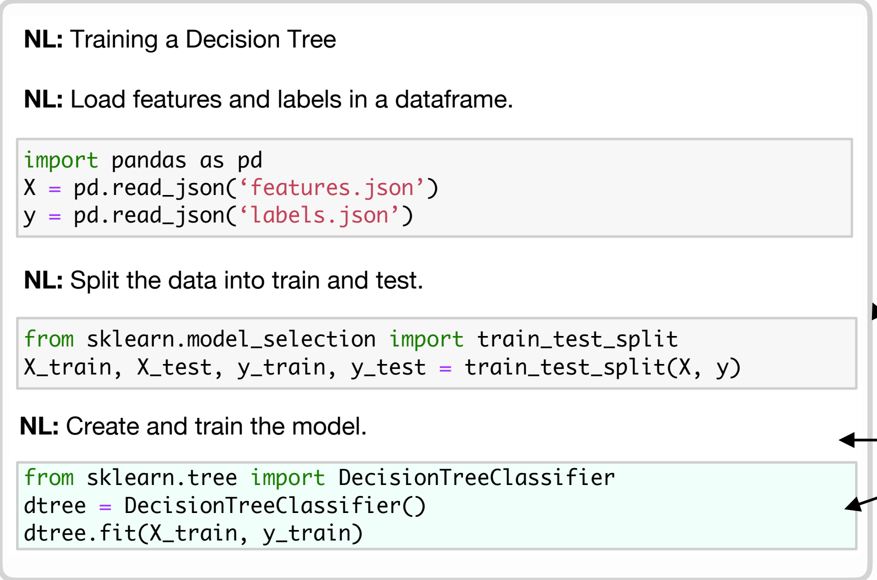
JuICe: A Large Scale Distantly Supervised Dataset for Open Domain Context-based Code Generation

Rajas Agashe, Srini Iyer, Luke Zettlemoyer
Paul G. Allen School of Computer Science & Engineering, Univ. of Washington



NL: Training a Decision Tree

NL: Load features and labels in a dataframe.

```
import pandas as pd
X = pd.read_json('features.json')
y = pd.read_json('labels.json')
```

NL: Split the data into train and test.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y)
```

NL: Create and train the model.

```
from sklearn.tree import DecisionTreeClassifier
dtree = DecisionTreeClassifier()
dtree.fit(X_train, y_train)
```

Task & Motivation

- Developers write complex programs piece by piece, thus we aim to learn models which allow developers to write pieces by themselves and use an interactive NL -> Code system to autogenerate other pieces.
- We release JuICe with the goal of enhancing interactive capabilities of models.

Input: All NL and Code cells above
Output: Target Code cell

Dataset

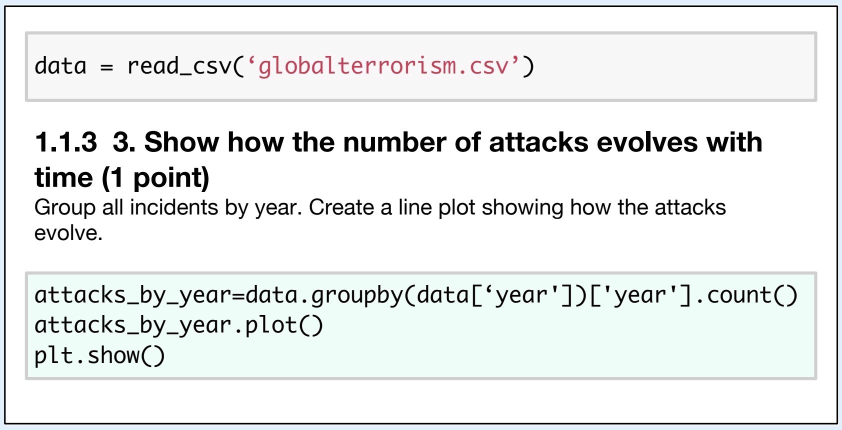
TRAIN	1,518,049 (Context, NL, Code)
DEV	1,744
TEST	1,981
Avg # Tokens: **38**	% contextual **61.9%**

Code: https://github.com/rajasagashe/JuICe

2. Human Curated

Dev and test sets are built from online programming assignments.

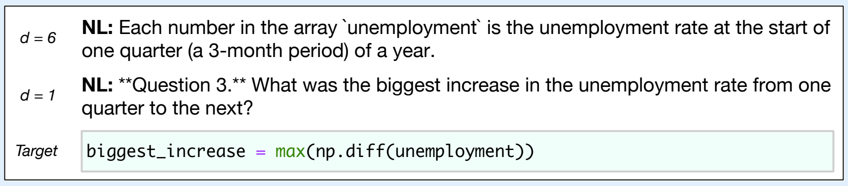
Examples:



```
data = read_csv('globalterrorism.csv')
```

1.1.3 3. Show how the number of attacks evolves with time (1 point)
Group all incidents by year. Create a line plot showing how the attacks evolve.

```
attacks_by_year=data.groupby(data['year'])['year'].count()
attacks_by_year.plot()
plt.show()
```

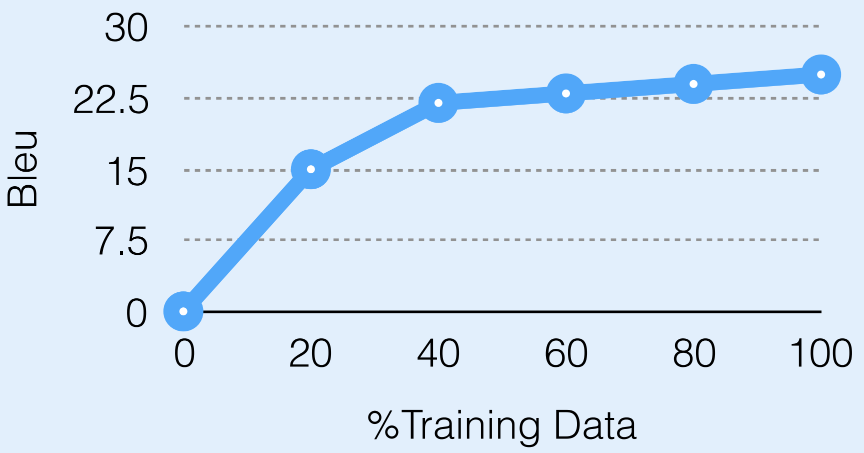


d = 6	**NL:** Each number in the array `unemployment` is the unemployment rate at the start of one quarter (a 3-month period) of a year.
d = 1	**NL:** **Question 3.** What was the biggest increase in the unemployment rate from one quarter to the next?
Target	`biggest_increase = max(np.diff(unemployment))`

Dataset Properties

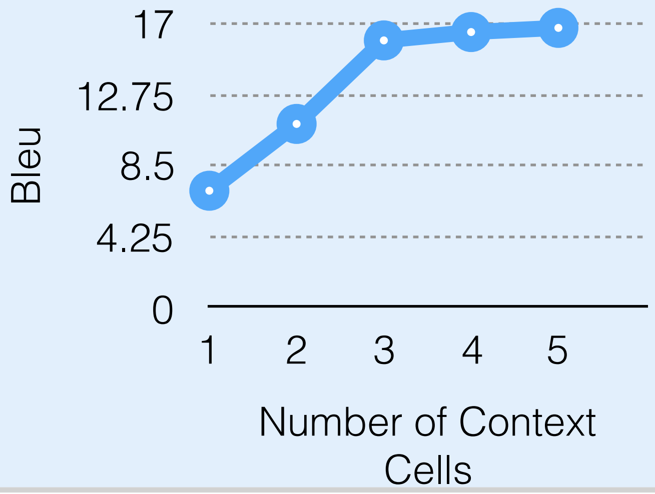
1. Large Scale

- We collect all public Jupyter notebooks on Github resulting in a large training set.
- More distantly supervised training data improves the model.



3. Context Based

- Models need to reason over multiple cells to generate target code. Adding more context cells improves the model.


Bleu
Number of Context Cells