goto; chicago

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

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

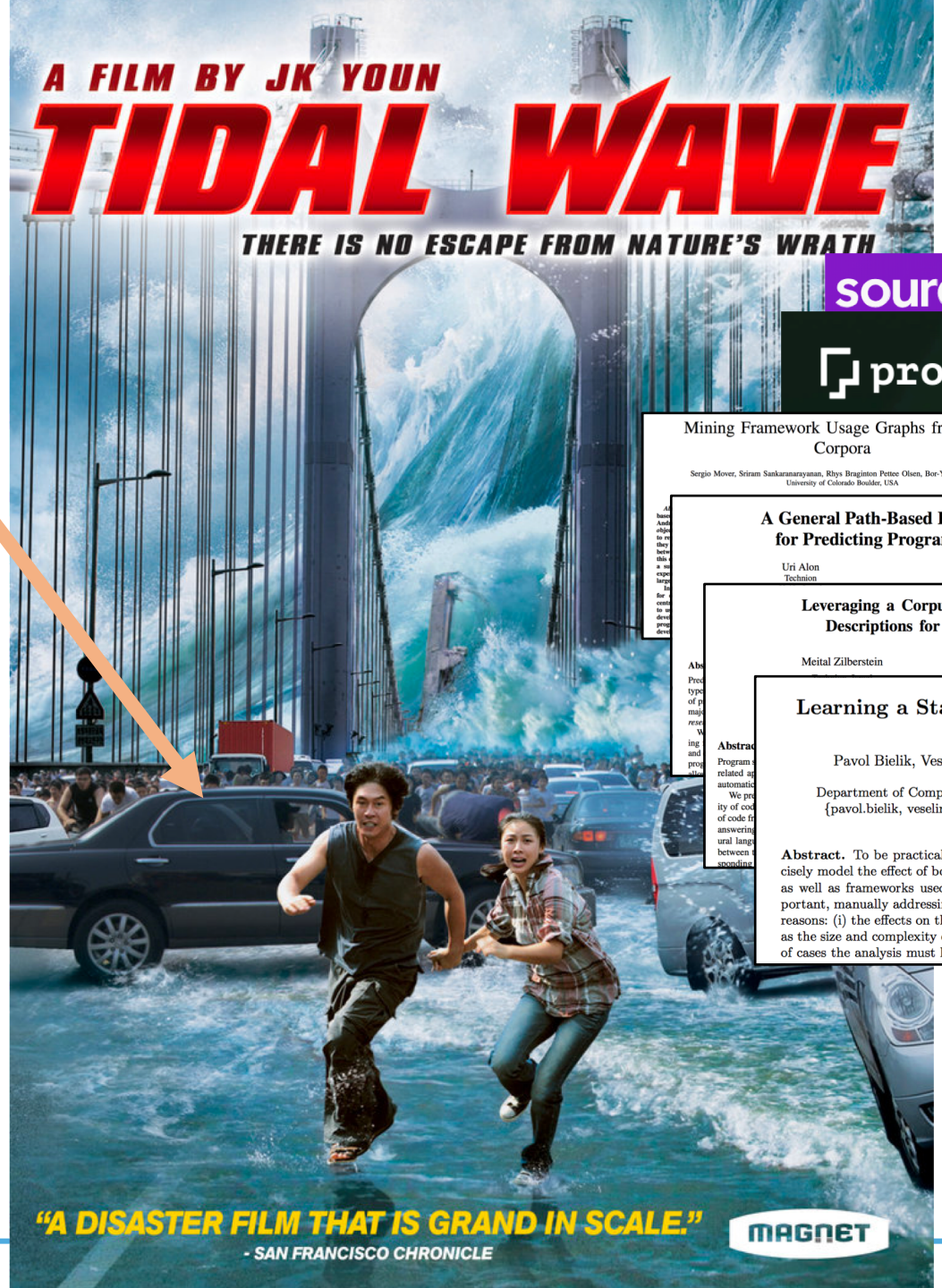
diffblue

DEEPCODE

source{d}

muse dev

Developers?



sourced

codota

prodo.ai

PCODE

ffblue

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

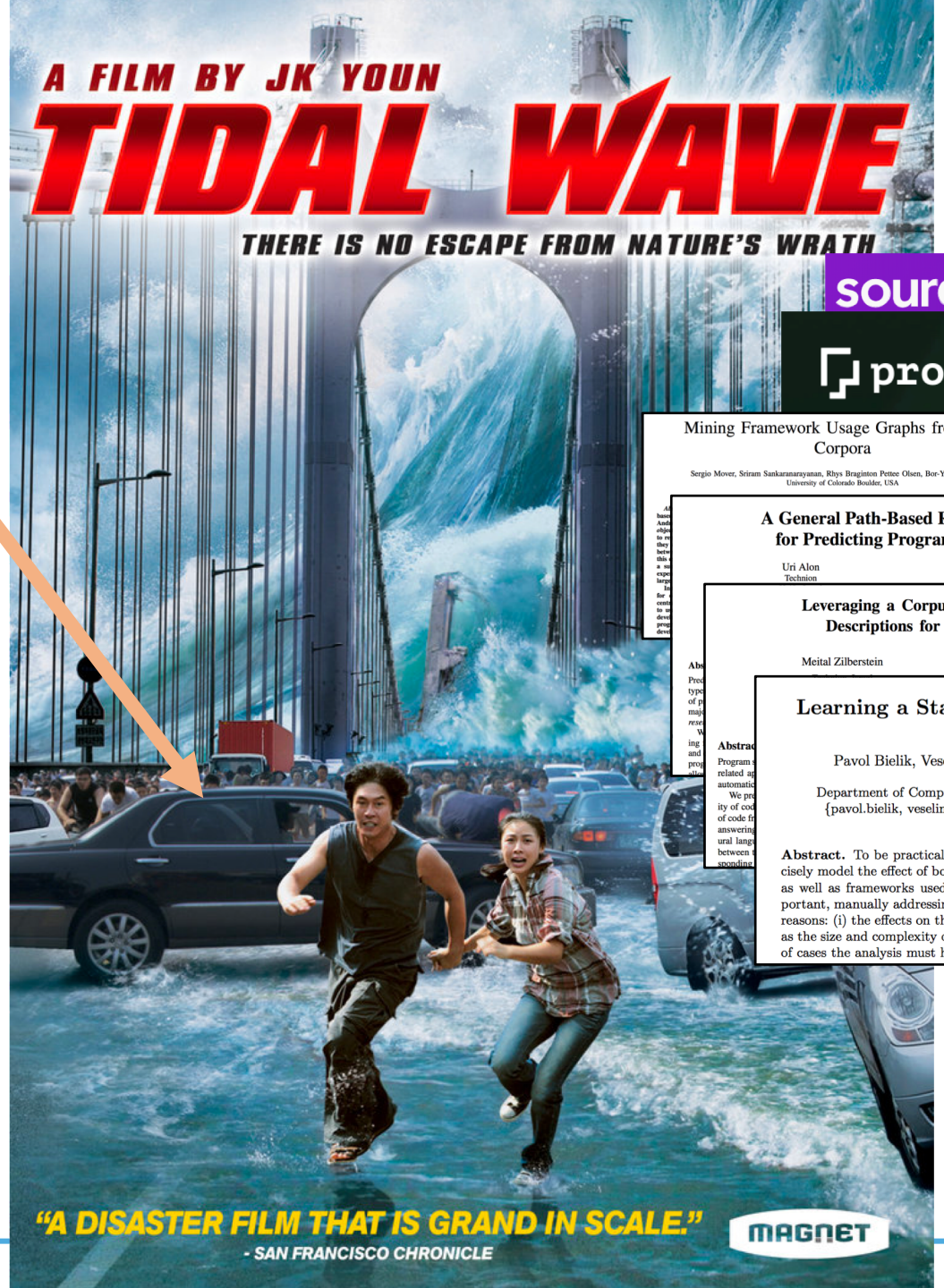
Department of Computer Science, ETH Zürich, Switzerland
{pavol.bielik, veselin.raychev, martin.vechev}@inf.ethz.ch

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

Developers?

… or developers?



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.

sourced
codota
prodo.ai
P.CODE
ffblue

muse dev

How Did We Get Here?

1 hour version: Easy!

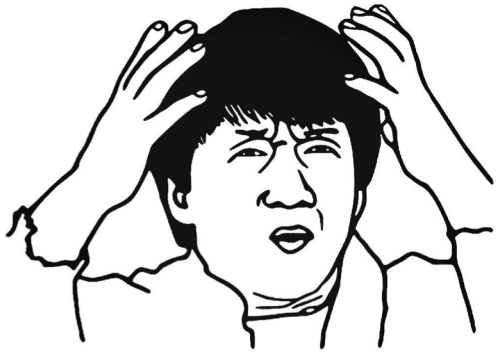
well…

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



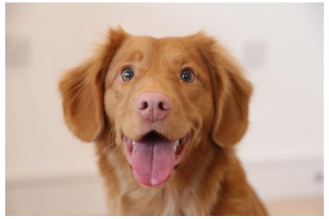
_____ : Images :: _____ : Code

Classification : Images :: _____ : Code

<u>ML Task</u>

Classification

or

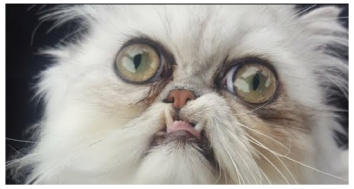


Classification : Images :: _____ : Code

<u>ML Task</u>

Classification

or



Normal Cat

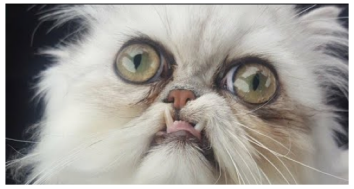
Memeable Cat

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

<u>ML Task</u>

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();
```

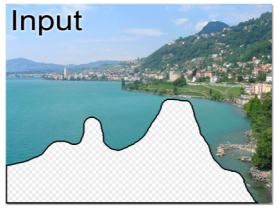
->

```
Scanner sc = new Scanner(new File(file));
st = sc.nextLine();
```

Completion : Images :: _____ : Code

<u>ML Task</u>

Image Completion

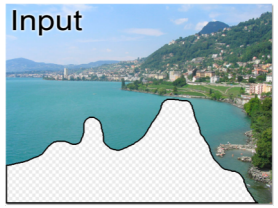


->

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
   }
}
```

->

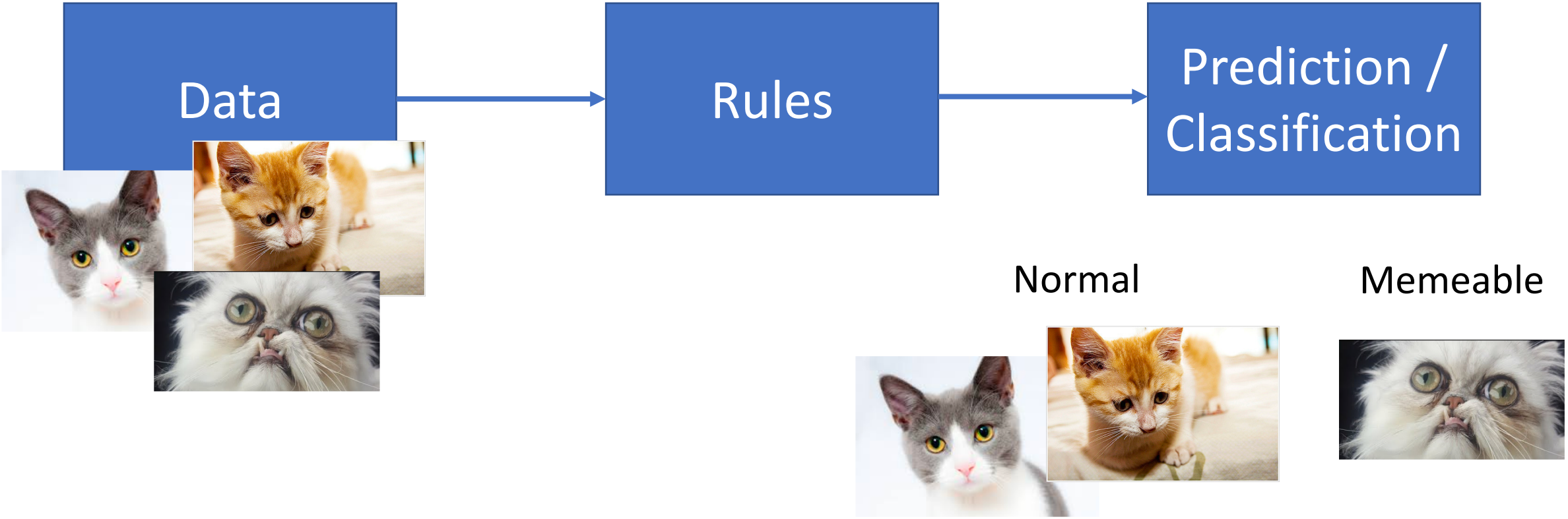
```
FileReader fr1;
BufferedReader br1;
String s1;
fr1 = new FileReader(file);
br1 = new BufferedReader(fr1);
while ((s1 = br1.readLine()) != null) {}
br1.close();
```

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

What is Machine Learning?

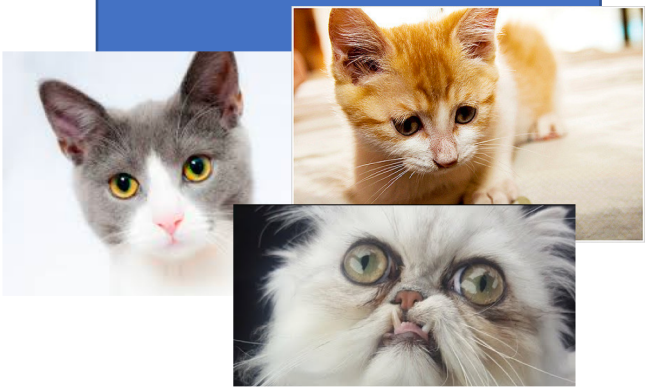
Deep Learning $\subset$ ANNs $\subset$ ML $\subset$ AI

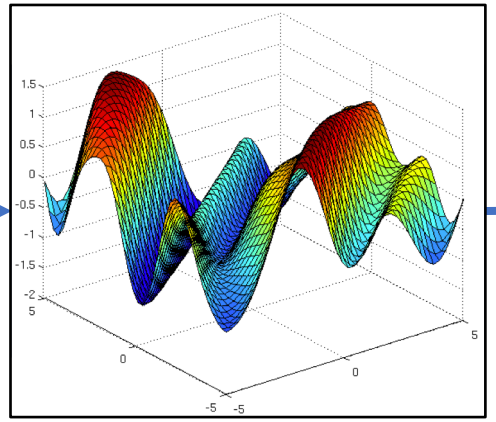
Artificial Intelligence



Machine Learning

Data



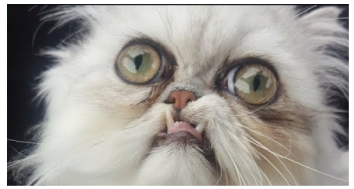


http://katbailey.github.io/post/gaussian-processes-for-dummies/

Prediction / Classification

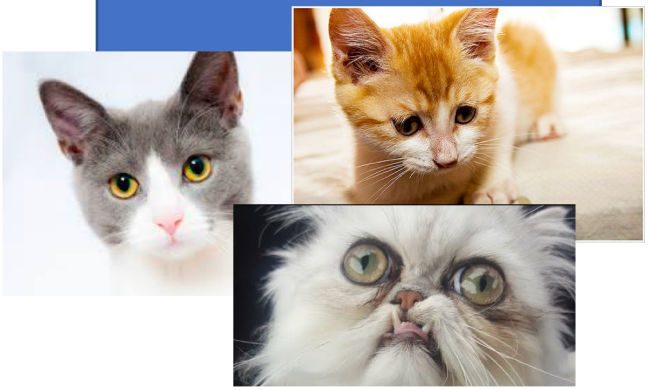
Normal

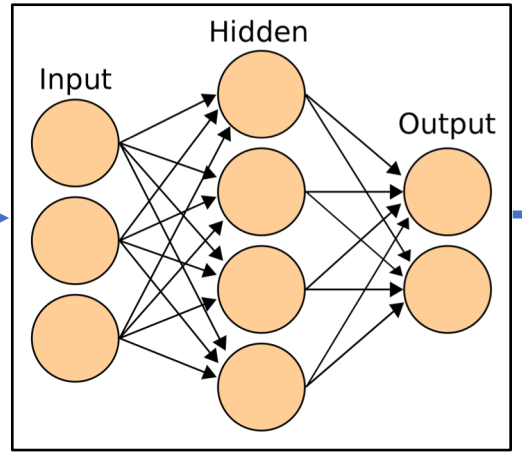
Memeable



Artificial Neural Networks

Data

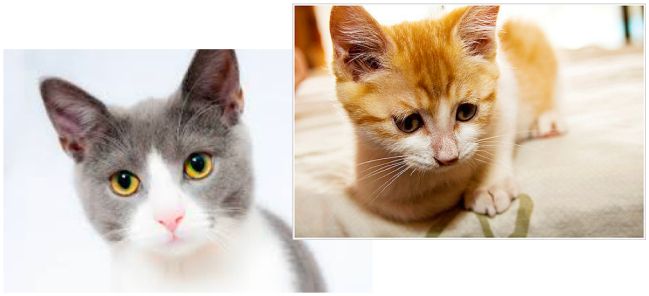




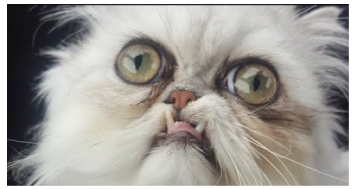
https://www.analyticsvidhya.com/blog/2014/10/ann-work-simplified/

Prediction / Classification

Normal

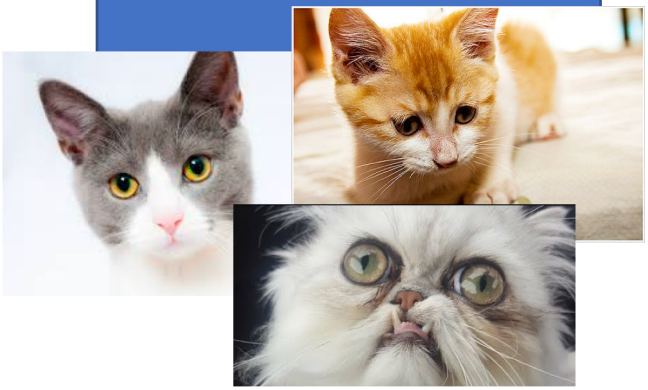


Memeable

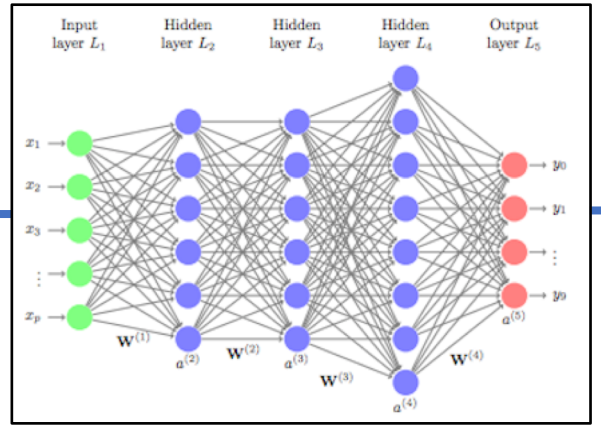


Deep Learning

Data



Lots (Millions) of Images

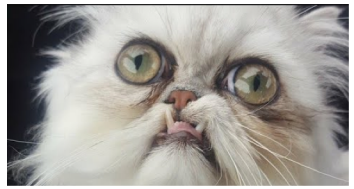


http://uc-r.github.io/feedforward_DNN

Prediction / Classification

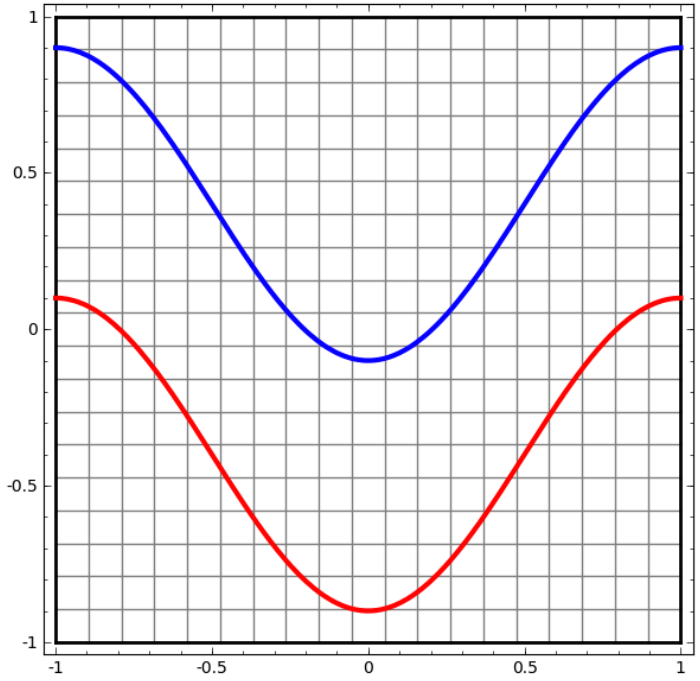
Normal

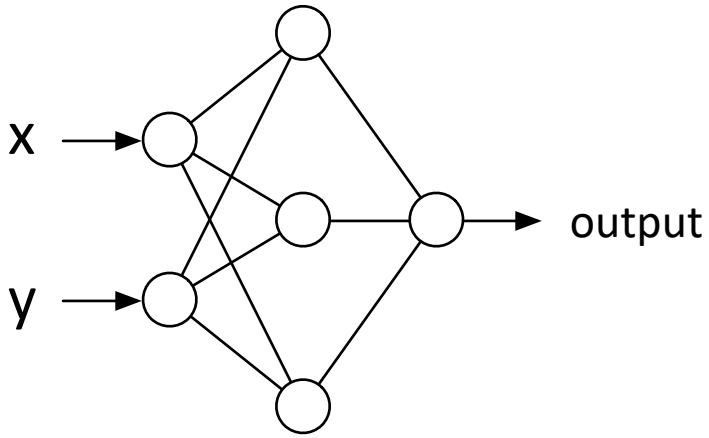
Memeable



How Do Neural Networks Work?

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

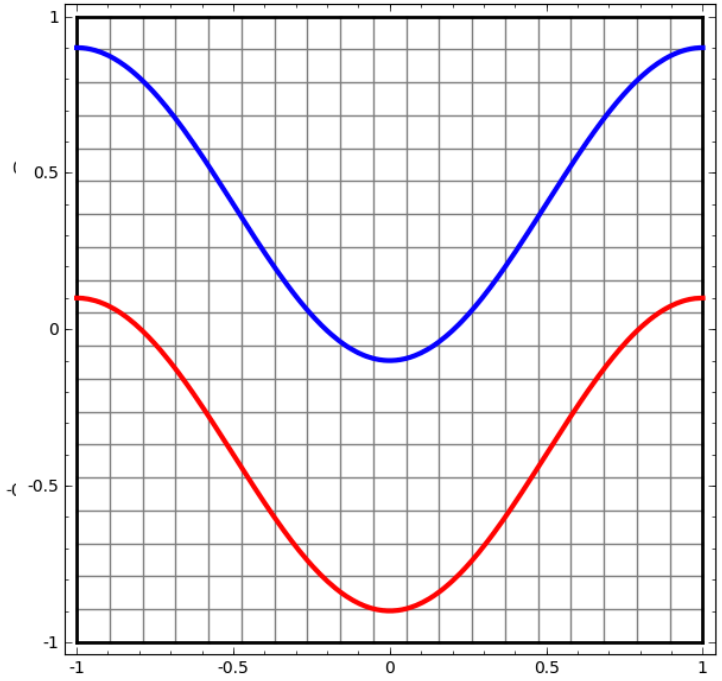


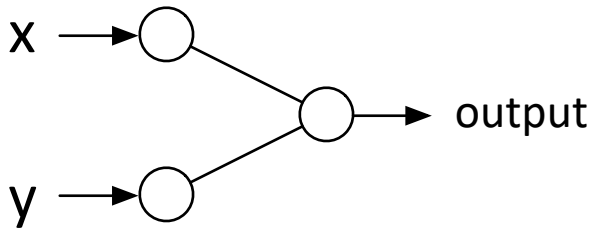


Red if output < 0, blue otherwise

How Do Neural Networks Work?

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



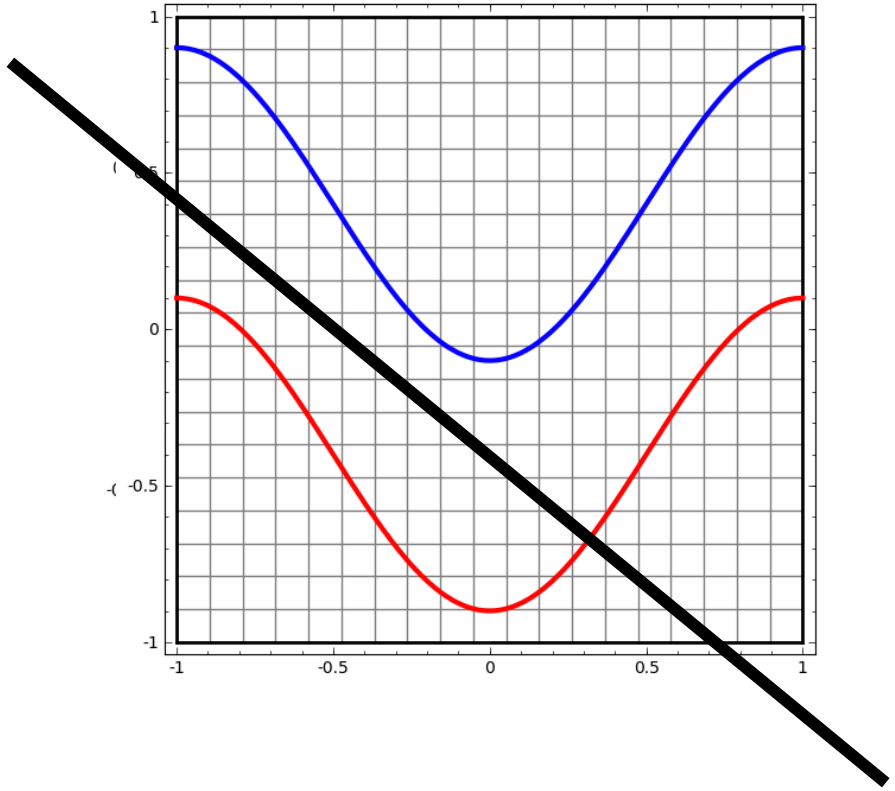


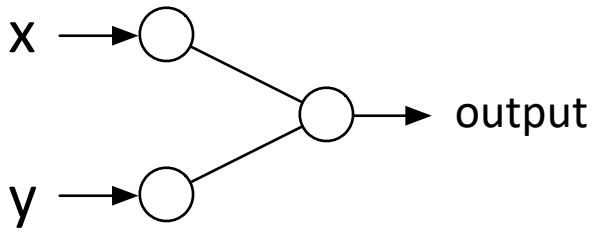
$$\text{output} = w_0 \cdot x + w_1 \cdot y$$

Red if output < 0, blue otherwise

How Do Neural Networks Work?

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



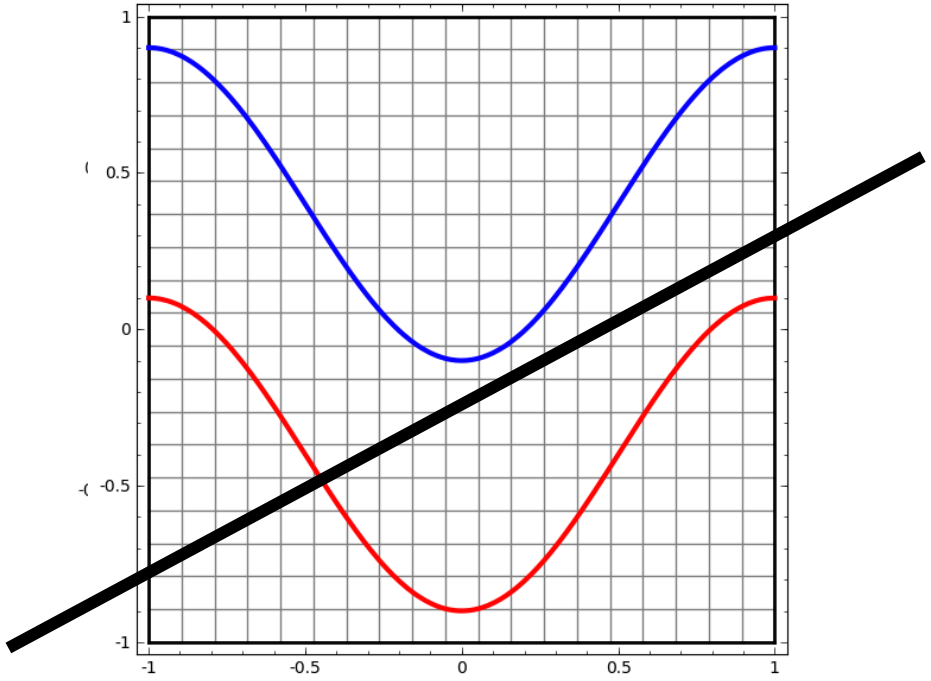


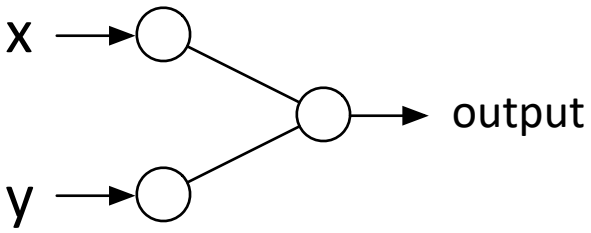
$$\text{output} = w_0 \cdot x + w_1 \cdot y$$

Red if output < 0, blue otherwise

How Do Neural Networks Work?

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



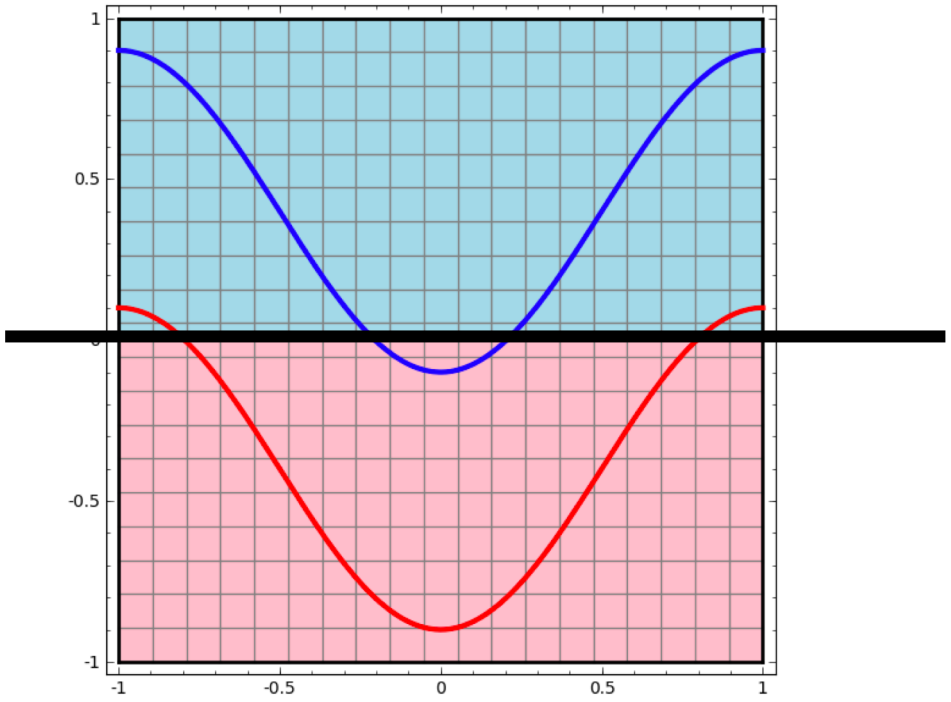


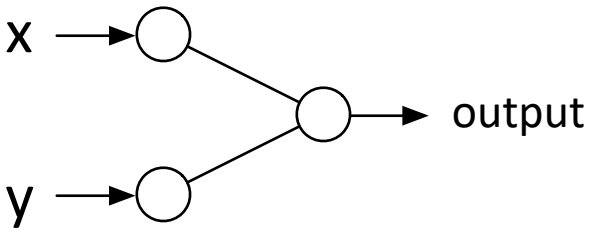
$$\text{output} = w_0 \cdot x + w_1 \cdot y$$

Red if output < 0, blue otherwise

How Do Neural Networks Work?

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



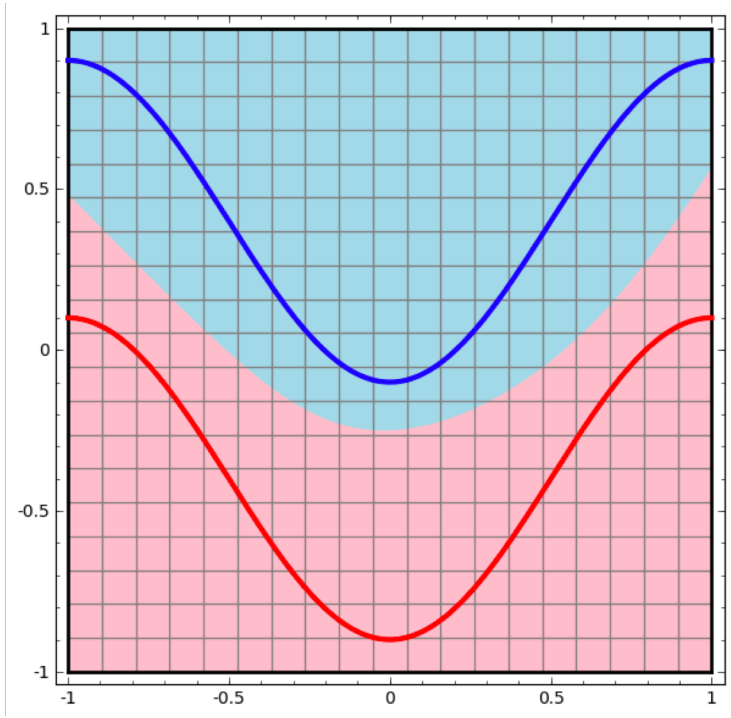


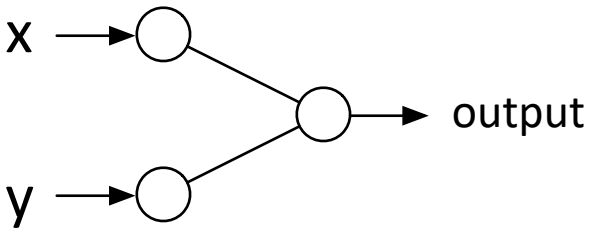
$$\text{output} = w_0 \cdot x + w_1 \cdot y$$

Red if output < 0, blue otherwise

How Do Neural Networks Work?

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



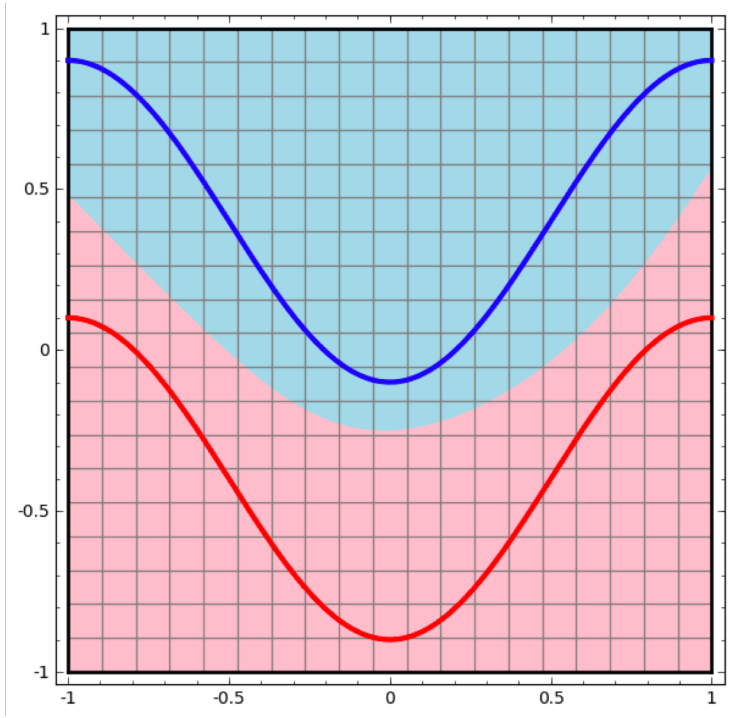


$$\text{output} = f(x, y)$$

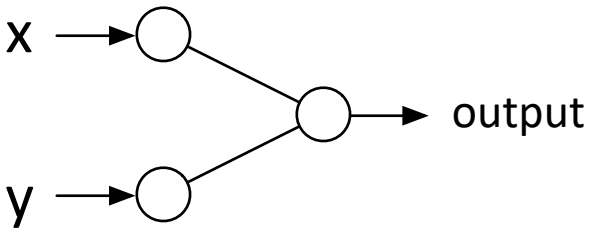
Red if output < 0, blue otherwise

How Do Neural Networks Work?

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



What is f?

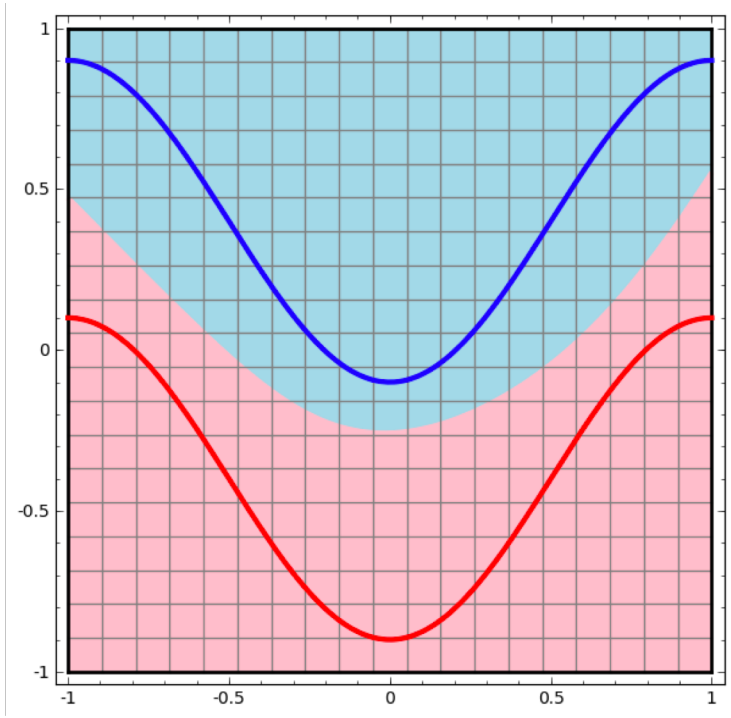


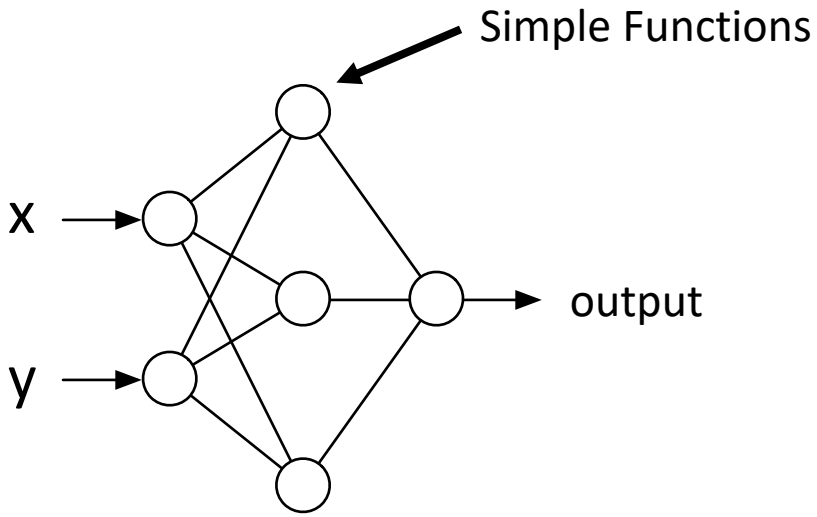
$$\text{output} = f(x, y)$$

Red if output < 0, blue otherwise

How Do Neural Networks Work?

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

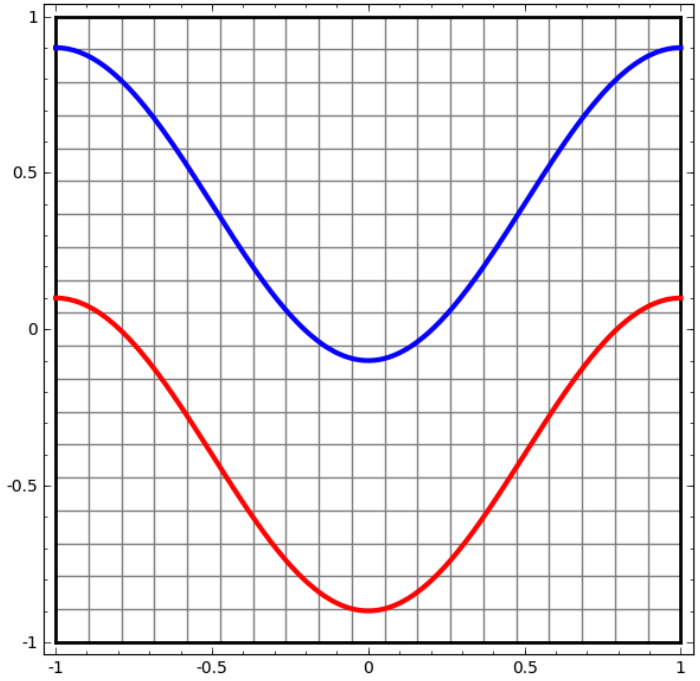


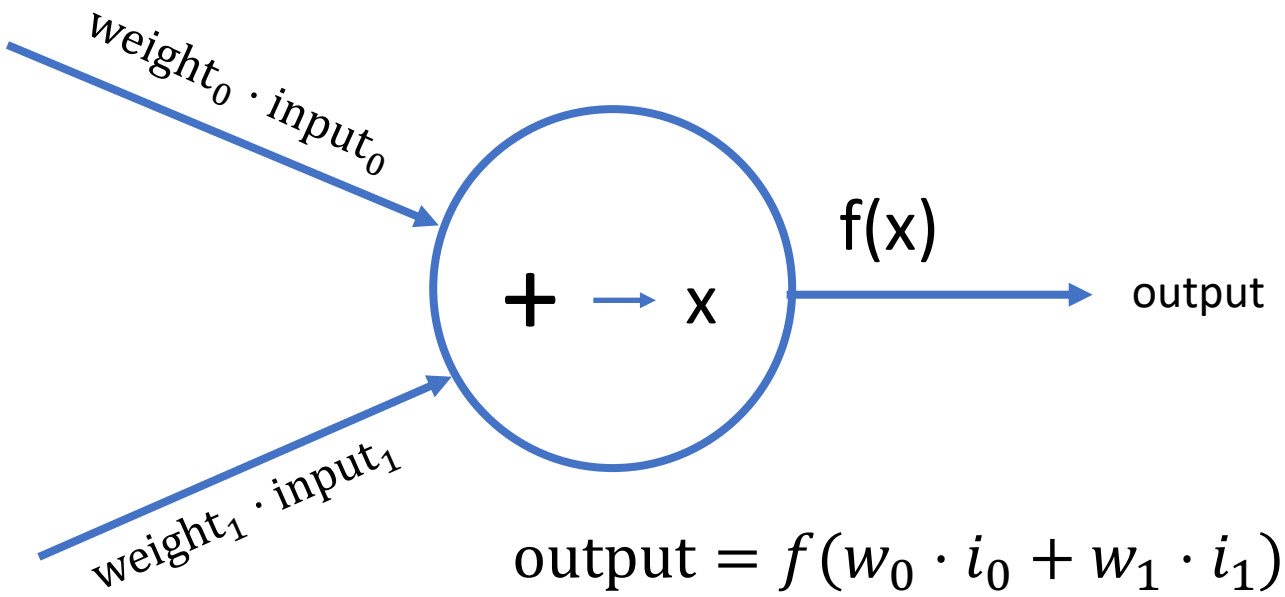


Red if output < 0, blue otherwise

How Do Neural Networks Work?

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

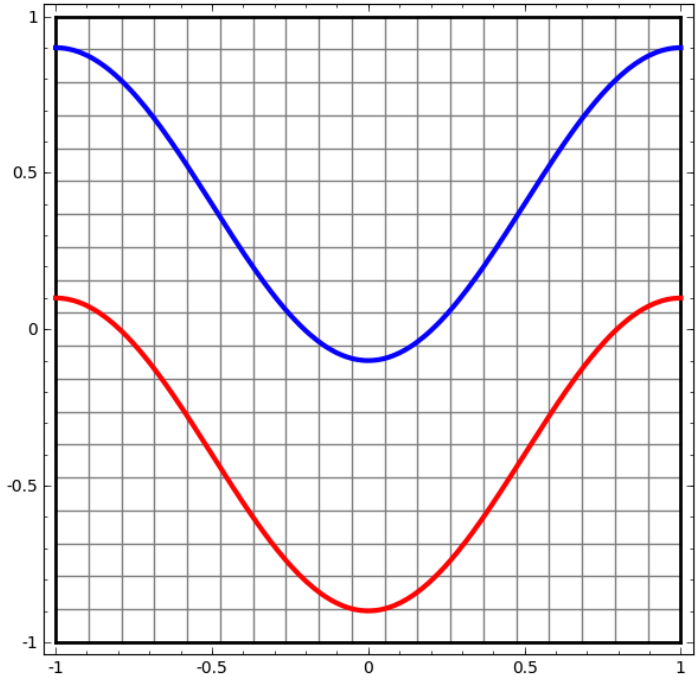


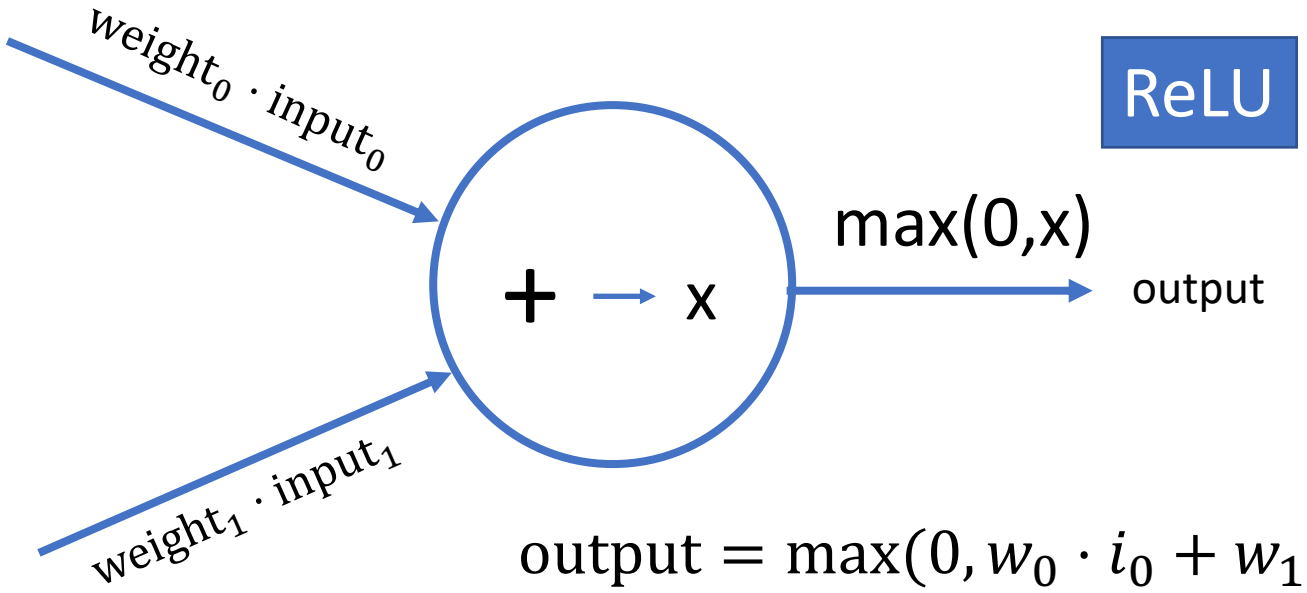


$$\text{output} = f(w_0 \cdot i_0 + w_1 \cdot i_1)$$

How Do Neural Networks Work?

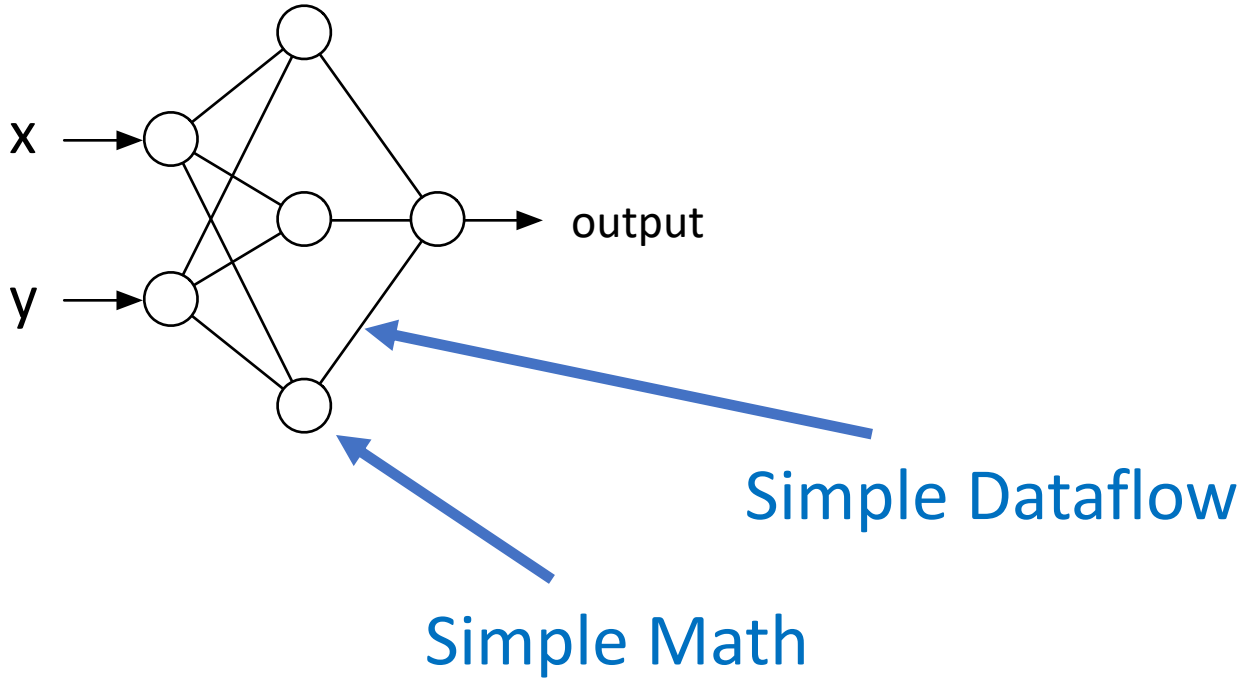
Images From: http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/



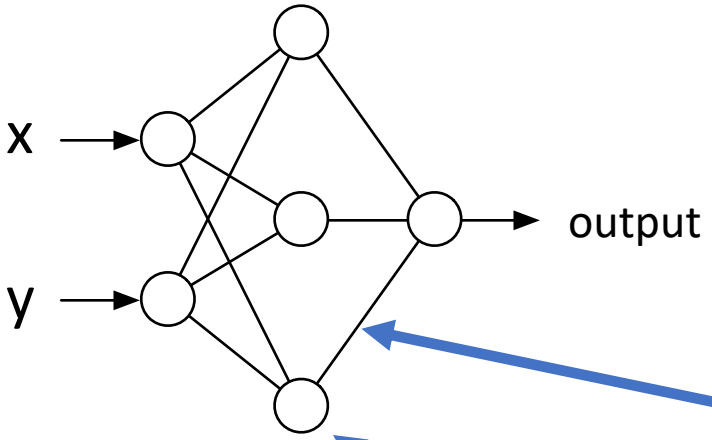


$$\text{output} = \max(0, w_0 \cdot i_0 + w_1 \cdot i_1)$$

Computationally – Dead Simple

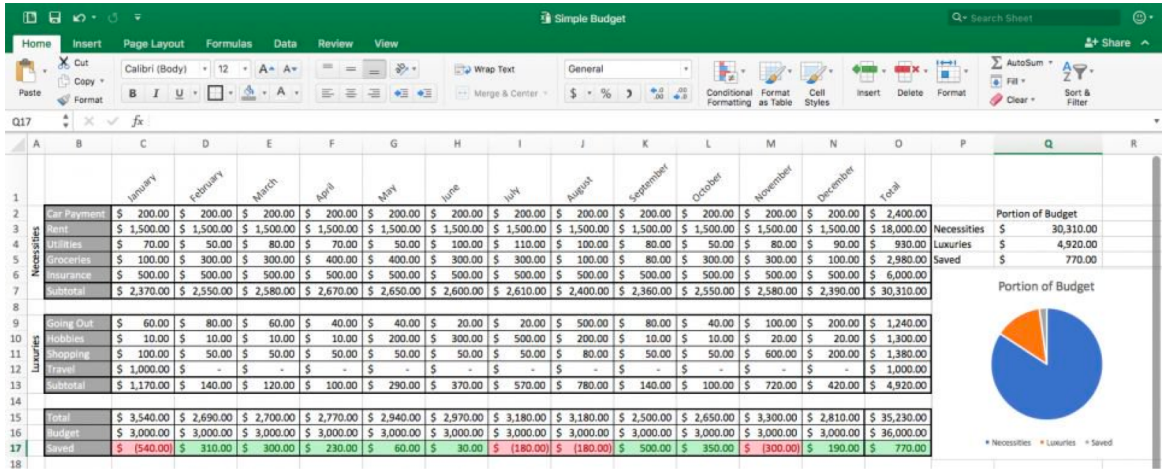


Computationally – Dead Simple



Simple Dataflow

Simple Math

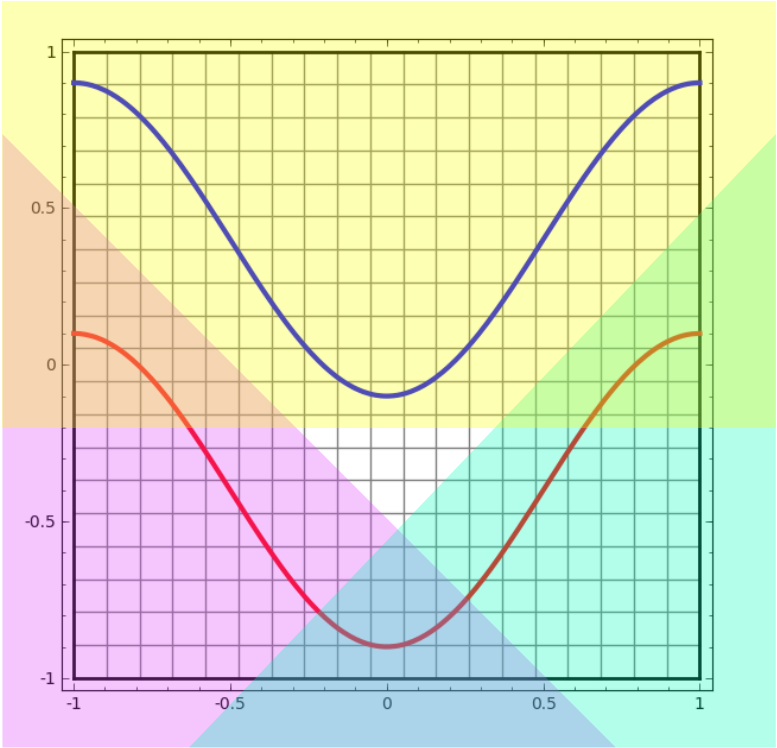


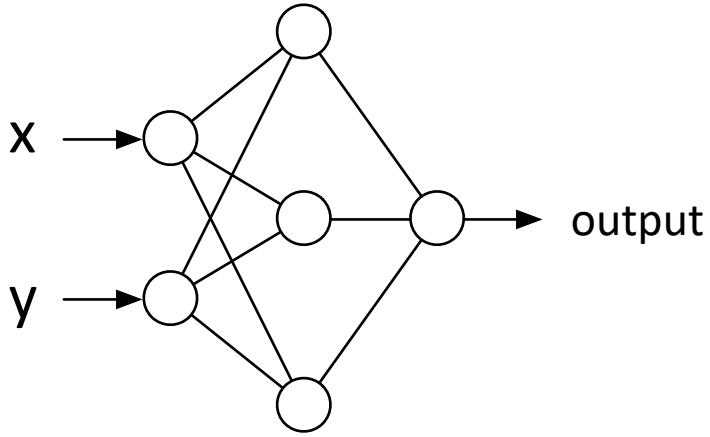
		January	February	March	April	May	June	July	August	September	October	November	December	Total
Necessities	Car Payment	$ 200.00	$ 200.00	$ 200.00	$ 200.00	$ 200.00	$ 200.00	$ 200.00	$ 200.00	$ 200.00	$ 200.00	$ 200.00	$ 200.00	$ 2,400.00
	Rent	$ 1,500.00	$ 1,500.00	$ 1,500.00	$ 1,500.00	$ 1,500.00	$ 1,500.00	$ 1,500.00	$ 1,500.00	$ 1,500.00	$ 1,500.00	$ 1,500.00	$ 1,500.00	$ 18,000.00
	Utilities	$ 70.00	$ 50.00	$ 80.00	$ 70.00	$ 50.00	$ 100.00	$ 110.00	$ 100.00	$ 80.00	$ 50.00	$ 80.00	$ 90.00	$ 930.00
	Groceries	$ 100.00	$ 300.00	$ 300.00	$ 400.00	$ 400.00	$ 300.00	$ 300.00	$ 100.00	$ 80.00	$ 300.00	$ 300.00	$ 100.00	$ 2,980.00
	Insurance	$ 500.00	$ 500.00	$ 500.00	$ 500.00	$ 500.00	$ 500.00	$ 500.00	$ 500.00	$ 500.00	$ 500.00	$ 500.00	$ 500.00	$ 6,000.00
	Subtotal	$ 2,370.00	$ 2,550.00	$ 2,580.00	$ 2,670.00	$ 2,650.00	$ 2,600.00	$ 2,610.00	$ 2,400.00	$ 2,360.00	$ 2,550.00	$ 2,580.00	$ 2,390.00	$ 30,310.00
Luxuries	Going Out	$ 60.00	$ 80.00	$ 60.00	$ 40.00	$ 40.00	$ 20.00	$ 20.00	$ 500.00	$ 80.00	$ 40.00	$ 100.00	$ 200.00	$ 1,240.00
	Hobbies	$ 10.00	$ 10.00	$ 10.00	$ 10.00	$ 200.00	$ 300.00	$ 500.00	$ 200.00	$ 10.00	$ 10.00	$ 20.00	$ 20.00	$ 1,300.00
	Shopping	$ 100.00	$ 50.00	$ 50.00	$ 50.00	$ 50.00	$ 50.00	$ 50.00	$ 80.00	$ 50.00	$ 50.00	$ 600.00	$ 200.00	$ 1,380.00
	Travel	$ 1,000.00	$ -	$ -	$ -	$ -	$ -	$ -	$ -	$ -	$ -	$ -	$ -	$ 1,000.00
	Subtotal	$ 1,170.00	$ 140.00	$ 120.00	$ 100.00	$ 290.00	$ 370.00	$ 570.00	$ 780.00	$ 140.00	$ 100.00	$ 720.00	$ 420.00	$ 4,920.00
	Total	$ 3,540.00	$ 2,690.00	$ 2,700.00	$ 2,770.00	$ 2,940.00	$ 2,970.00	$ 3,180.00	$ 3,180.00	$ 2,500.00	$ 2,650.00	$ 3,300.00	$ 2,810.00	$ 35,230.00
	Budget	$ 3,000.00	$ 3,000.00	$ 3,000.00	$ 3,000.00	$ 3,000.00	$ 3,000.00	$ 3,000.00	$ 3,000.00	$ 3,000.00	$ 3,000.00	$ 3,000.00	$ 3,000.00	$ 36,000.00
	Saved	$ (540.00)	$ 310.00	$ 300.00	$ 230.00	$ 60.00	$ 30.00	$ (180.00)	$ (180.00)	$ 500.00	$ 350.00	$ (300.00)	$ 190.00	$ 770.00

Portion of Budget	
Necessities	$ 30,310.00
Luxuries	$ 4,920.00
Saved	$ 770.00

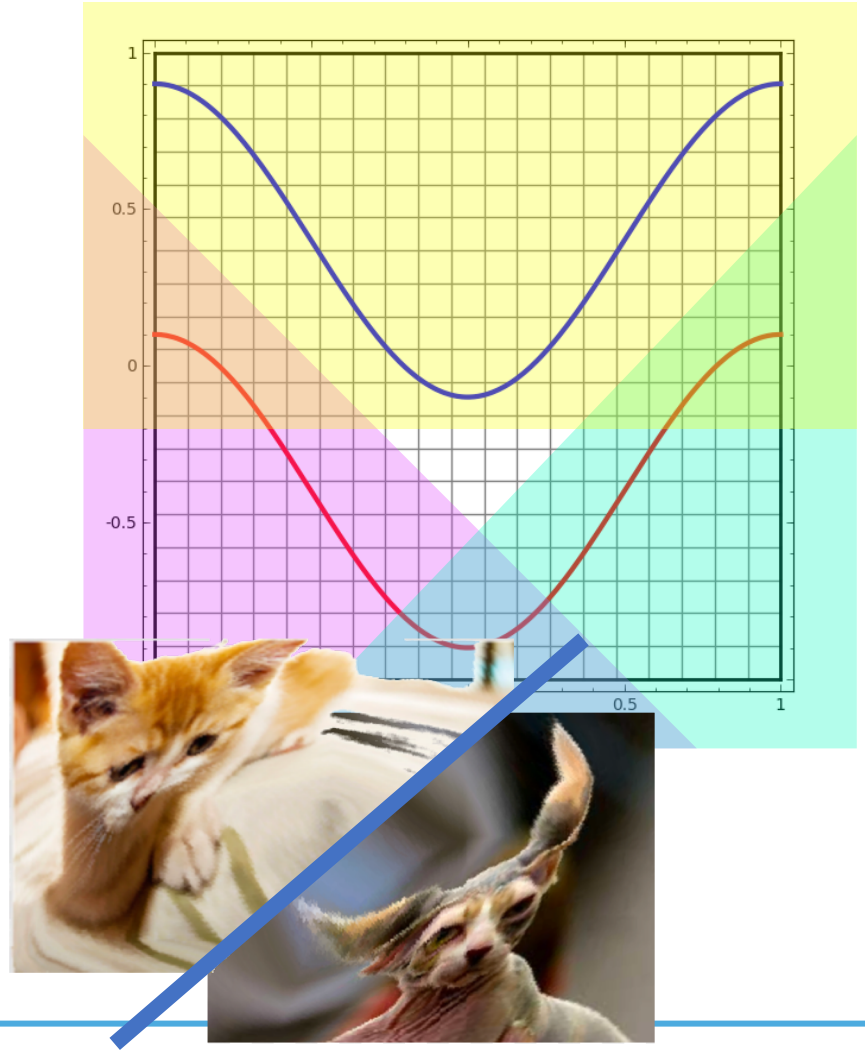
Demo Time!

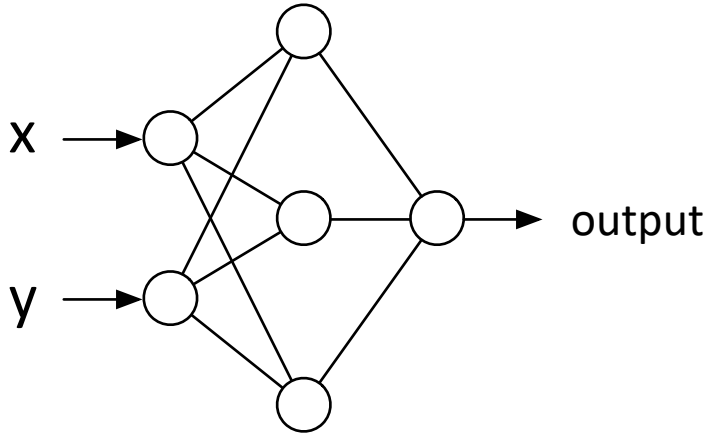
Learning Boundaries



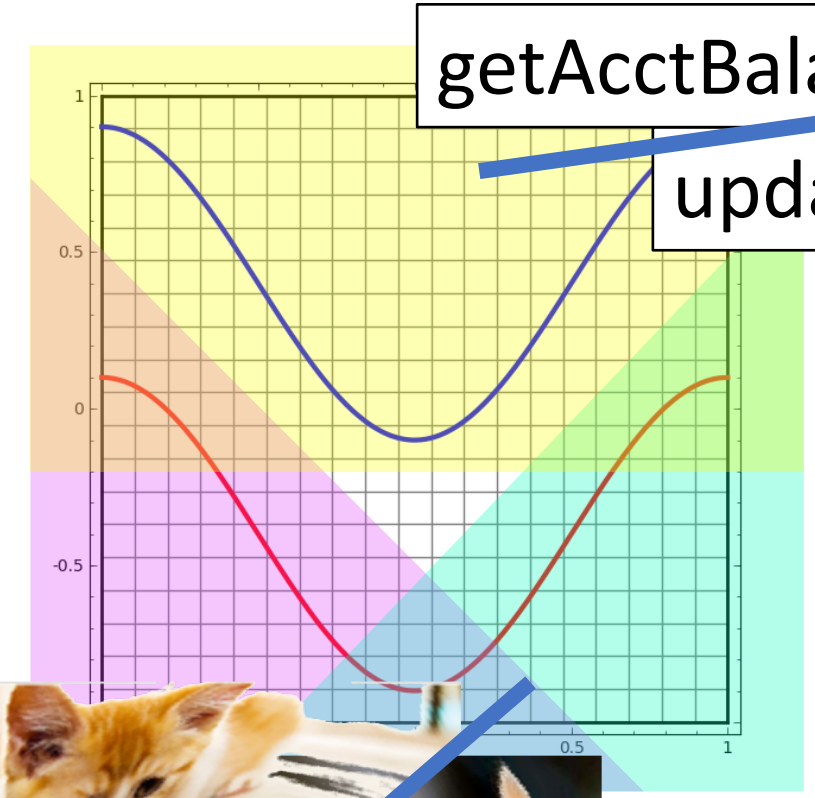


Learning Boundaries



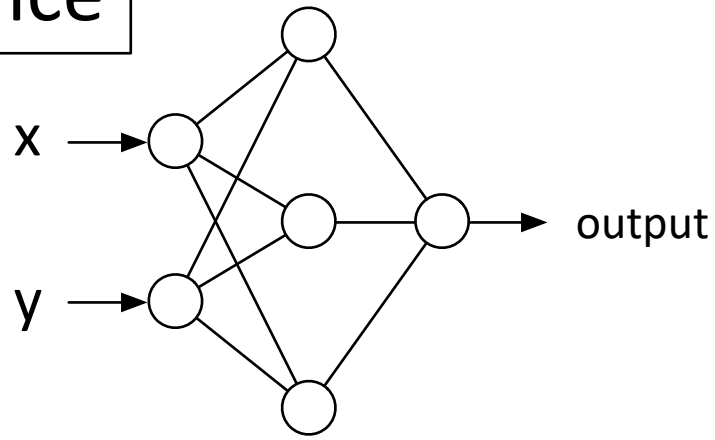


Learning Boundaries



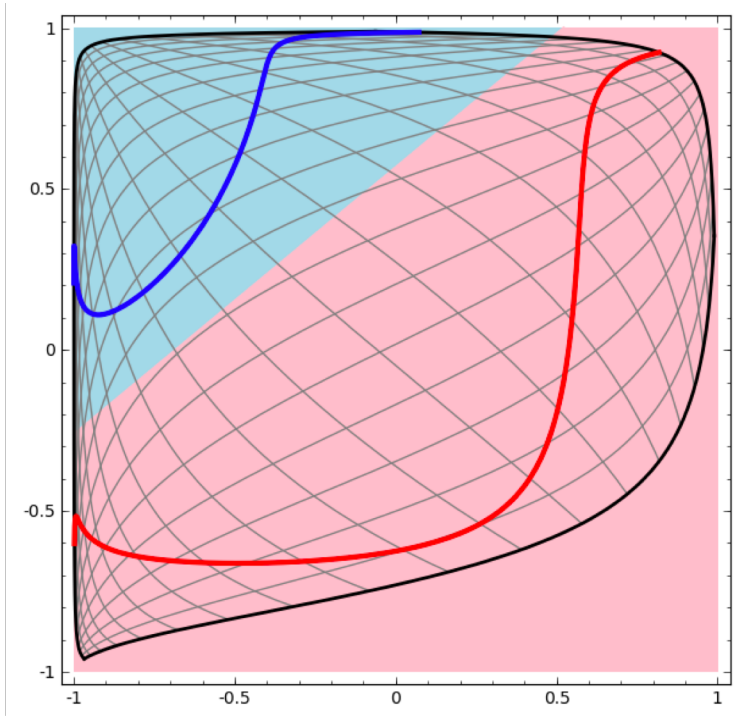
getAcctBalance

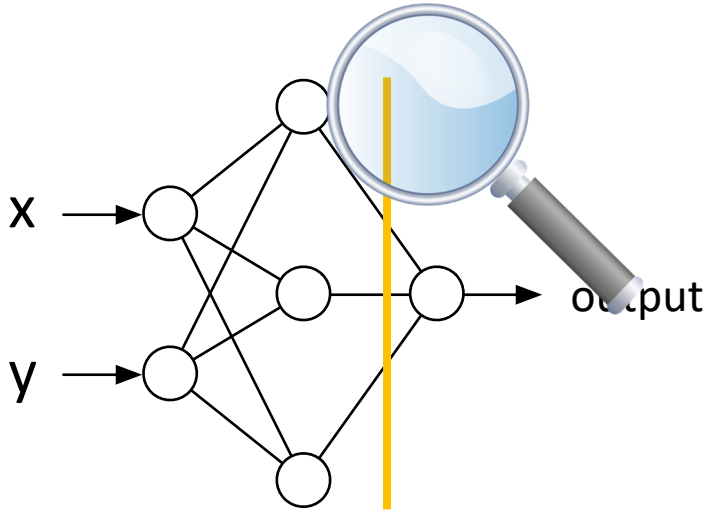
updateAcctBalance



How Do Neural Networks Work?

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



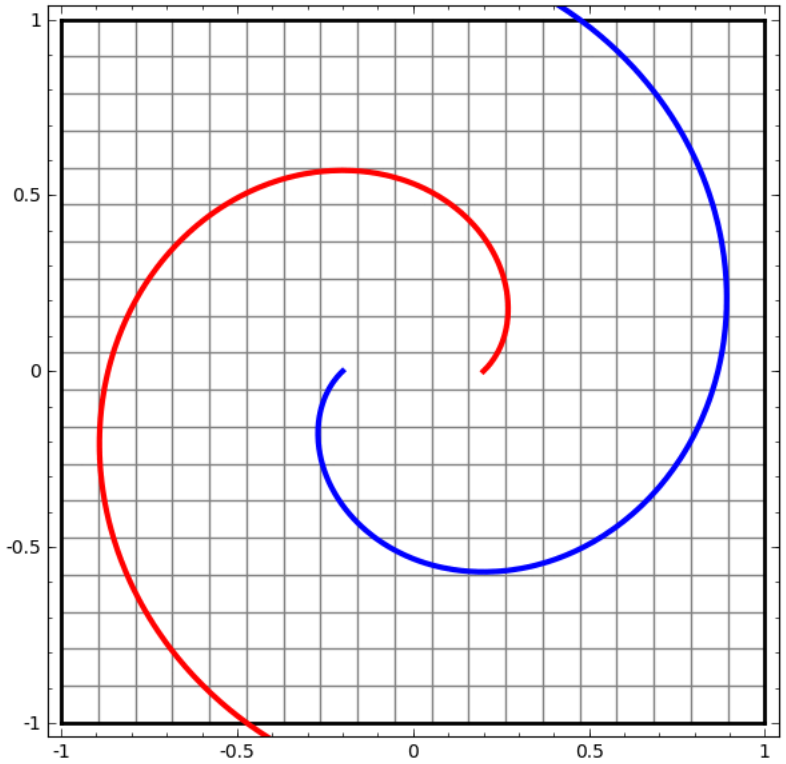


$$\text{output} = w_0 \cdot i_0 + w_1 \cdot i_1 + w_2 \cdot i_2$$

Red if output < 0, blue otherwise

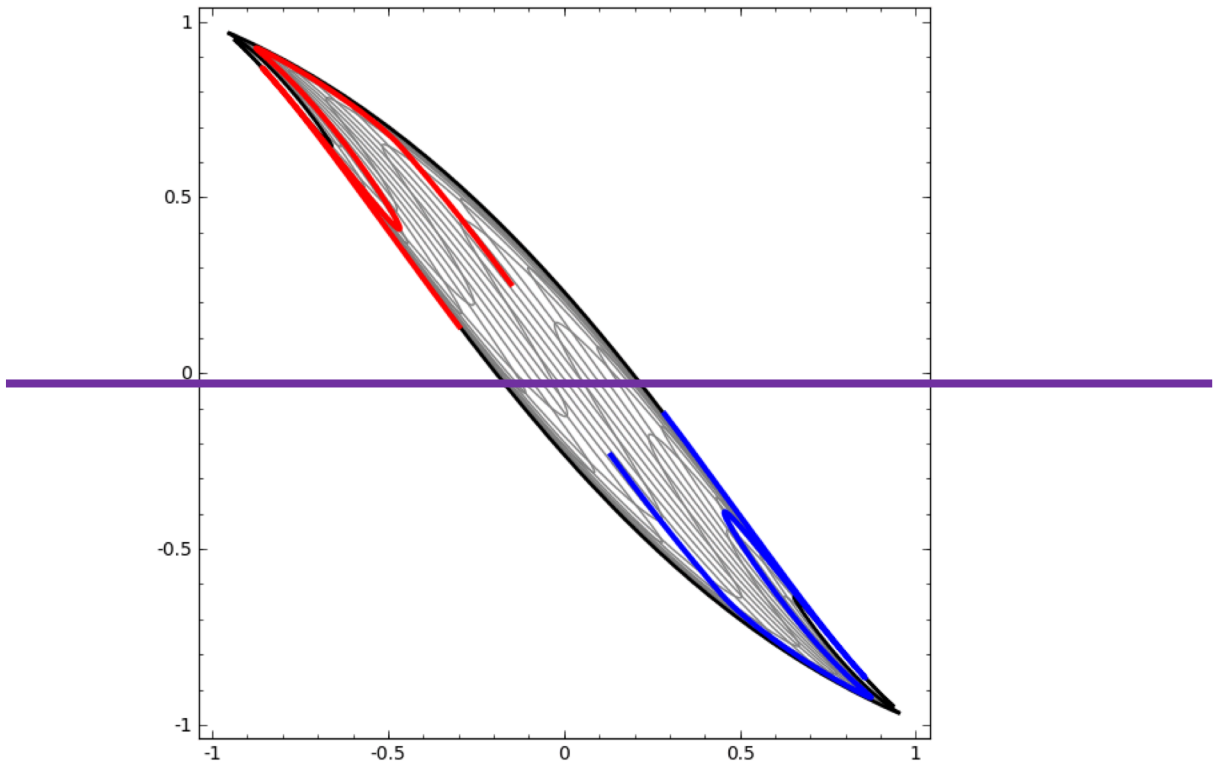
How Do Neural Networks Work?

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



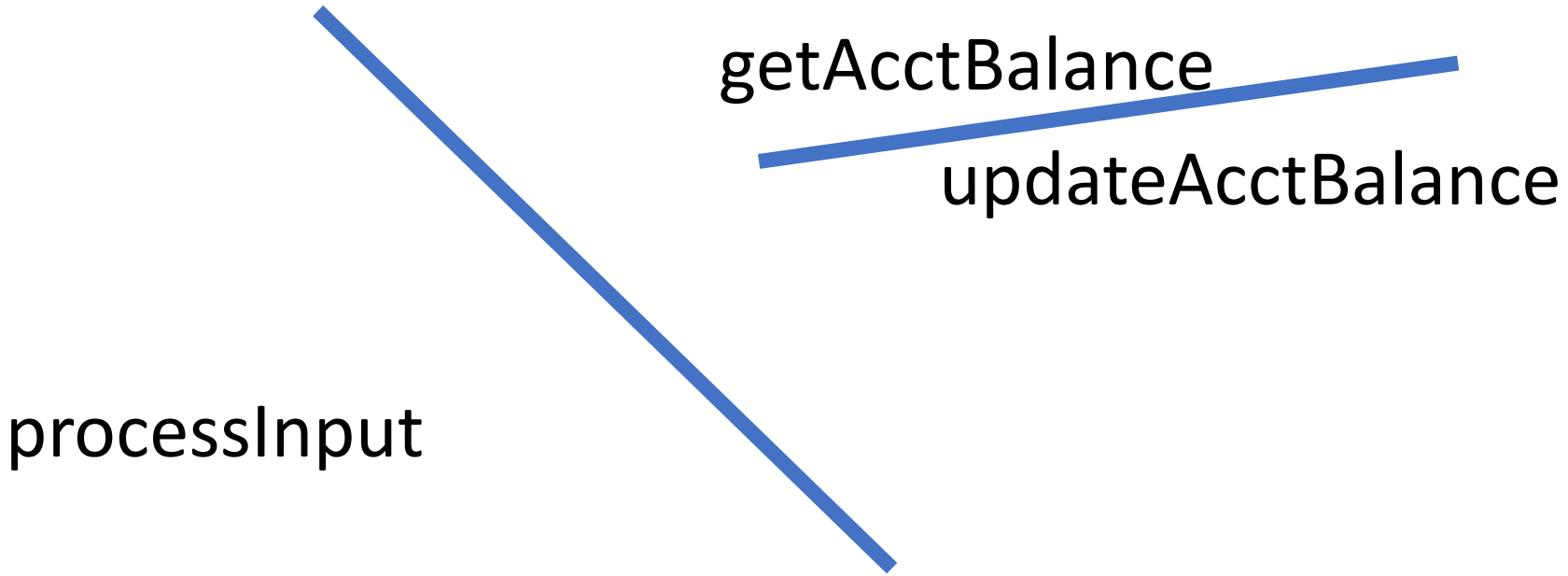
How Do Neural Networks Work?

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



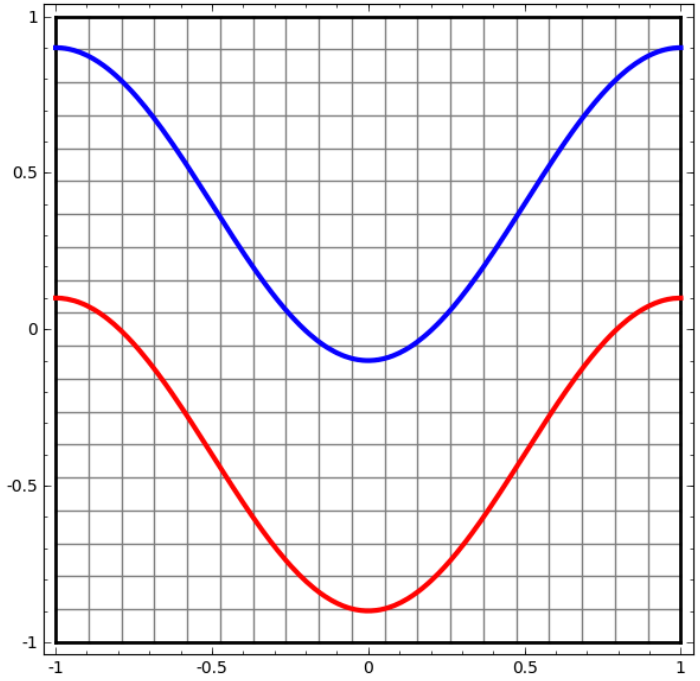
How Do Neural Networks Work?

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

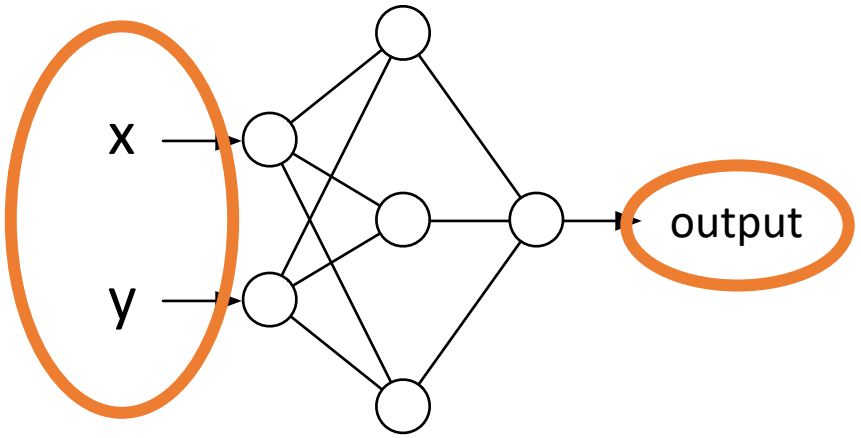


Neural Network Usage

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

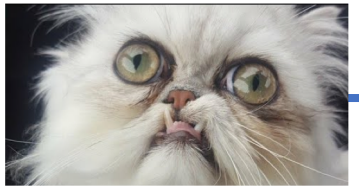


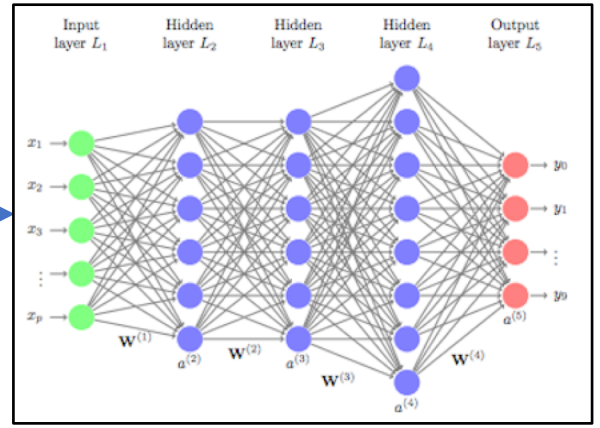
What goes here for images / code?



Red if output > 0, blue otherwise

Neural Network Input



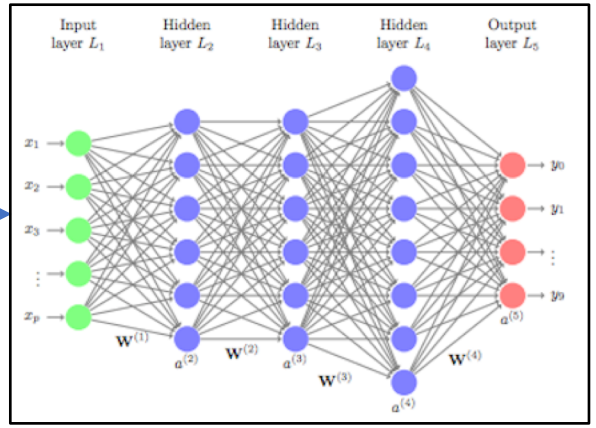


How to connect data to NN inputs?

Neural Networks for **Code**

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

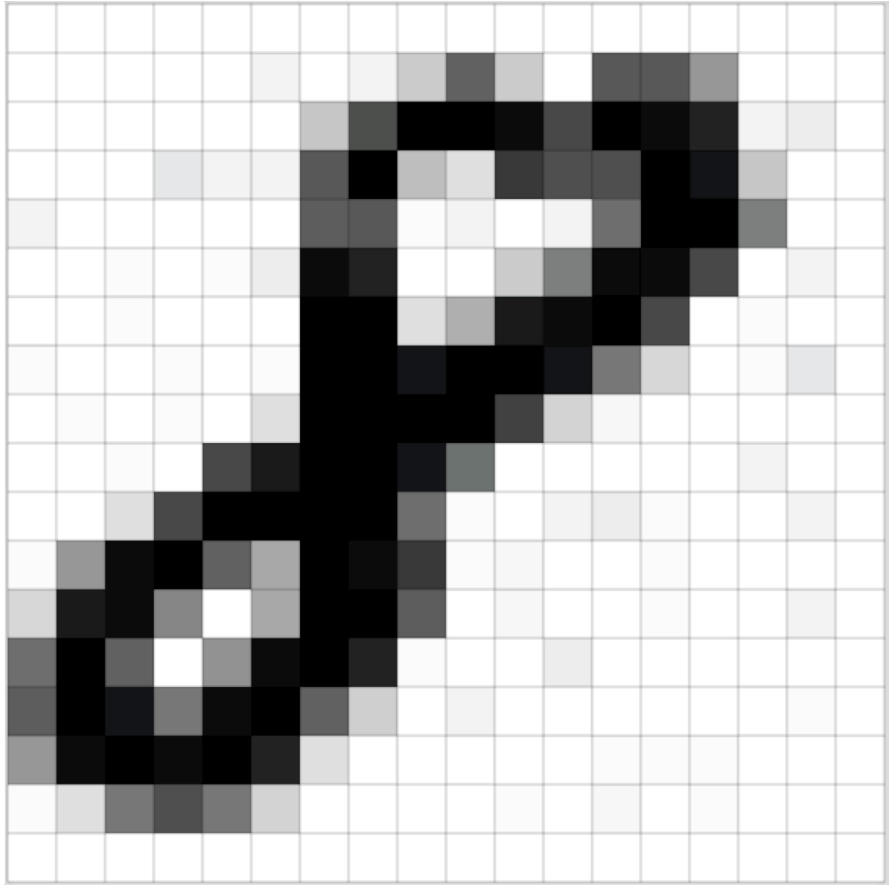
???



Representations

https://medium.com/@ageitgey/machine-learning-is-fun-part-3-deep-learning-and-convolutional-neural-networks-f40359318721

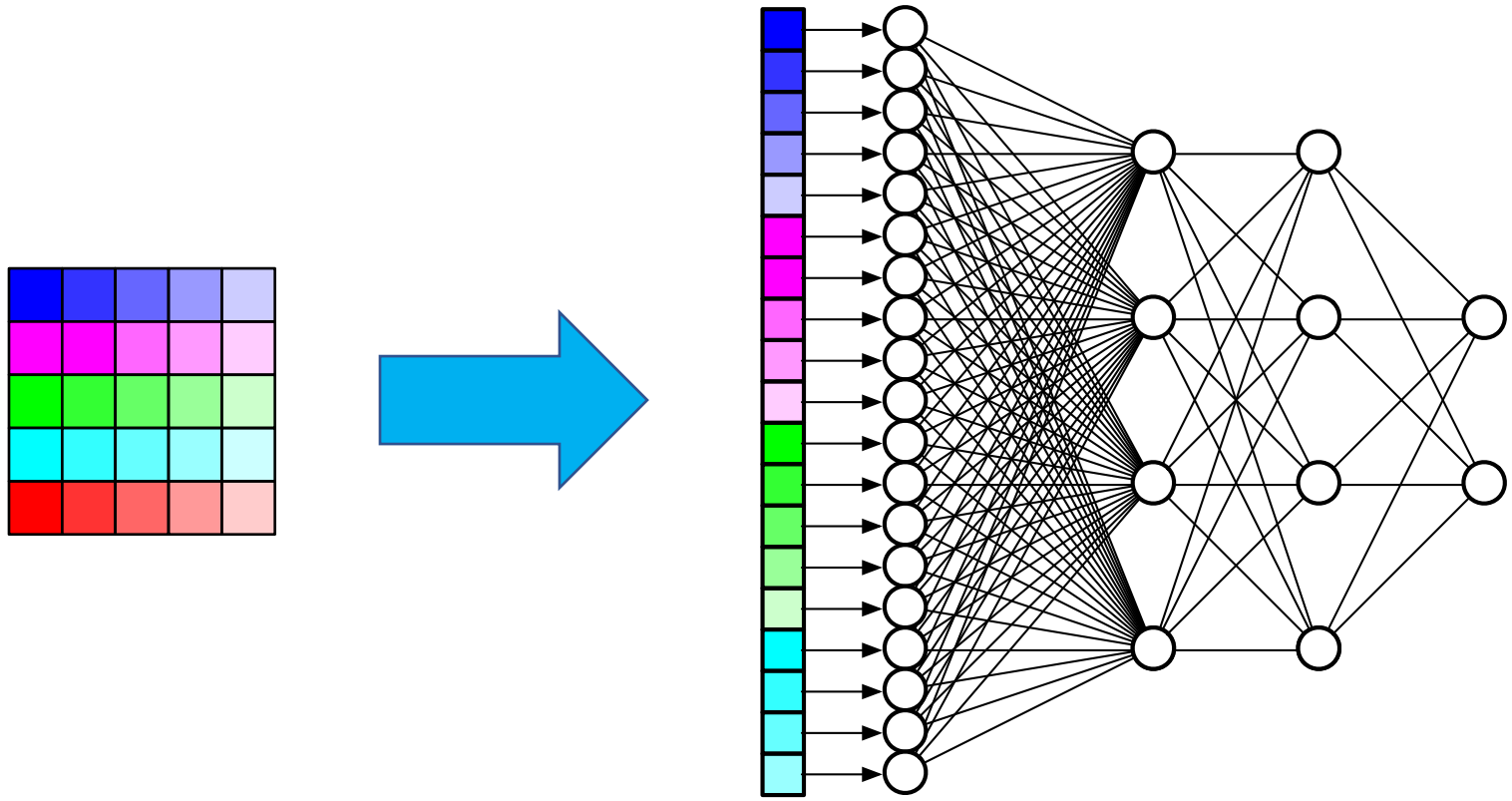
18 x 18



324 integers

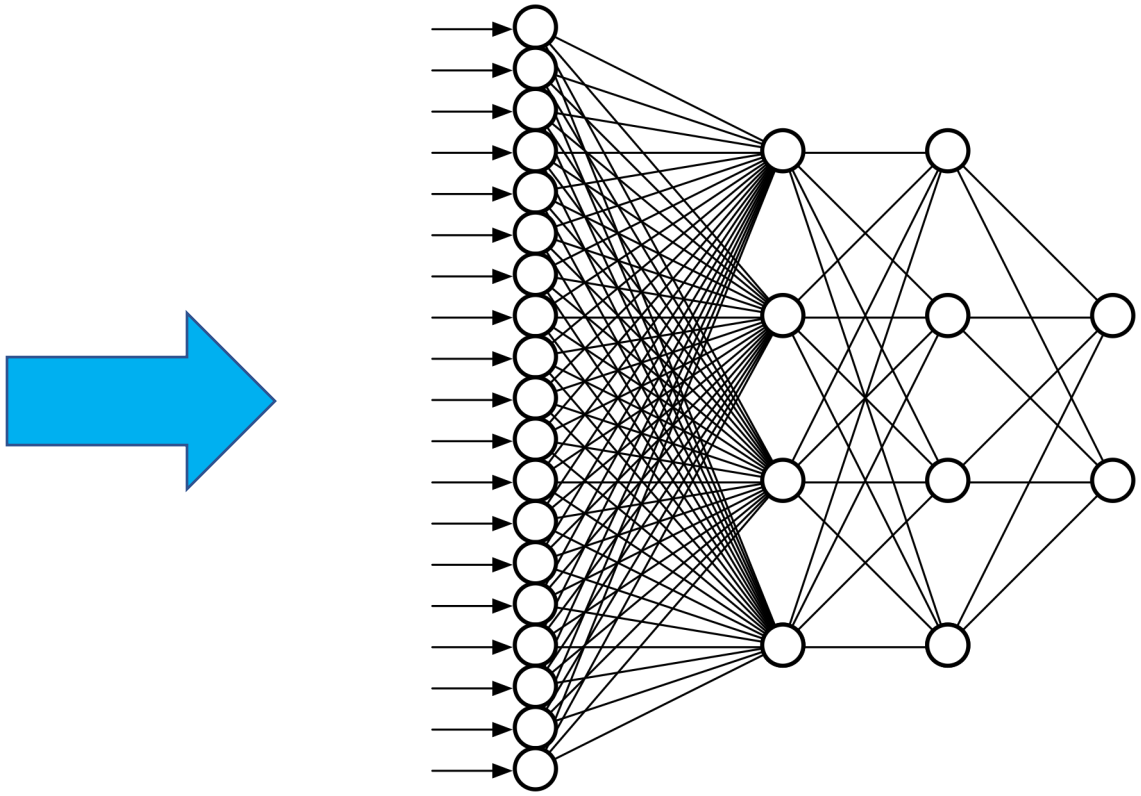
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	12	0	11	39	137	37	0	152	147	84	0	0	0
0	0	1	0	0	0	41	160	250	255	235	162	255	238	206	11	13	0
0	0	0	16	9	9	150	251	45	21	184	159	154	255	233	40	0	0
10	0	0	0	0	0	145	146	3	10	0	11	124	253	255	107	0	0
0	0	3	0	4	15	236	216	0	0	38	109	247	240	169	0	11	0
1	0	2	0	0	0	253	253	23	62	224	241	255	164	0	5	0	0
6	0	0	4	0	3	252	250	228	255	255	234	112	28	0	2	17	0
0	2	1	4	0	21	255	253	251	255	172	31	8	0	1	0	0	0
0	0	4	0	163	225	251	255	229	120	0	0	0	0	0	11	0	0
0	0	21	162	255	255	254	255	126	6	0	10	14	6	0	0	9	0
3	79	242	255	141	66	255	245	189	7	8	0	0	5	0	0	0	0
26	221	237	98	0	67	251	255	144	0	8	0	0	7	0	0	11	0
125	255	141	0	87	244	255	208	3	0	0	13	0	1	0	1	0	0
145	248	228	116	235	255	141	34	0	11	0	1	0	0	0	1	3	0
85	237	253	246	255	210	21	1	0	1	0	0	6	2	4	0	0	0
6	23	112	157	114	32	0	0	0	0	2	0	8	0	7	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Feeding to Neural Network



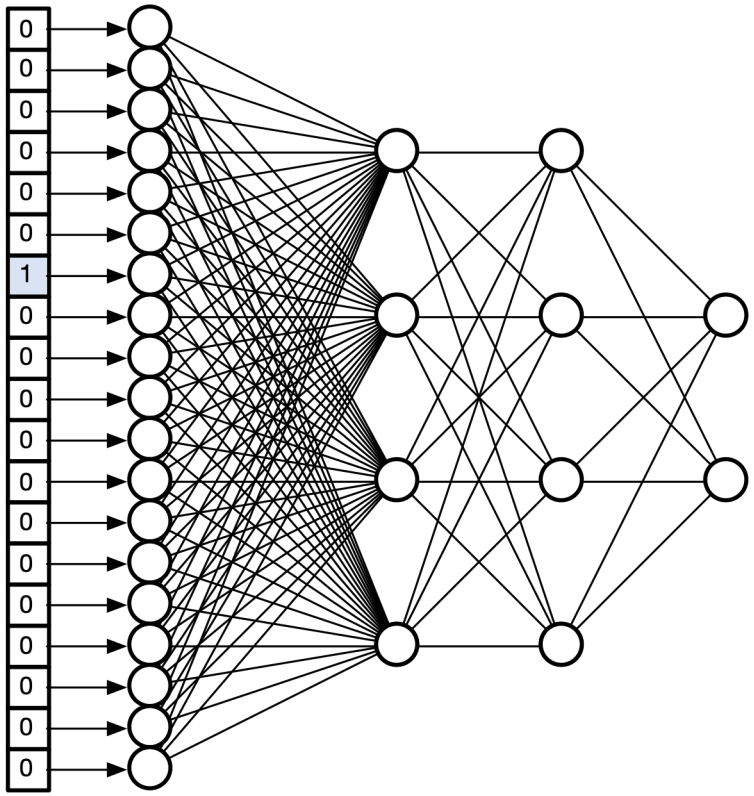
Feeding Text to Neural Network

The big brown bear is sitting in a chair



Feeding Text to Neural Network

a	0
	0
bear	0
big	0
brown	0
	0
chair	1
	0
in	0
is	0
	0
	0
sitting	0
	0
The	0



Feeding Text to Neural Network

a

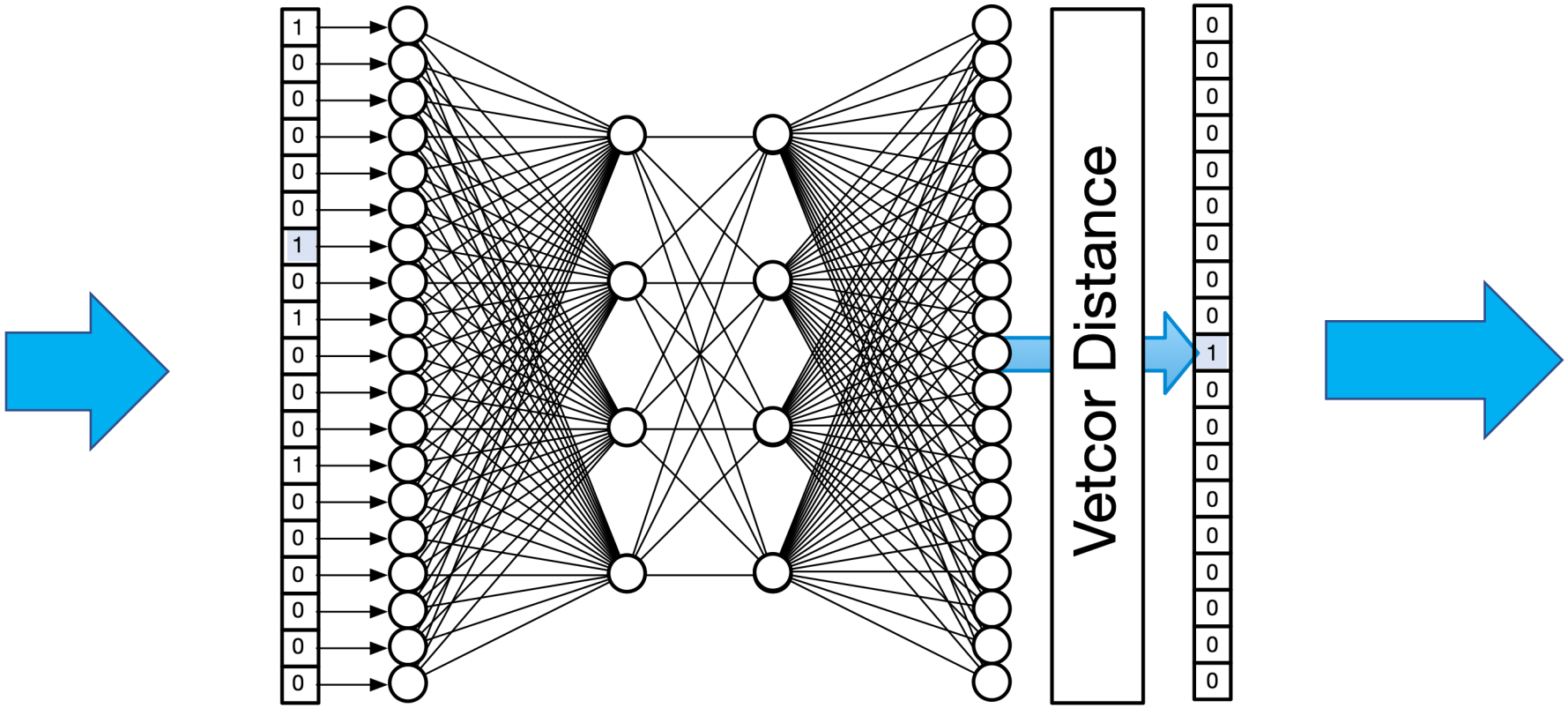
bear
big
brown

chair

in
is

sitting

The



a

bear
big
brown

chair

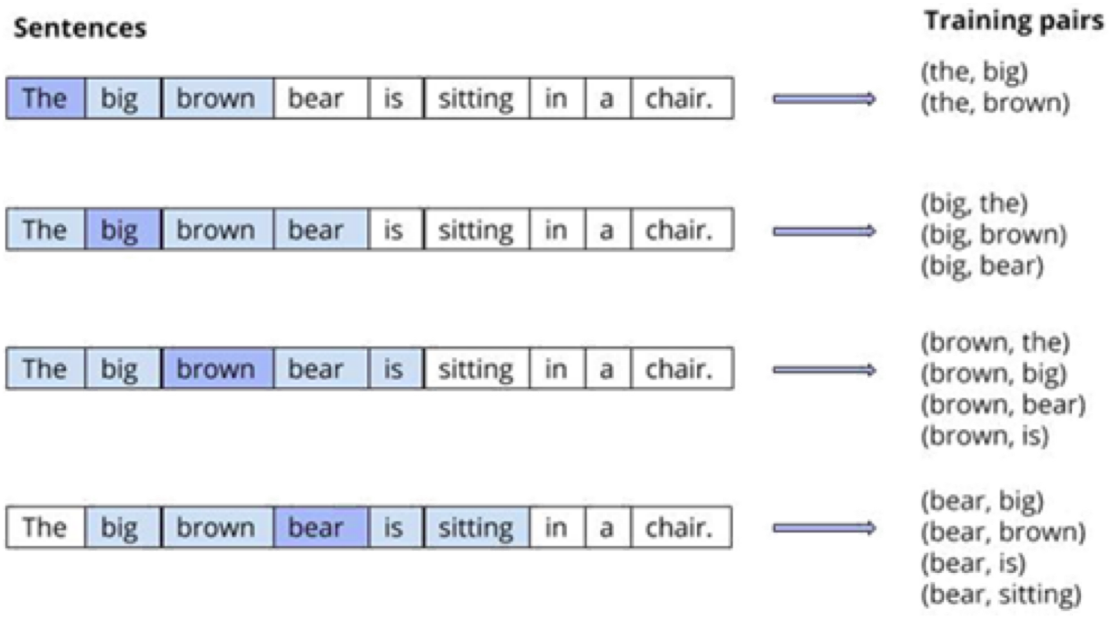
in
is

sitting

The

Neural Networks for Text

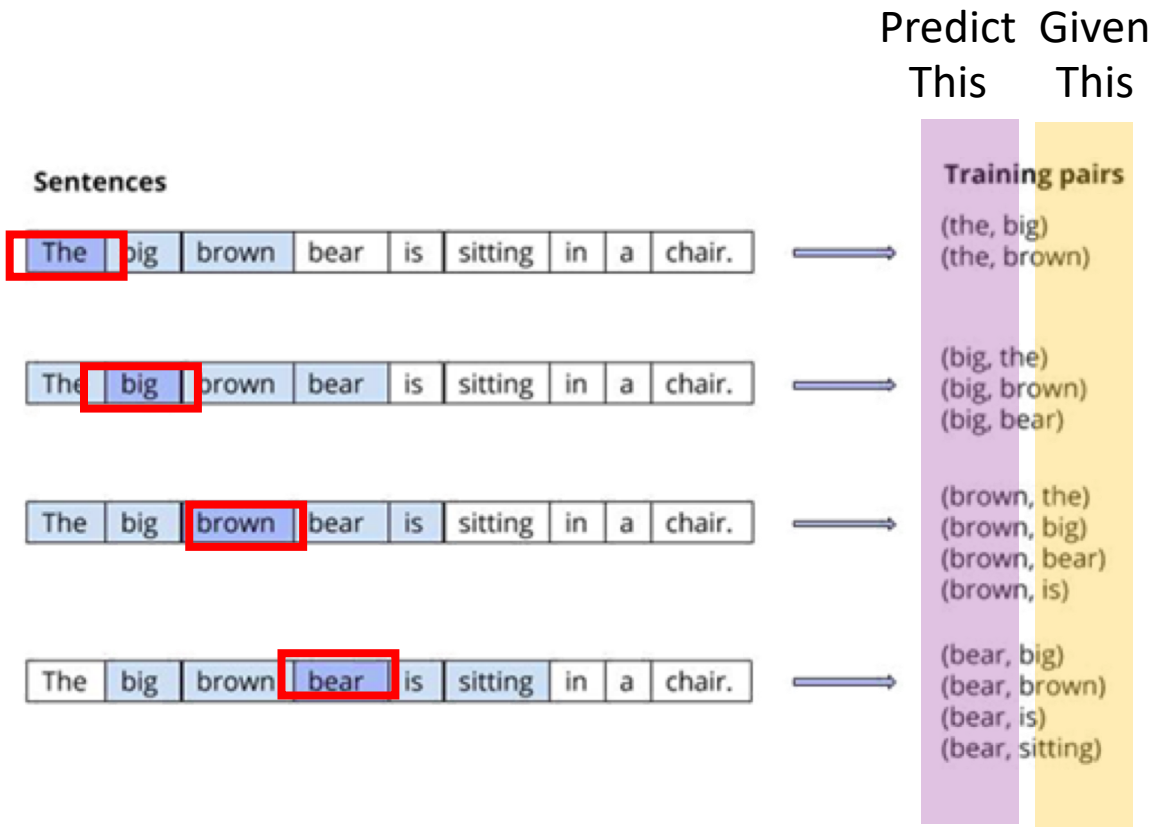
word2vec by Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean



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

Neural Networks for Text

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



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

Feeding Text to Neural Network

(Actually More Complicated)

a

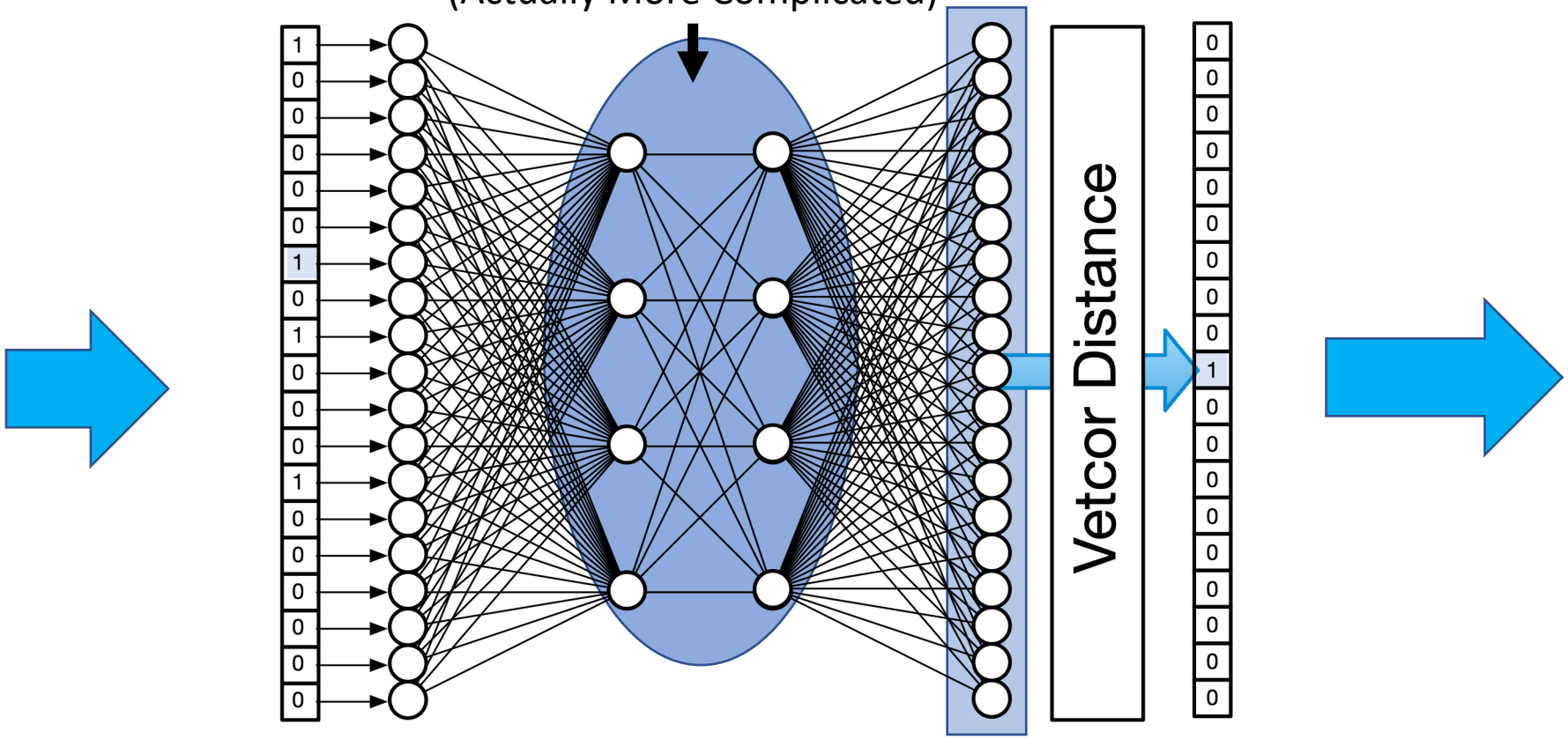
bear
big
brown

chair

in
is

sitting

The



Vector Representation
of Meaning

a

bear
big
brown

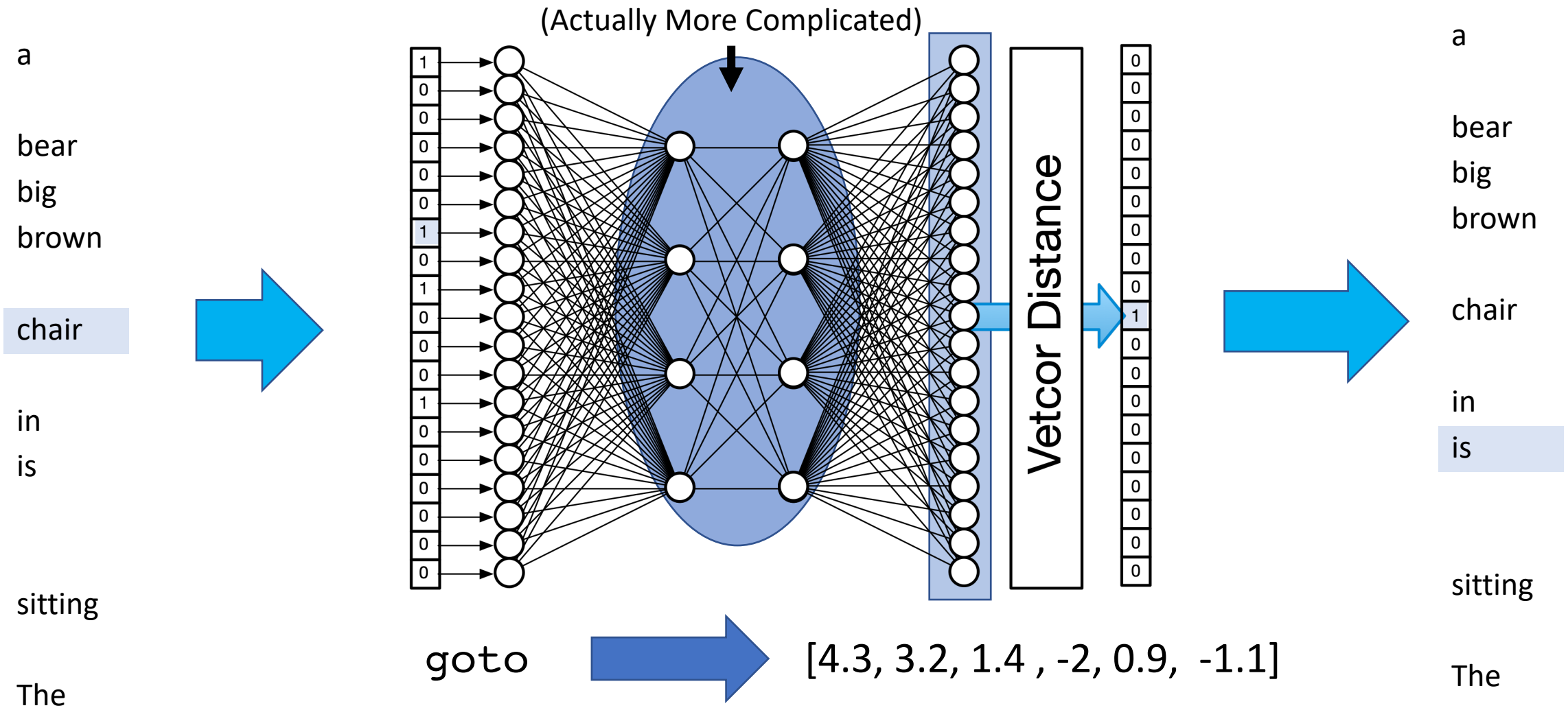
chair

in
is

sitting

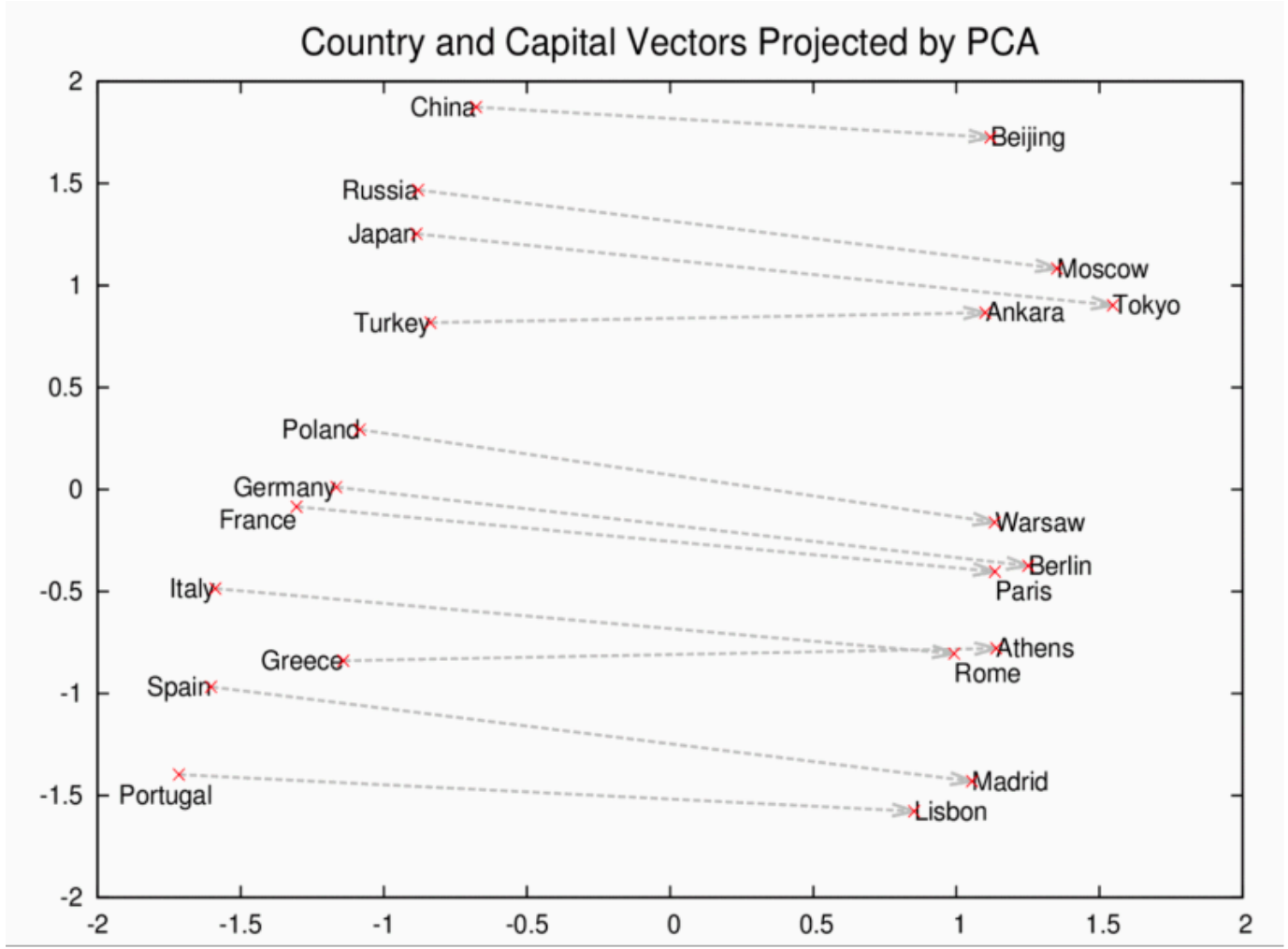
The

Feeding Text to Neural Network



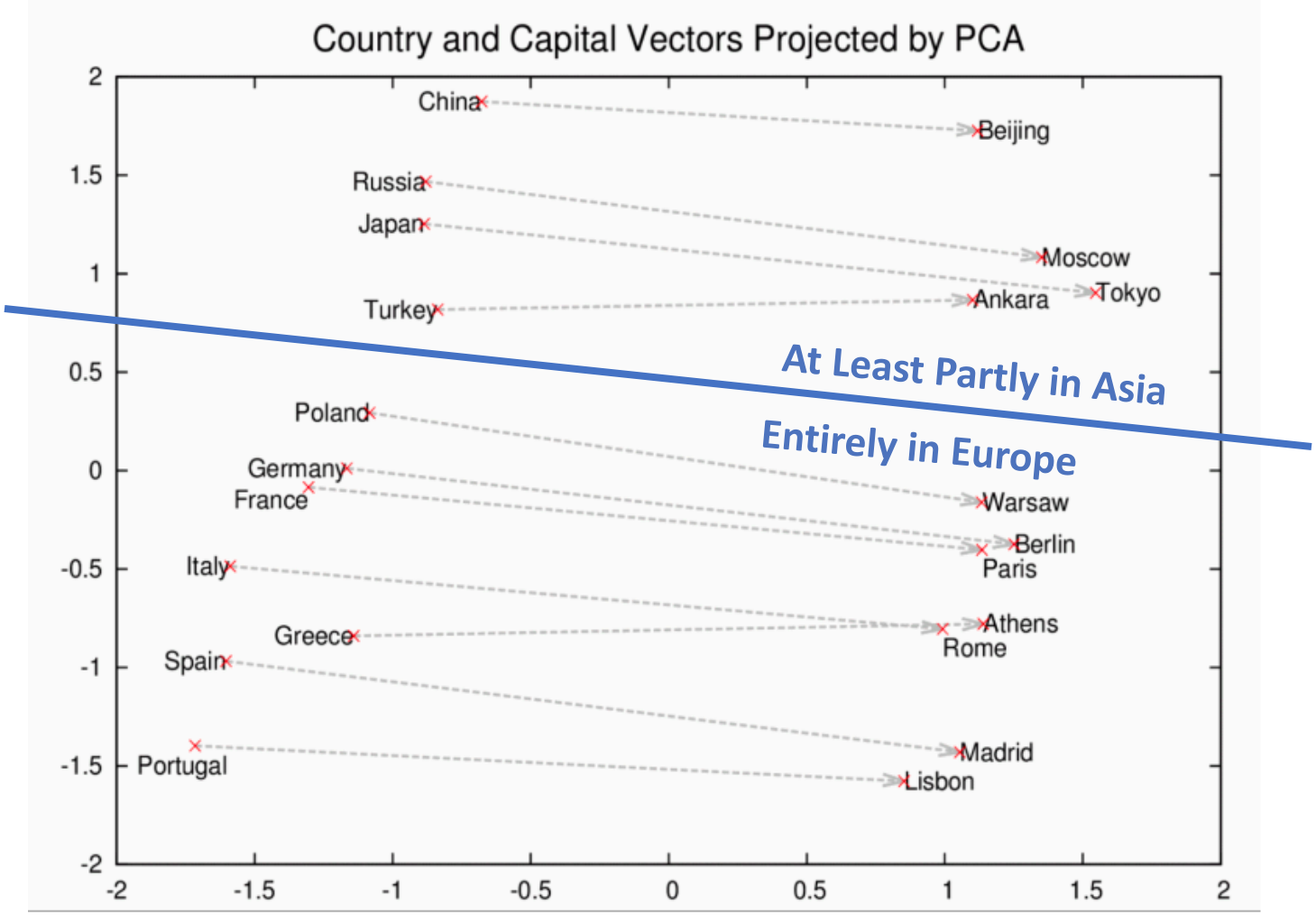
word2vec

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



word2vec

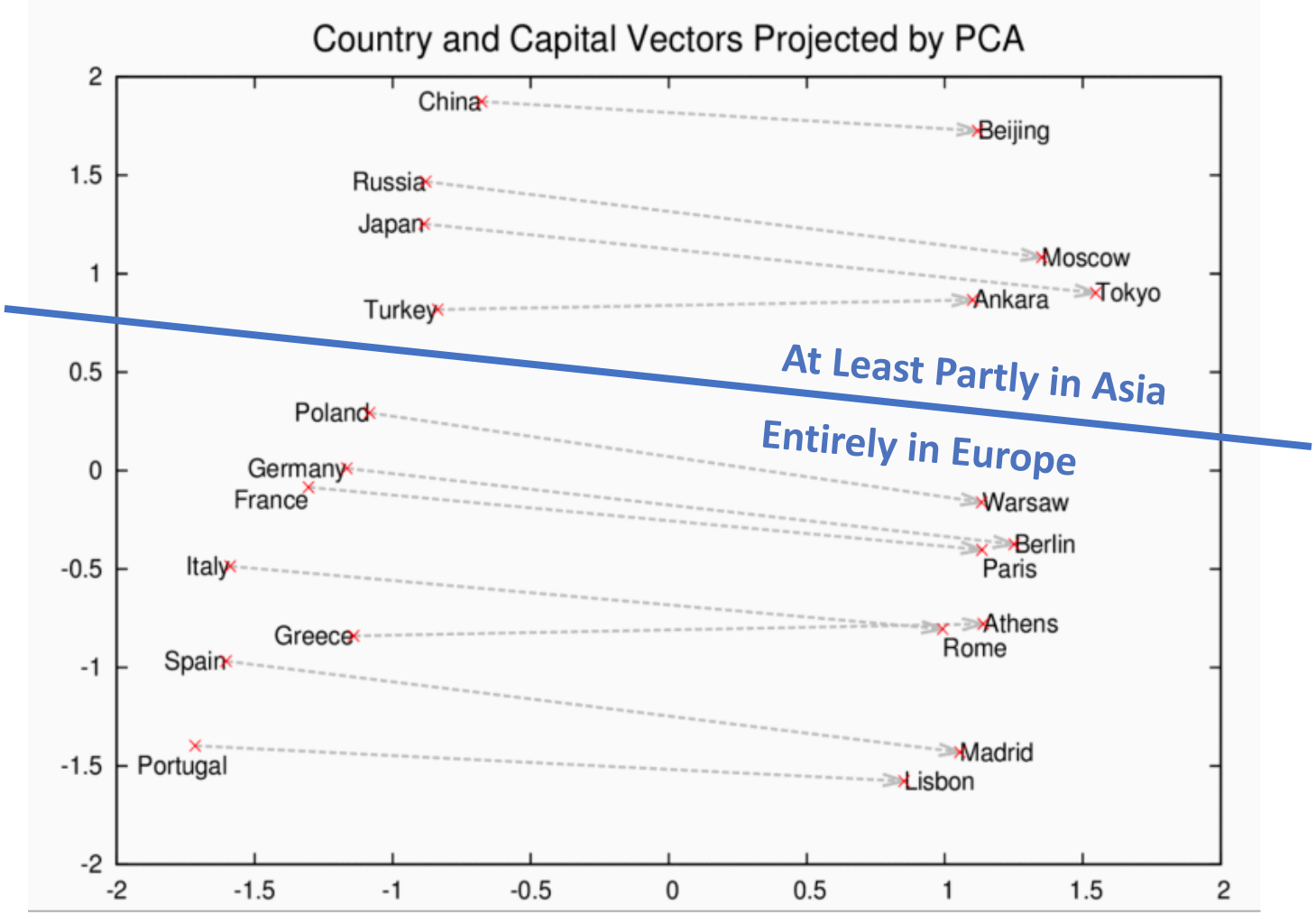
by Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean



muse dev

word2vec

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

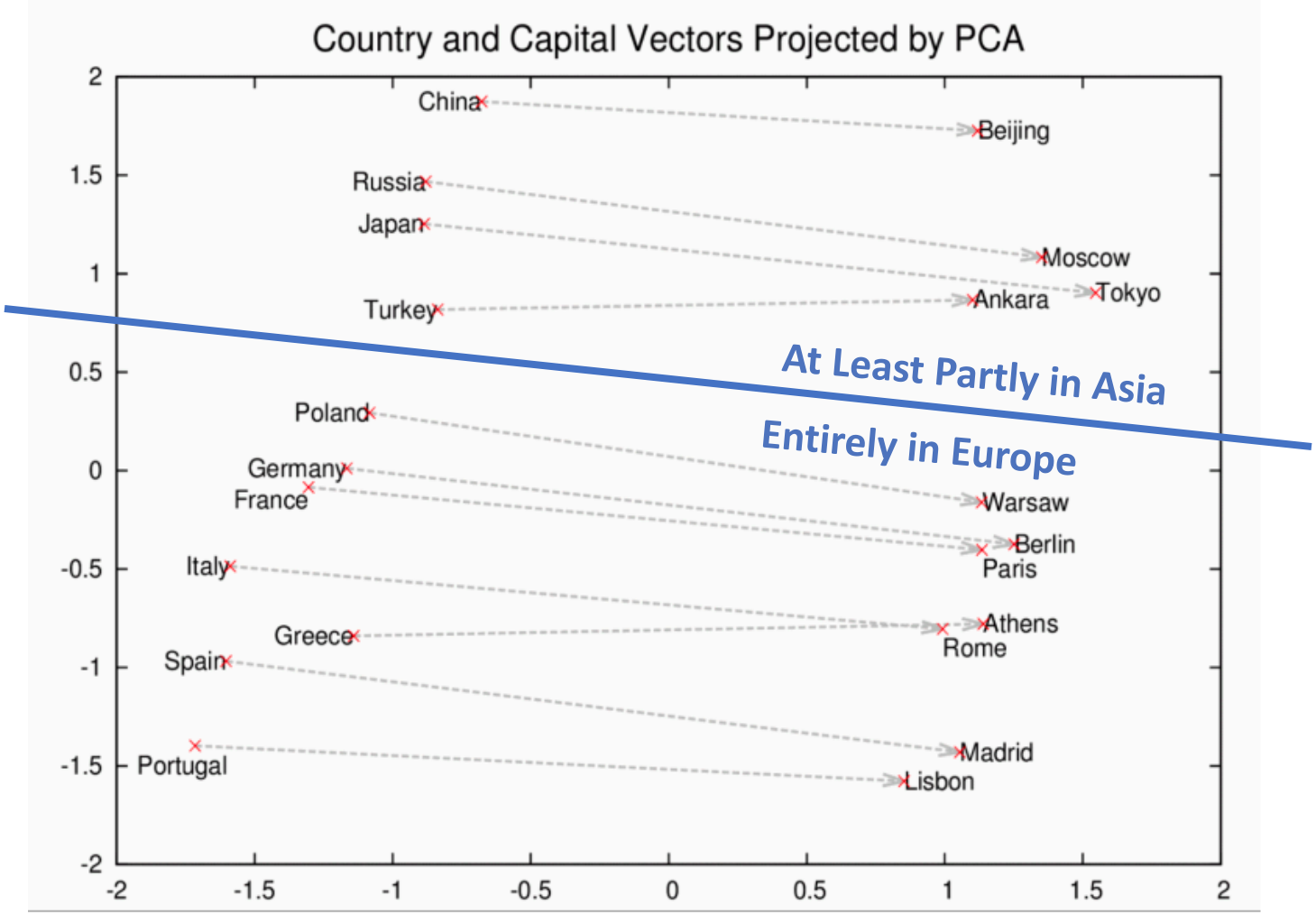


<u>Distance Captures Similarity</u>

Russia is closer to China than to Italy

word2vec

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



<u>Distance Captures Similarity</u>

Russia is closer to China than to Italy

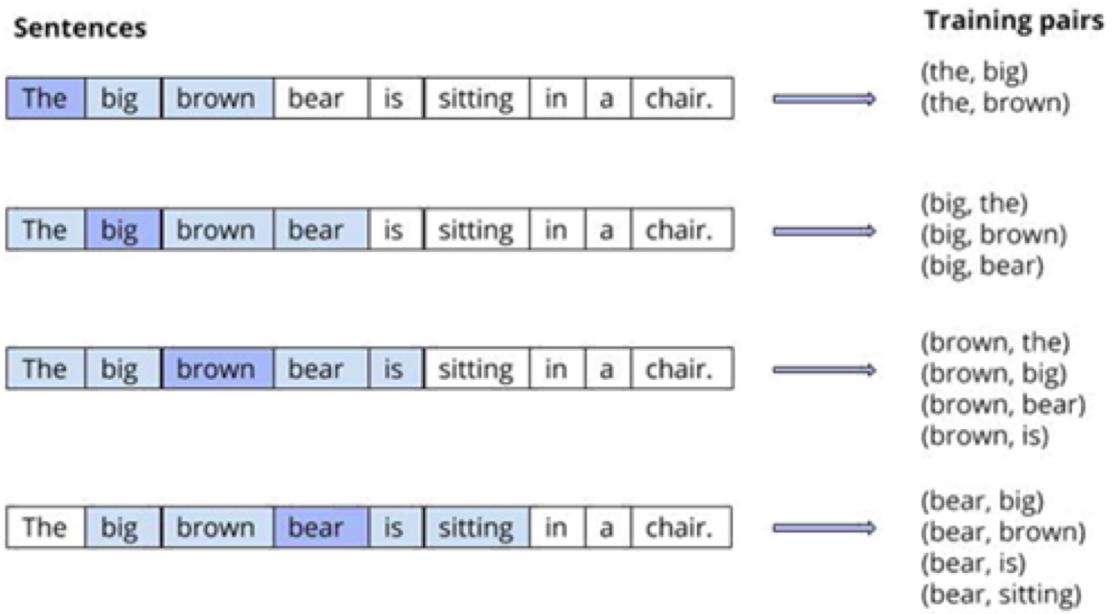
<u>Math Creates Analogies</u>

Russia – Moscow + Paris = France

(Russia:Moscow :: Paris:France)

Neural Networks for **Code**

code2vec: Learning Distributed Representations of Code



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

Neural Networks for **Code**

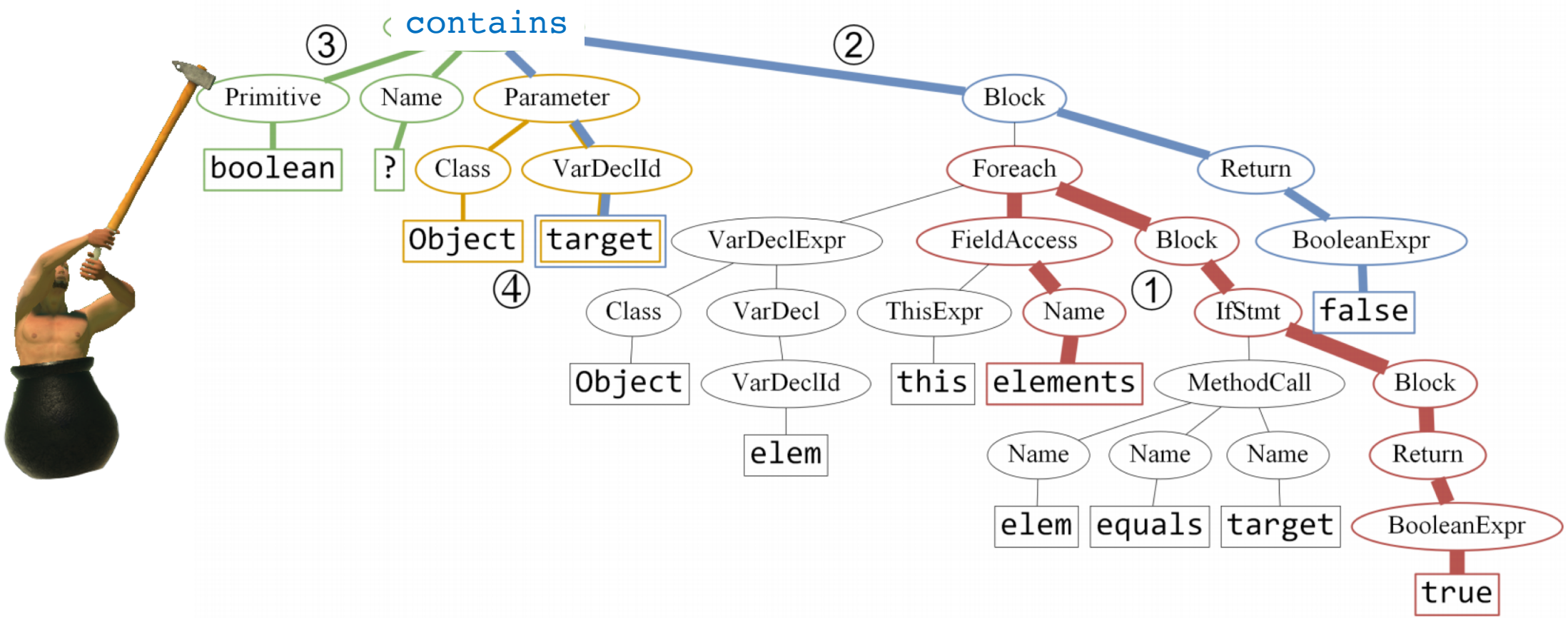
code2vec: Learning Distributed Representations of Code

```
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 Networks for **Code**

code2vec: Learning Distributed Representations of Code

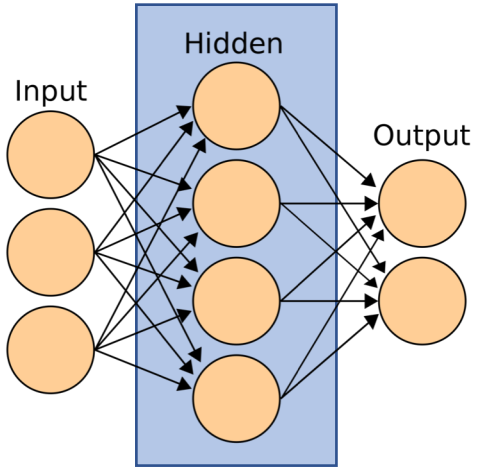


(elements, Name↑FieldAccess↑Foreach↓Block↓IfStmt↓Block↓Return↓BooleanExpr, true)

muse dev

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()) {
  …
}
```

→

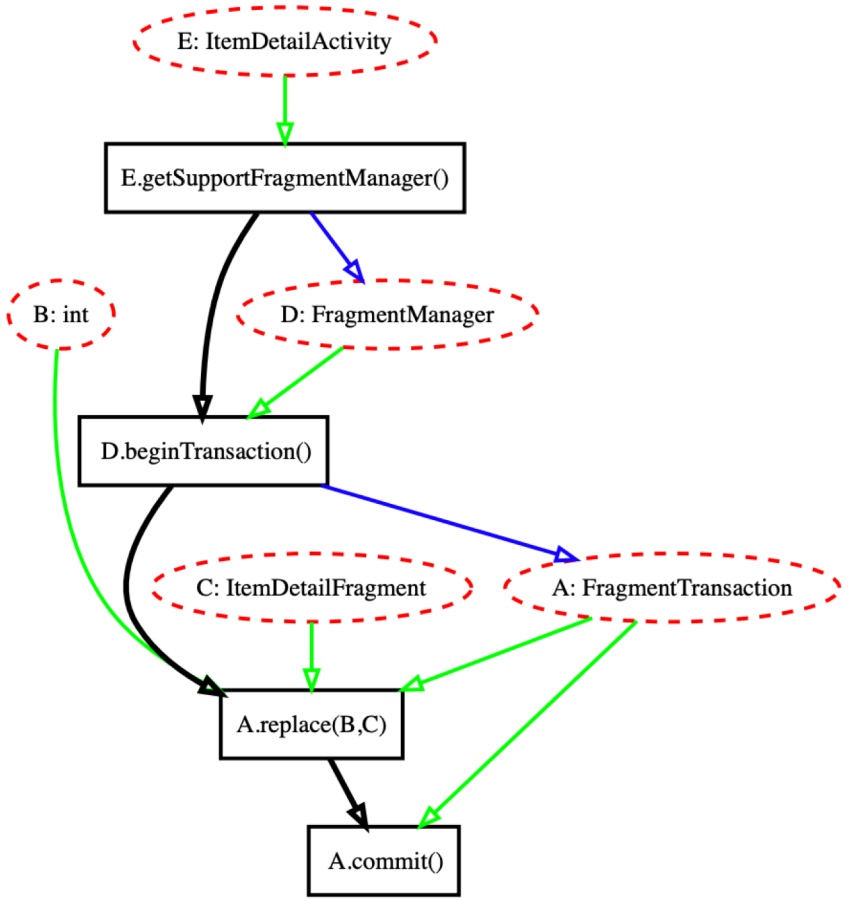
```
while
(iter.hasNext()) {
  …
}
```

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
- Similar to what new languages and frameworks enable, but with distinct capabilities.

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

Email:
stephen@muse.dev

Muse Dev
https://muse.dev

muse dev