Comparing Large Scale Language Models for Source Code

Rafael - Michael Karampatsis
CDT in Data Science
School of Informatics, University of Edinburgh

Motivation

Good Language models for code could potentially contribute to:

- Code completion
- Retrieval of code from natural language
- Retrieval of natural language from code snippets
- Identify buggy code
- Translate code from one programming language to another
- Automatic code generation
- Speech recognition of code

Data

Open source projects mined from **Github** in 4 programming languages.

1. Java
2. C
3. Python
4. Javascript

About 400 Million lines of code (LOC) for each language.
Split into three sets: Training set (150M loc), validation set and test set.
Some held out data for each language.

N-gram Language Model

What is a language model???

the cat sat on the mat

Chain Rule

P(the cat sat on the mat) = P(the) P(cat | the) P(sat | the cat) P(on | the cat sat) P(the | the cat sat on) P(mat | the cat sat on the)

$$P(w_1, w_2, \ldots, w_n) = \prod P(w_i \mid w_1, w_2, \ldots, w_{i-1})$$

N-gram Language Model

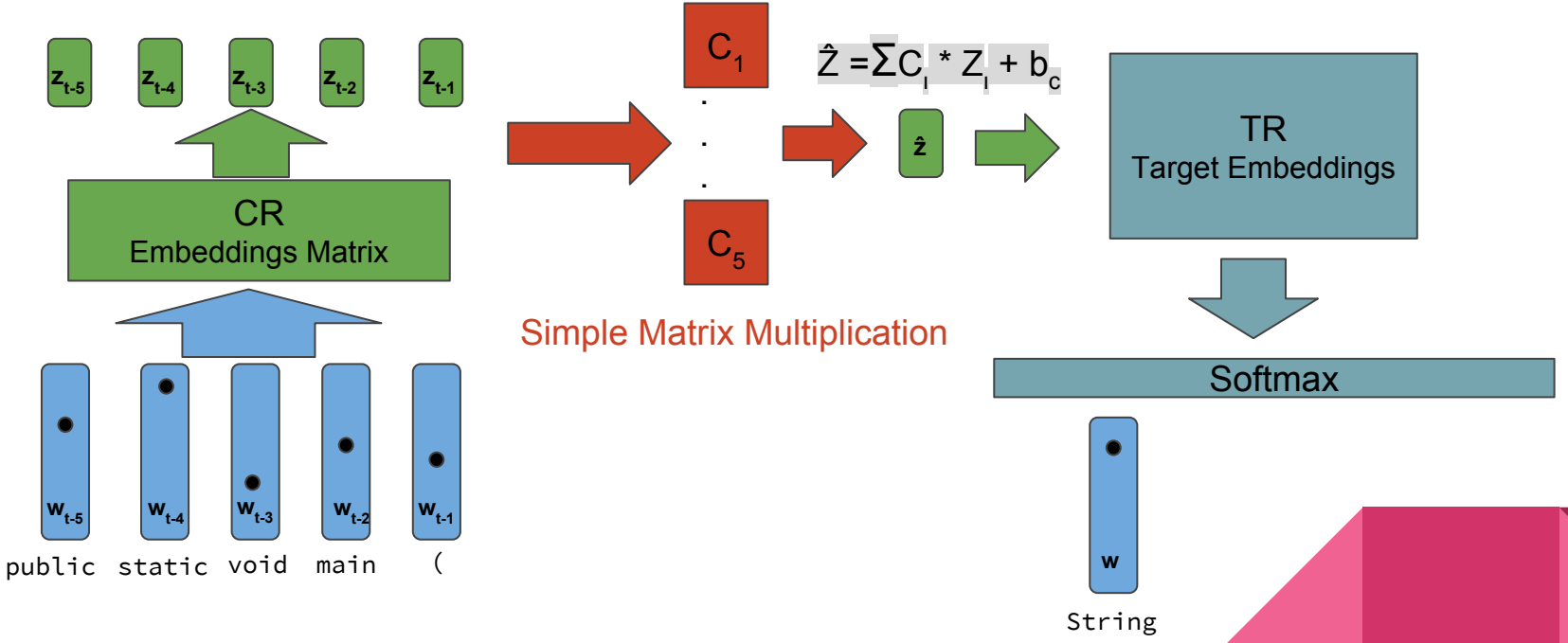
N-gram Language Model

Consider only a context (history) of ***N-1*** words.

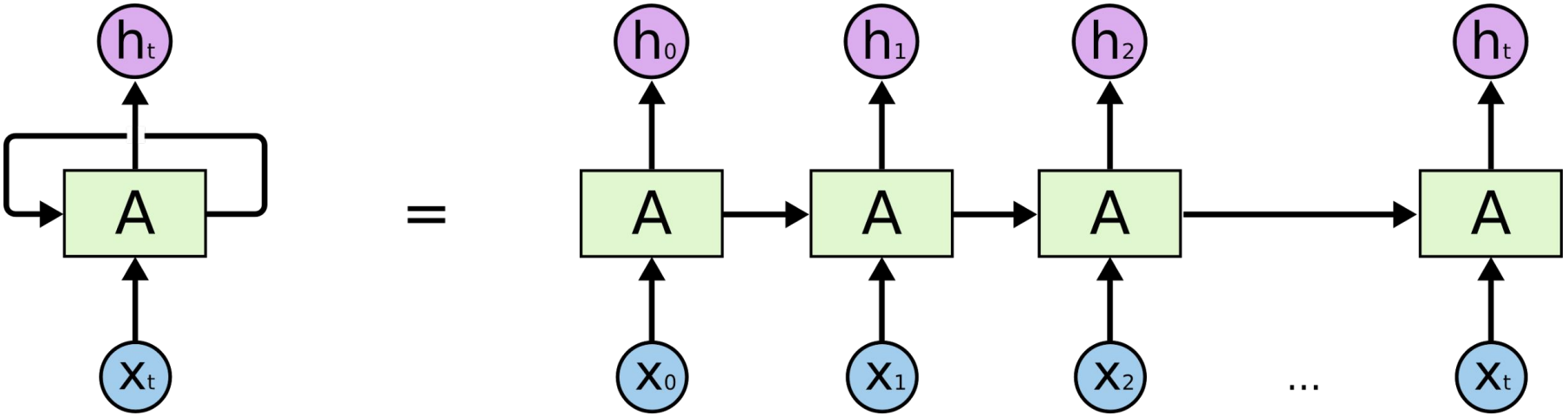
Example: 3-gram Language Model

P(the cat sat on the mat) = P(the) P(cat | the) P(sat | the cat) P(on | cat sat) P(the | sat on) P (mat | on the)

Log-Bilinear Language Model (LBL)

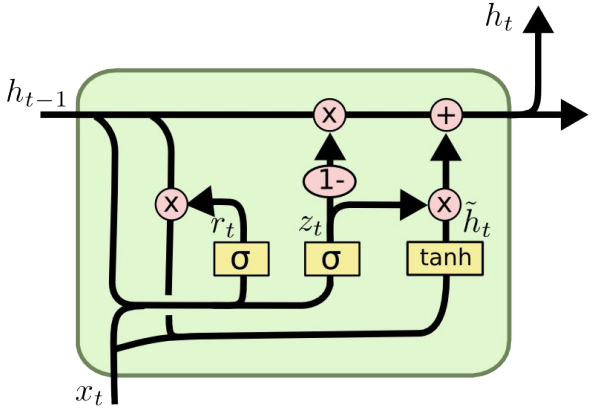


Gated Recurrent Unit Neural Network (GRU)



- Great at processing sequences
- Can handle long-term dependencies

Gated Recurrent Unit Neural Network (GRU)



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

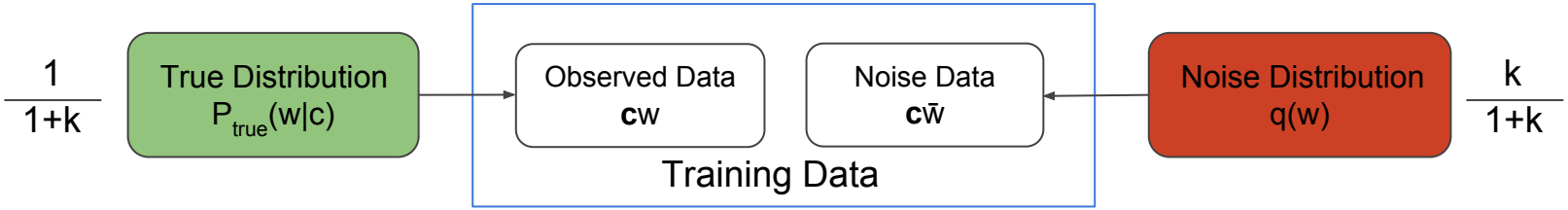
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

- Introduced by *Cho et al.* in 2014
- Simpler variation of LSTMs that works equally well (sometimes better).
- Combines the forget and input gates into a single “update gate.”
- Merges the cell state and hidden state.

Noise Contrastive Estimation (NCE)




```
public static void
```

```
public static private
public static }
          …
public static return
```

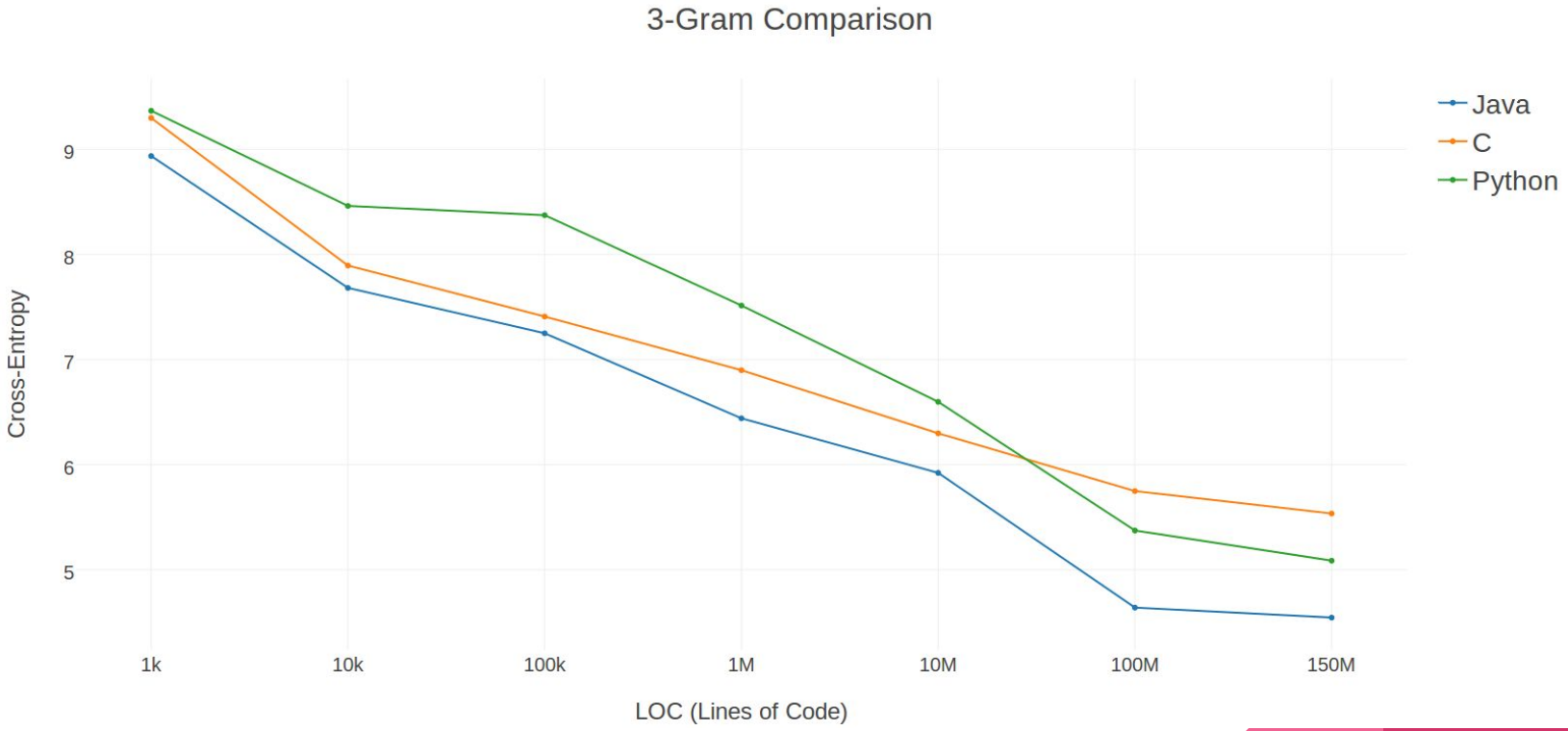
$$\frac{P(w \text{ is true} \mid \mathbf{c}w)}{P(w \mid \mathbf{c}) + kq(w)}$$

$$\frac{P(\bar{w}_j \text{ is noise} \mid \mathbf{c}\bar{w}_j)}{P(\bar{w}_j \mid \mathbf{c}) + kq(w)}$$

Experimental Setup

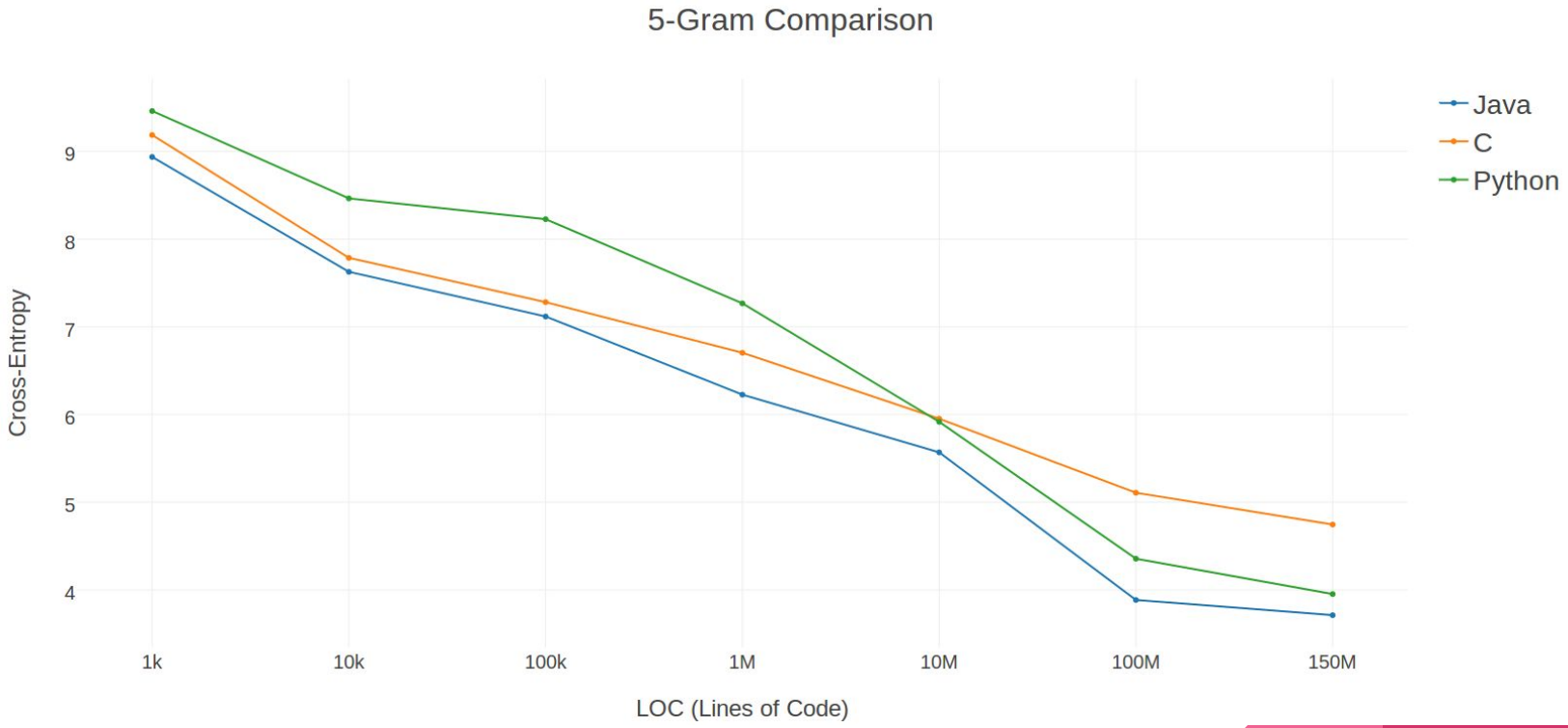
1. Data is tokenized and then split into training, validation and test.
2. Training data of variable sizes. 1k, 10k, 100k, 1M, 10M, 100M, 150M loc.
3. The vocabulary of the 150M loc is extracted and all rare tokens (frequency≤5) are mapped into a special **unk** token. Helps the model to learn the distribution of tokens unseen in the training data and reduces vocabulary size. Vocabulary ≃2M.
4. Hyper parameters are tuned using the development set.

Results



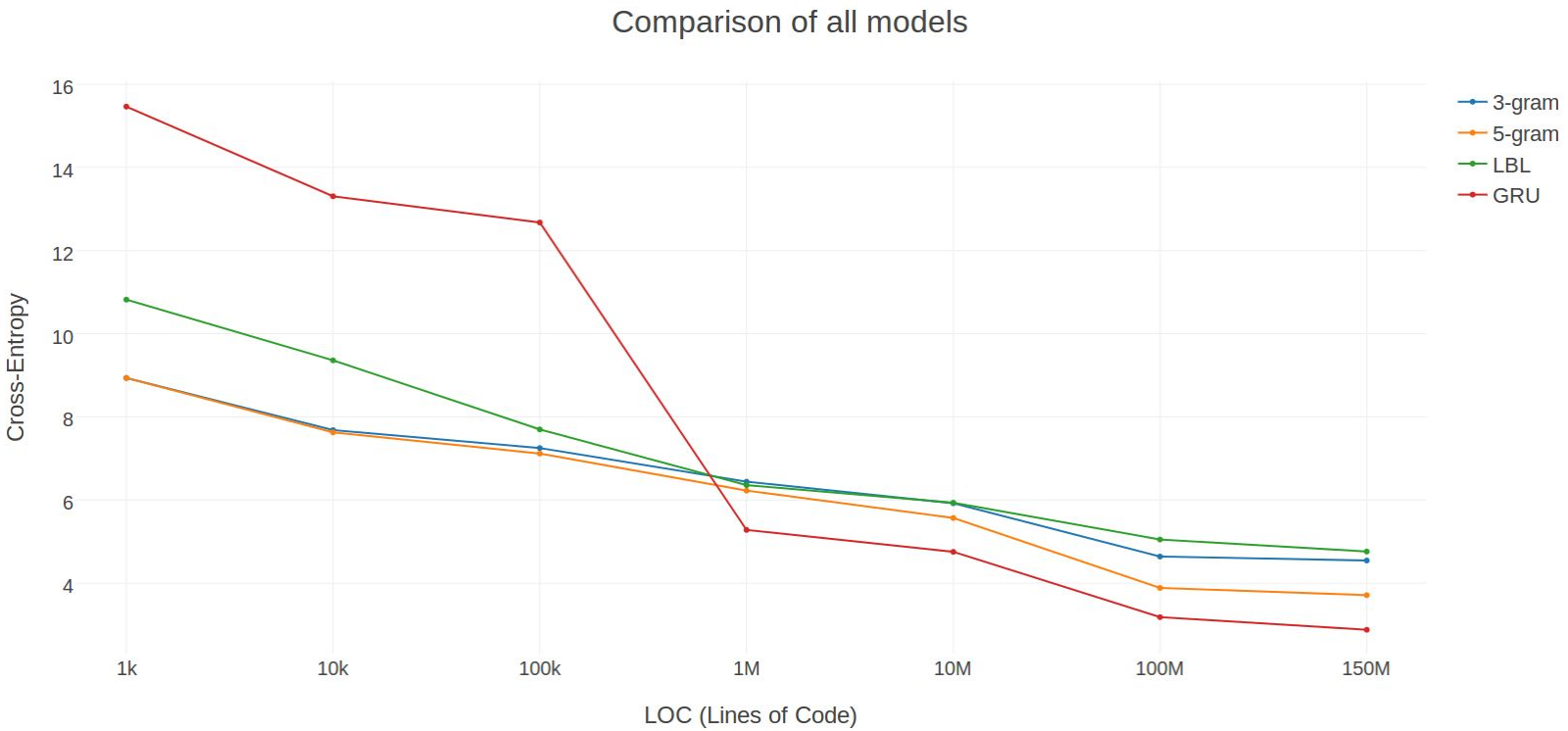
DATASCIENCE
EPSRC CENTRE FOR DOCTORAL TRAINING

Results



DATASCIENCE
EPSRC CENTRE FOR DOCTORAL TRAINING

Results


Comparison of all models
Cross-Entropy
16
14
12
10
8
6
4
1k
10k
100k
1M
10M
100M
150M
LOC (Lines of Code)
3-gram
5-gram
LBL
GRU

Code Generation Example

Kneser-Ney 3-gram 150M LOC

package org . osid . repository . impl . setCaptionWidth (30 , new <unk> ( ) { this . connection . getEndPoint ( ) throws Exception { CpuRawData defaultDomainId) { <unk> (entry . getValue ( )) ;

import java.util.Arrays ; import com.hp.hpl.jena.util.iterator.ExtendedIterator ; public class <unk> extends java . lang . String outServerName ! = null : outTuple loadNext (27) " : ") ; Node v $5 , v) ; final int __PERIOD_ISSET_ID = now ( )) ;

public static String <unk> ( ) throws Exception { createTable (TABLE) ; this . isShowing ( ) + " sh " , ctx , procedureName) proxy . host ( ) , new CardType [ ] A (I1 . contains (. , 2) ; unicodePstmt . close ( ) { comments . getList ( ) ; selenium . clickAt ("] ")) { qs . getQuestVarById (0 , 1 , stats . VersionedUpdateElementType ibutton = idmFieldName ;

Code Generation Example

1 Layer GRU 150M LOC

```
package rgftc overlayEnvelope TripleCollectionInfo overdueTimes ID_DOUBLE_MARK alternate clReview
SINGLE_QUOTED_STRING__TEXT searchblogsentrycp = fromEllipsoid ; }
public DI _small_ ( ) { return null; }
public DVar _a_ ( ) { ( Keystore ) fLastModified = 1 ; } }                                                    public
static Object getInstance ( Object obj , so ( org . eclipse . gmf . thrift . gmf . getLogEnterTimestamp . canceller . Value . length ) ;

protected boolean initialized = true ; protected Configuration ( ) { super ( ) ; }
protected Class getSelectorType ( ) { return m_contains ; }

public DSmall add ( char [ ] ) { int eventType = WILDCARD_NAME . PUTFIELD ( ) ; }

public boolean canProvideCapability ( storedDetailChunks . . equals ( ( ) ) { return true ; }
public int searchQueryHint ( ) { int [ ] ret = null ; }
public int getID ( ) { return getConditional ( ) ; } }
```

DATASCIENCE
EPSRC CENTRE FOR DOCTORAL TRAINING

Questions?


```
def questions_handler(questions):
    successful_talk = True
    for question in questions:
        answer = answer(question)
        print answer
        if not audience_likes(answer):
            successful_talk = False
    return successful_talk
```

DATASCIENCE
EPSRC CENTRE FOR DOCTORAL TRAINING