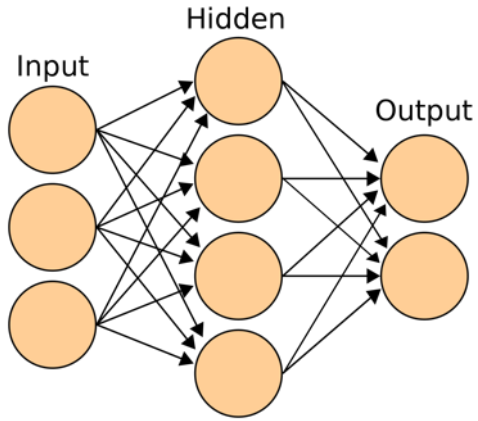
word2vec by Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean



https://www.smartcat.io/blog/2017/word2vec-the-world-of-word-vectors/

Goal: Similar **Contexts** -> Similar **Vectors**

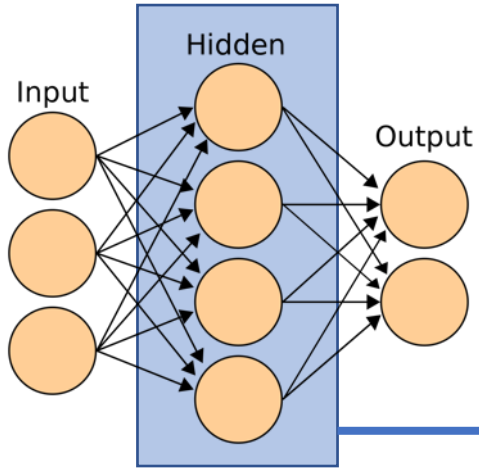
Nearby Words



Predict target word

Goal: Similar **Contexts** -> Similar **Vectors**

Nearby Words



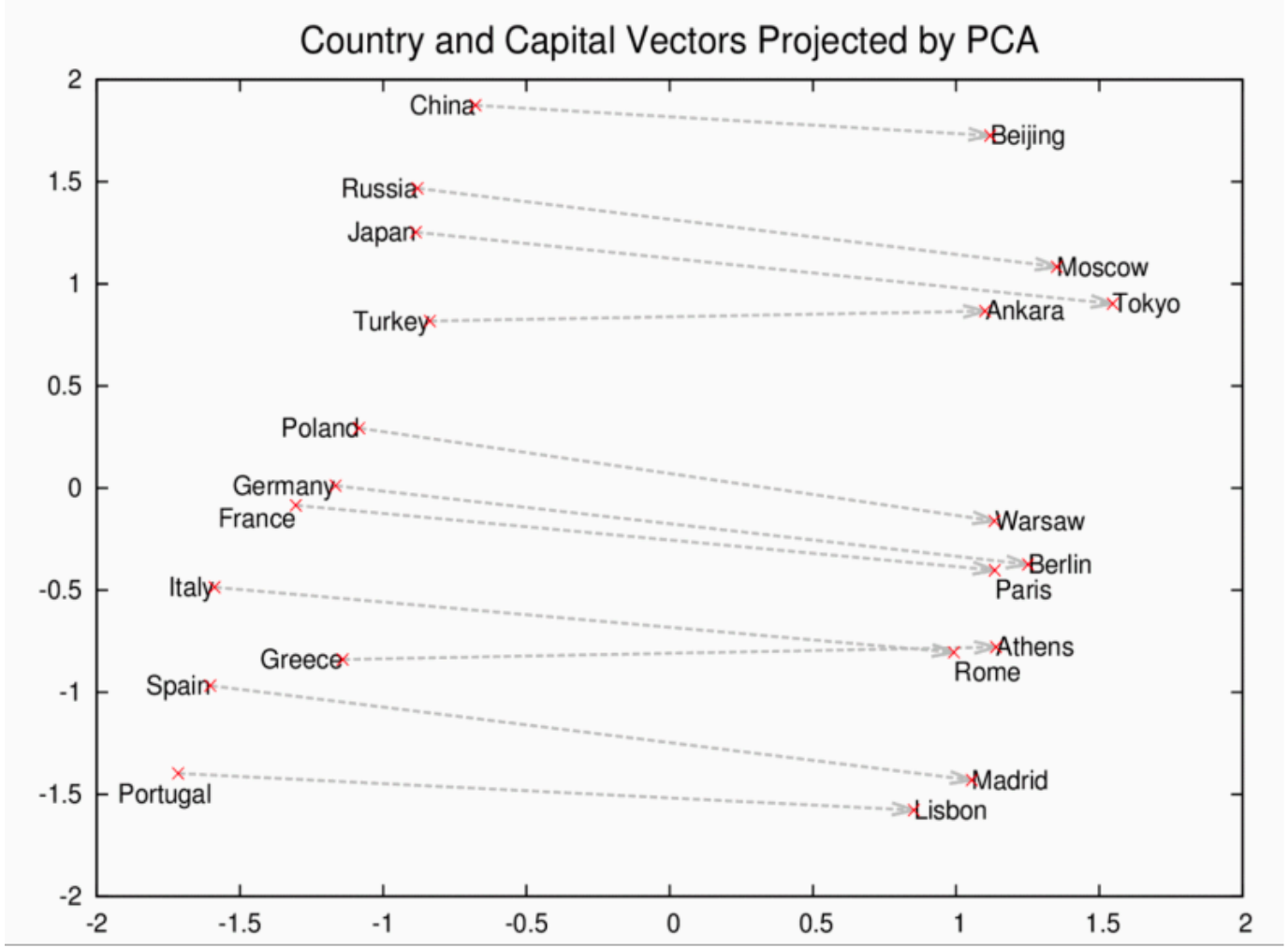
Predict target word

Vector Representation of Meaning

`DroidCon` → [4.3, 3.2, 1.4 , -2, 0.9, -1.1]

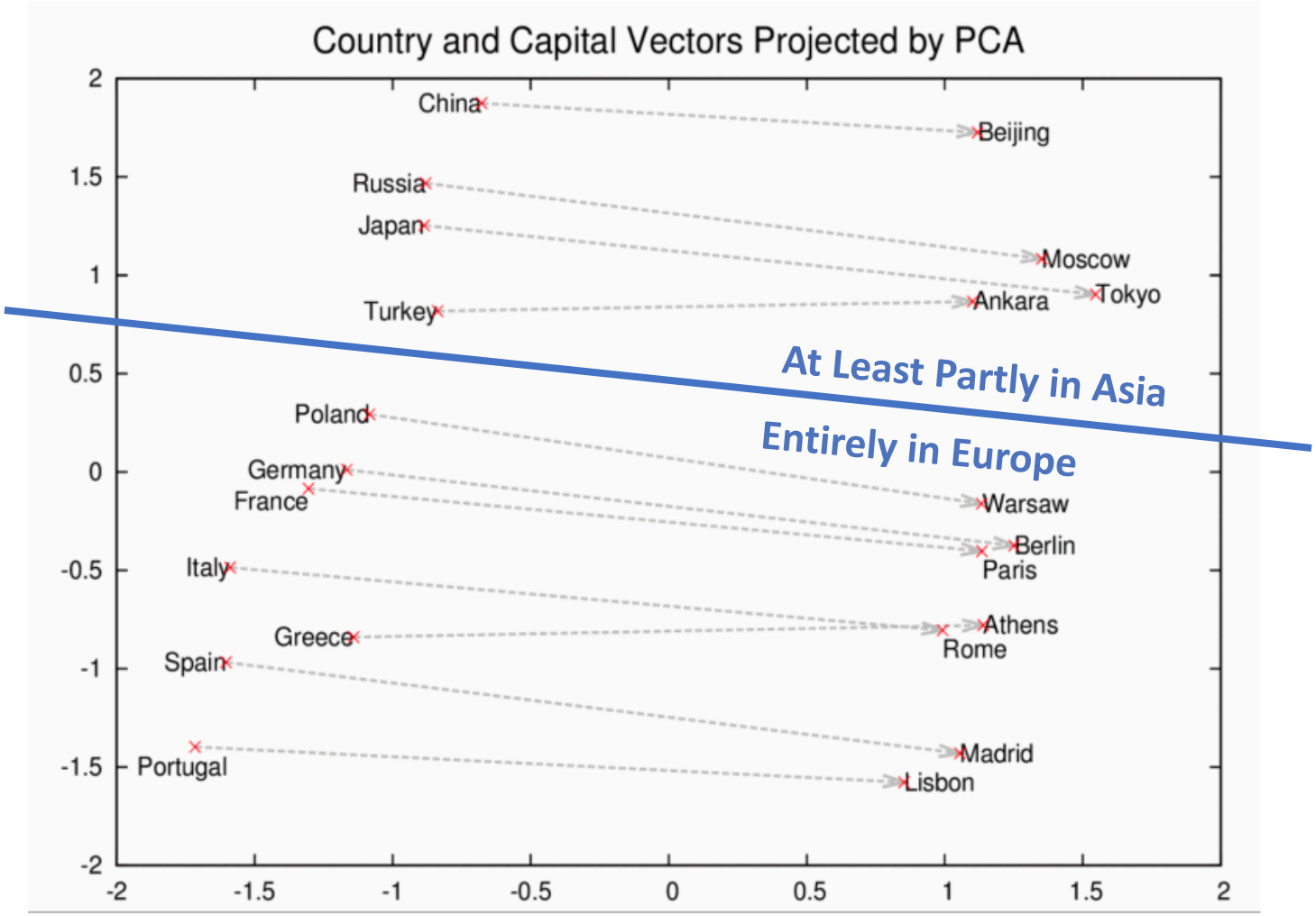
word2vec

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



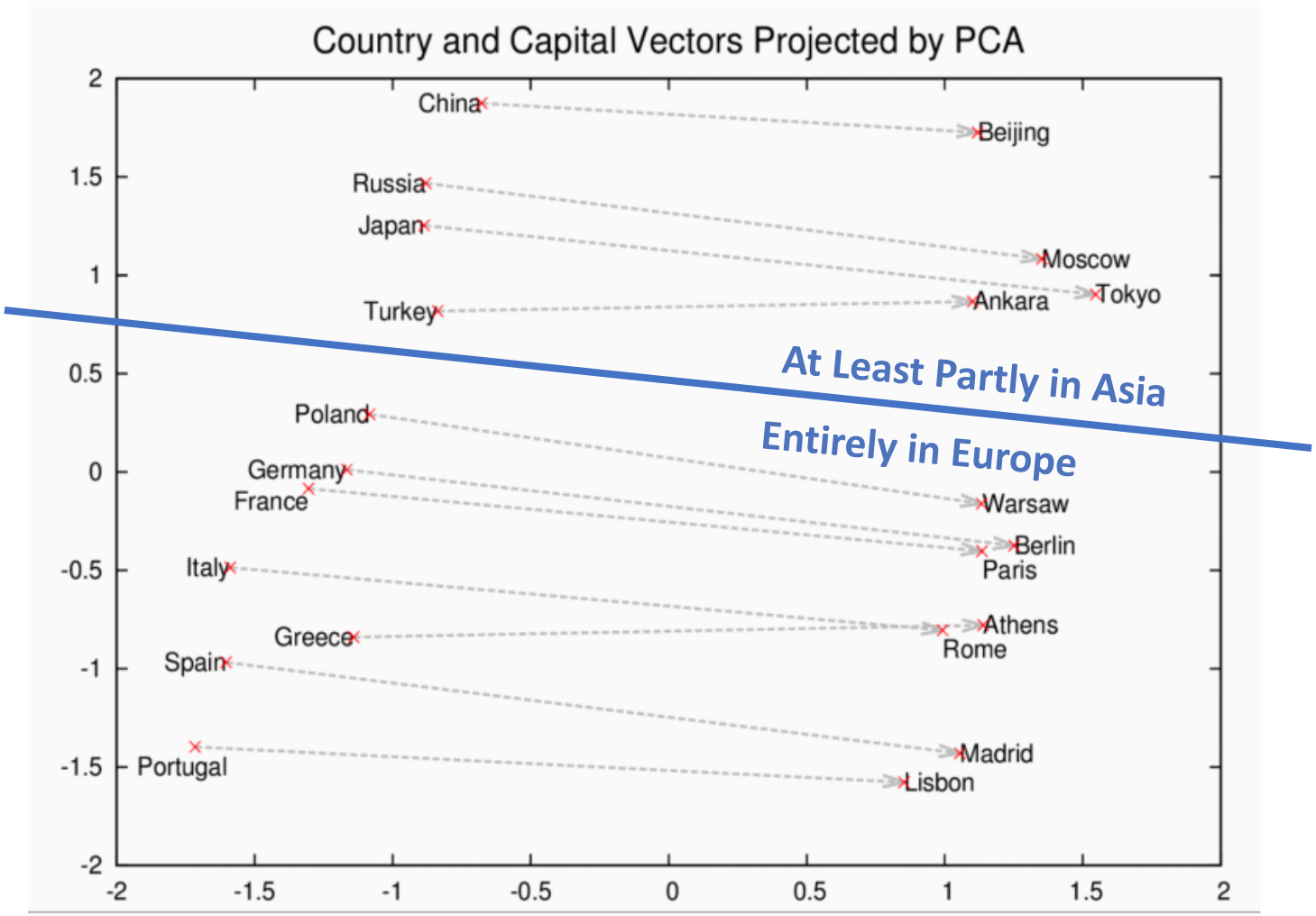
word2vec

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



word2vec

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

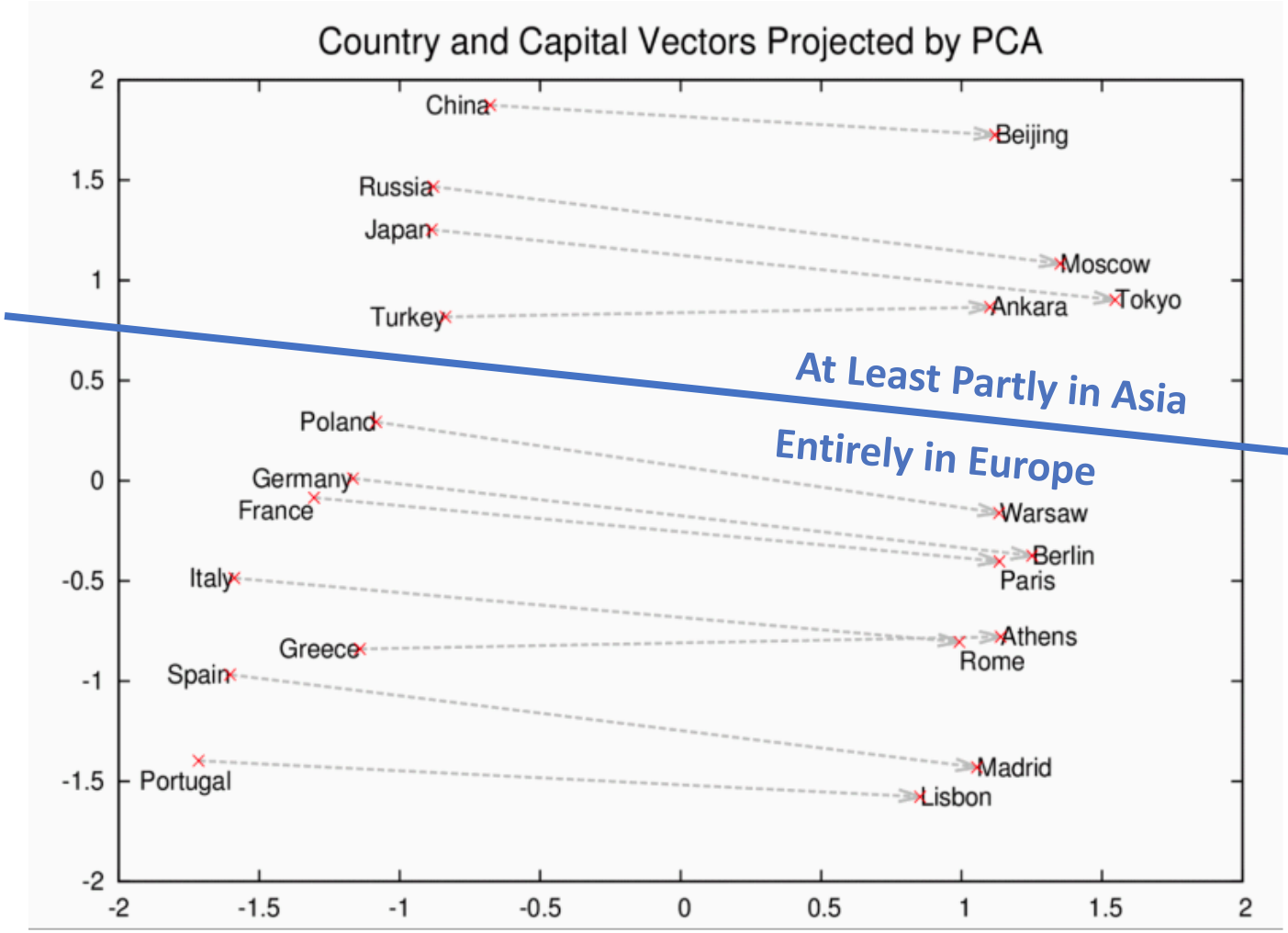


Distance Captures Similarity

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

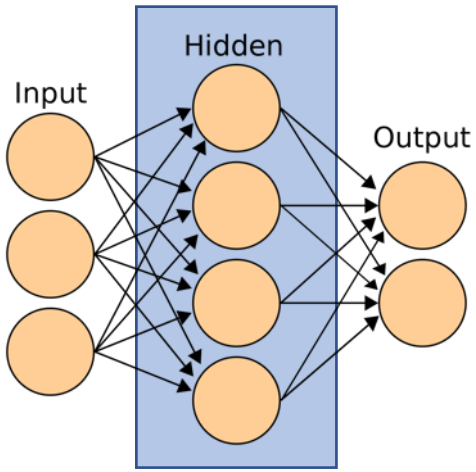
Math Creates Analogies

Russia – Moscow + Paris = France

(Russia:Moscow :: Paris:France)

Goal: Similar **Contexts** -> Similar **Vectors**

Paths



Predict **Method Name** from context

code2vec

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

Distance Captures Similarity

`count` is similar to `getCount`

Math Works Out

`equals` + `toLower` = `equalsIgnoreCase`

`remove` + `add` = `update`

`setHeaders` + `setRequestBody` = `createHttpPost`

Analogies

`open` : `connect` :: `close` : `disconnect`

`receive` : `download` :: `send` : `upload`

Labeling Functionality

```
while (iter.has_next()) {
  elem = iter.next();
  if(elem = v)
    return iter;
}
return null;
```

findElement

Other Applications of These Models

- Better variable names

```
while (iter.hasNext()) {
  elem = iter.next();
  if(elem = v)
    return iter;
}
return null;
```

```
while (iter.hasNext()) {
  curr_elem = iter.next();
  if(curr_elem = v)
    return iter;
}
return null;
```

- Code comments

```
// search collection for v
```

- Code completion

```
while (iter.has_next())
  // search for v with iter
return null;
```

```
while (iter.hasNext()) {
  elem = iter.next();
  if(elem = v)
    return iter;
}
return null;
```

Other Applications of These Models

- Better method names

getElem  find / search

- Correction of mis-remembered APIs

```
while (!iter.isEmpty()) {
  …
}
```

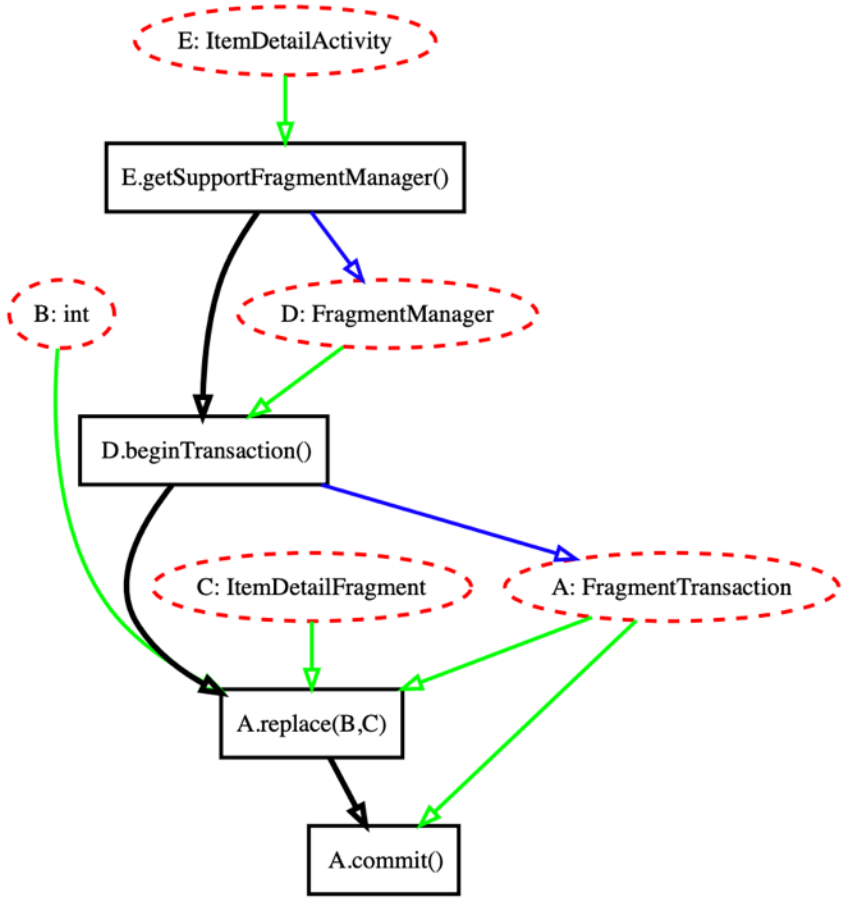
```
while
(iter.hasNext()) {
  …
}
```

Not uncommon to have one model support multiple applications.

Richer Representations

There’s more to code than syntax.

control + data flow



Application: Learn API Usage Patterns

Mining Framework Usage Graphs from App Corpora
https://github.com/cuplv/biggroum

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

Try It!

- http://askbayou.com/
- https://code2vec.org/
- https://code2seq.org/
- https://github.com/src-d/awesome-machine-learning-on-source-code

Realism

"a young boy is holding a baseball bat."

Contact Me

Twitter:
@stephenmagill

Web:
https://muse.dev