

Build a Customized NLP Service

Interim Presentation

Friday, July 9, 2021

Presented by: Yongchao Zhou

Motivation

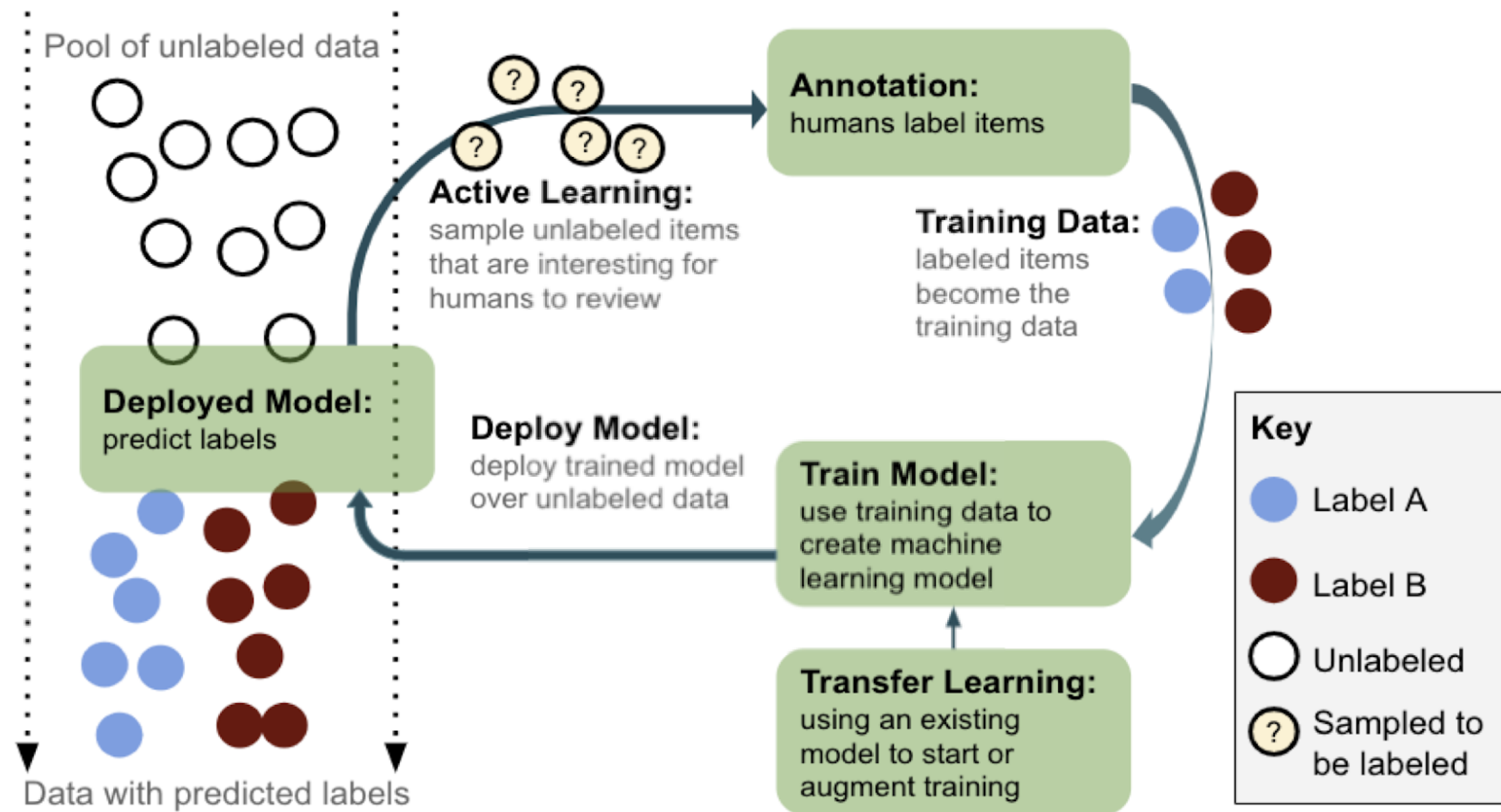


Figure 1: Machine Learning Life Cycle

- ML Application Life Cycle
 - Data Collection
 - Data Annotation
 - Model Training & Evaluation
 - Model Deployment
- ML Tools
 - Active Learning
 - Transfer Learning

Motivation

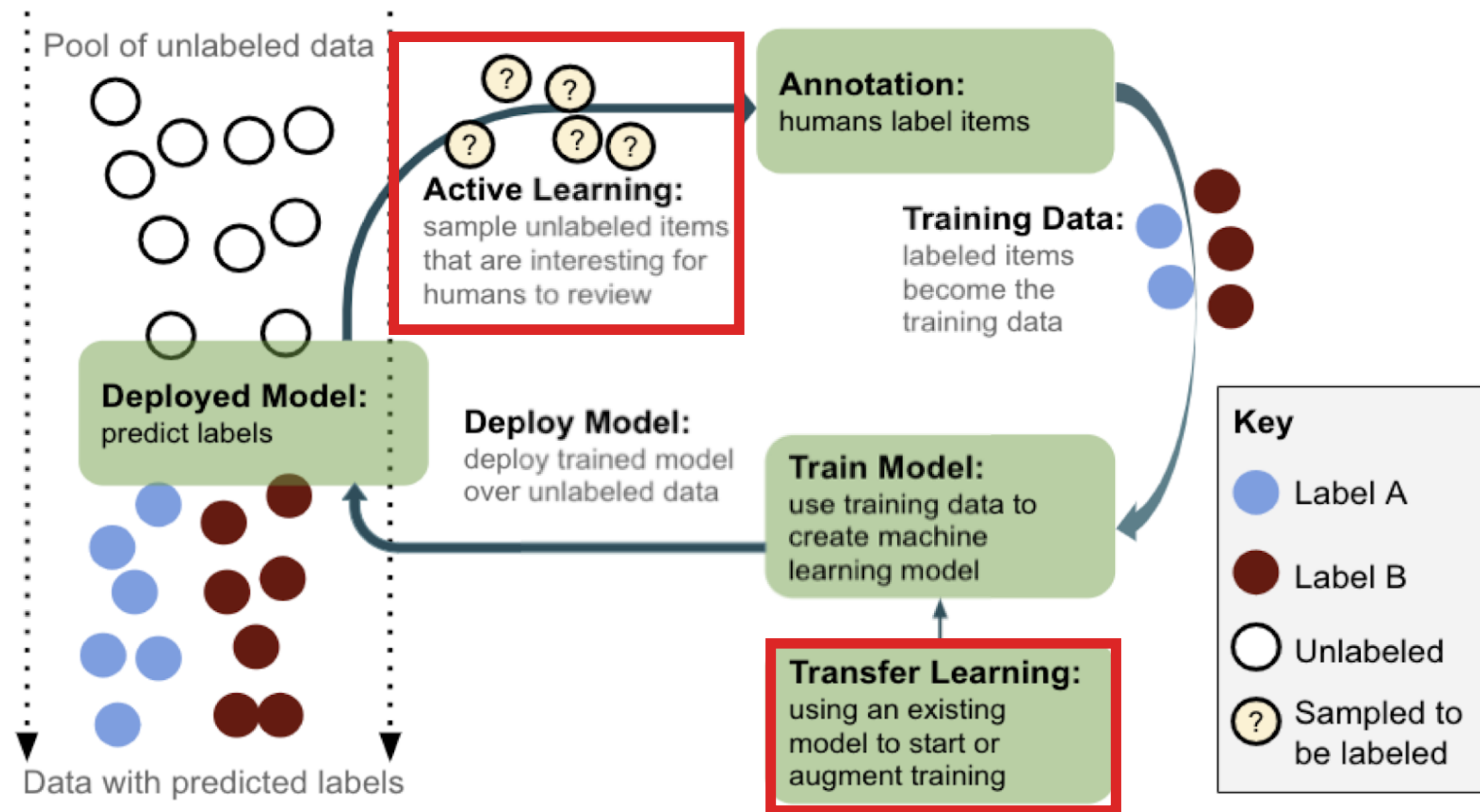


Figure 1: Machine Learning Life Cycle

- ML Application Life Cycle
 - Data Collection
 - Data Annotation
 - Model Training & Evaluation
 - Model Deployment
- ML Tools
 - Active Learning
 - Transfer Learning

NER service

- Inference API
- Graphical User Interface

NER service

- Inference API

```
>>> import NerModel
>>> model = NerModel(args.model_dir)
>>> model(args.text)

{'text': 'Geoffrey Everest Hinton (born 6 December 1947)',
 'ents': [{'token': 'Geoffrey Everest Hinton', 'start': 0, 'end': 20, 'label': 'PERSON'},
           {'token': '6', 'start': 30, 'end': 31, 'label': 'DATE'},
           {'token': 'December 1947', 'start': 32, 'end': 45, 'label': 'DATE'},
           {'token': 'British', 'start': 52, 'end': 59, 'label': 'NATIONALITY'}]}
```

- Graphical User Interface

NER service

- Inference API

```
>>> import NerModel
>>> model = NerModel(args.model_dir)
>>> model(args.text)

{'text': 'Geoffrey Everest Hinton (born 6 December 1947)',
 'ents': [{'token': 'Geoffrey Everest Hinton', 'start': 0, 'end': 20, 'label': 'PERSON'},
          {'token': '6', 'start': 30, 'end': 31, 'label': 'CARDINAL'},
          {'token': 'December 1947', 'start': 32, 'end': 45, 'label': 'DATE'},
          {'token': 'British', 'start': 52, 'end': 59, 'label': 'NORP'}]}
```

- Graphical User Interface

Geoffrey Everest Hinton (born 6 December 1947) is a British-Canadian cognitive psychologist and computer scientist, most noted for his work on artificial neural networks.

Max characters 1500

170

ANALYZE

RESET

Geoffrey Everest Hinton **PERSON** (born 6 **CARDINAL**

December 1947 **DATE**) is a **British** **NORP**-Canadian cognitive psychologist and computer scientist, most noted for his work on artificial neural networks.

How to build a customized NER service?

- Data Collection -> Data Annotation -> Model Training & Evaluation -> Model Deployment (API/GUI)

How to build a customized NER service?

- Data Collection -> Data Annotation -> Model Training & Evaluation -> Model Deployment (API/GUI)

How to build a customized NER service?

- Data Collection -> **Data Annotation** -> Model Training & Evaluation -> Model Deployment (API/GUI)

```
>>> raw_dataset[0]
"Geoffrey Everest Hinton (born 6 December 1947) is a British-Canadian cognitive psychologist and computer scientist, mos
```



```
>>> train_dataset[0]
{'text': 'Geoffrey Everest Hinton (born 6 December 1947) is a British-Canadian cognitive psychologist and computer scier
'annotation': [{'token': 'Geoffrey', 'label': 'PERSON', 'iob': 3}, {'token': 'Everest', 'label': 'PERSON', 'iob': 1}, {'
    {'token': 'born', 'label': 'null', 'iob': 2}, {'token': '6', 'label': 'CARDINAL', 'iob': 3}, {'token': 'Decembe
    {'token': ')', 'label': 'null', 'iob': 2}, {'token': 'is', 'label': 'null', 'iob': 2}, {'token': 'a', 'label':
    {'token': '-', 'label': 'null', 'iob': 2}, {'token': 'Canadian', 'label': 'null', 'iob': 2}, {'token': 'cogniti
    {'token': 'and', 'label': 'null', 'iob': 2}, {'token': 'computer', 'label': 'null', 'iob': 2}, {'token': 'scier
    {'token': 'most', 'label': 'null', 'iob': 2}, {'token': 'noted', 'label': 'null', 'iob': 2}, {'token': 'for',
    {'token': 'work', 'label': 'null', 'iob': 2}, {'token': 'on', 'label': 'null', 'iob': 2}, {'token': 'artificial
    {'token': 'networks', 'label': 'null', 'iob': 2}, {'token': '.', 'label': 'null', 'iob': 2}
    ... ]}
```

How to build a customized NER service?

- Data Collection -> **Data Annotation** -> Model Training & Evaluation -> Model Deployment (API/GUI)

```
>>> raw_dataset[0]
"Geoffrey Everest Hinton (born 6 December 1947) is a British-Canadian cognitive psychologist and computer scientist, mos
```



```
>>> train_dataset[0]
{'text': 'Geoffrey Everest Hinton (born 6 December 1947) is a British-Canadian cognitive psychologist and computer scier
'annotation': [{'token': 'Geoffrey', 'label': 'PERSON', 'iob': 3}, {'token': 'Everest', 'label': 'PERSON', 'iob': 1}, {'
    {'token': 'born', 'label': 'null', 'iob': 2}, {'token': '6', 'label': 'CARDINAL', 'iob': 3}, {'token': 'Decembe
    {'token': ')', 'label': 'null', 'iob': 2}, {'token': 'is', 'label': 'null', 'iob': 2}, {'token': 'a', 'label':
    {'token': '-', 'label': 'null', 'iob': 2}, {'token': 'Canadian', 'label': 'null', 'iob': 2}, {'token': 'cogniti
    {'token': 'and', 'label': 'null', 'iob': 2}, {'token': 'computer', 'label': 'null', 'iob': 2}, {'token': 'scier
    {'token': 'most', 'label': 'null', 'iob': 2}, {'token': 'noted', 'label': 'null', 'iob': 2}, {'token': 'for',
    {'token': 'work', 'label': 'null', 'iob': 2}, {'token': 'on', 'label': 'null', 'iob': 2}, {'token': 'artificial
    {'token': 'networks', 'label': 'null', 'iob': 2}, {'token': '.', 'label': 'null', 'iob': 2}
    ... ]}
```

How to build a customized NER service?

- Data Collection -> **Data Annotation** -> Model Training & Evaluation -> Model Deployment (API/GUI)

```
>>> raw_dataset[0]
"Geoffrey Everest Hinton (born 6 December 1947) is a British-Canadian cognitive psychologist and computer scientist, mos
```



```
>>> train_dataset[0]
{'text': 'Geoffrey Everest Hinton (born 6 December 1947) is a British-Canadian cognitive psychologist and computer scier
'annotation': [{'token': 'Geoffrey', 'label': 'PERSON', 'iob': 3}, {'token': 'Everest', 'label': 'PERSON', 'iob': 1}, {'
    {'token': 'born', 'label': 'null', 'iob': 2}, {'token': '6', 'label': 'CARDINAL', 'iob': 3}, {'token': 'Decembe
    {'token': ')', 'label': 'null', 'iob': 2}, {'token': 'is', 'label': 'null', 'iob': 2}, {'token': 'a', 'label':
    {'token': '-', 'label': 'null', 'iob': 2}, {'token': 'Canadian', 'label': 'null', 'iob': 2}, {'token': 'cogniti
    {'token': 'and', 'label': 'null', 'iob': 2}, {'token': 'computer', 'label': 'null', 'iob': 2}, {'token': 'scier
    {'token': 'most', 'label': 'null', 'iob': 2}, {'token': 'noted', 'label': 'null', 'iob': 2}, {'token': 'for',
    {'token': 'work', 'label': 'null', 'iob': 2}, {'token': 'on', 'label': 'null', 'iob': 2}, {'token': 'artificial
    {'token': 'networks', 'label': 'null', 'iob': 2}, {'token': '.', 'label': 'null', 'iob': 2}
    ... ]}
```

Annotation Interface

Annotation Interface

Label Definition

☒ ORG

☒ PRODUCT

☒ GPE

☒ LOC

☒ PERSON

☐ NORP

☐ FACILITY

☐ EVENT

☐ LAW

☐ LANGUAGE

☐ ART

☐ DATE

☐ TIME

☐ MONEY

☐ QUANTITY

☐ ORDINAL

☐ CARDINAL

☐ PERCENT

SELECT ALL

DEFAULT

RESET

Annotation Configuration

☒ Auto Suggestion

☒ Active Learning

Auto Suggestion Strength: 0

100

Uncertainty and Diversity Tradeoff: Uncertainty

Diversity

Model

en_core_web_sm

Annotation Interface

Label Definition

☒ ORG

☒ PRODUCT

☒ GPE

☒ LOC

☒ PERSON

☐ NORP

☐ FACILITY

☐ EVENT

☐ LAW

☐ LANGUAGE

☐ ART

☐ DATE

☐ TIME

☐ MONEY

☐ QUANTITY

☐ ORDINAL

☐ CARDINAL

☐ PERCENT

SELECT ALL

DEFAULT

RESET

Annotation Configuration

☒ Auto Suggestion

☒ Active Learning

Auto Suggestion Strength: 0

100

Uncertainty and Diversity Tradeoff: Uncertainty

Diversity

Model

en_core_web_sm

Annotation

←

→

✓

✕

★

🚫

☒ ORG

☐ PRODUCT

☐ GPE

☐ LOC

☐ PERSON

Geoffrey Everest Hinton CC **FRS FRSC** **PRODUCT** (born 6 December 1947) is a British - Canadian cognitive psychologist and computer scientist , most noted for his work on artificial neural networks . Since 2013 , he has divided his time working for Google (Google Brain) and

the University of Toronto **ORG** . In 2017 , he co - founded and became the Chief Scientific Advisor of **the Vector Institute** **ORG** in **Toronto** **GPE** . With **David Rumelhart** **PERSON** and **Ronald J. Williams** **PERSON** , **Hinton** **PERSON** was co - author of a highly cited paper published in 1986 that popularized the backpropagation algorithm for training multi - layer neural networks , although they were not the first to propose the approach . **Hinton** **ORG** is viewed as a leading figure in the deep learning community . The dramatic image - recognition milestone of the **AlexNet** **ORG** designed in collaboration with his students **Alex Krizhevsky** **PERSON** and **Ilya Sutskever** **PERSON** for the **ImageNet** **ORG** challenge 2012 was a breakthrough in the field of computer vision . **Hinton** **PERSON** received the 2018 Turing Award , together with **Yoshua Bengio** **PERSON** and **Yann LeCun** **PERSON** , for their work on deep learning . They are sometimes referred to as the " Godfathers of AI " and " Godfathers of Deep Learning " , and have continued to give public talks together .

Data Quality Control (In Progress)

```
>>> db.examples[0]
{'text': 'Geoffrey Everest Hinton (born 6 December 1947)',
 'annotations': [{
   'user1': [{
     'token': 'Geoffrey', 'label': 'PERSON', 'iob': 1},
     {'token': 'Everest', 'label': 'PERSON', 'iob': 1},
     {'token': 'Hinton', 'label': 'PERSON', 'iob': 1},
     {'token': 'born', 'label': 'null', 'iob': 2},
     {'token': '6', 'label': 'CARDINAL', 'iob': 3},
     {'token': 'December', 'label': 'DATE', 'iob': 3},
     ... ],
   'user2': [{
     'token': 'Geoffrey', 'label': 'null', 'iob': 1},
     {'token': 'Everest', 'label': 'null', 'iob': 2},
     {'token': 'Hinton', 'label': 'null', 'iob': 2},
     {'token': 'born', 'label': 'null', 'iob': 2},
     {'token': '6', 'label': 'DATE', 'iob': 3},
     {'token': 'December', 'label': 'DATE', 'iob': 3},
     ... ]}]
```

Data Quality Control (In Progress)

```
>>> db.examples[0]
{'text': 'Geoffrey Everest Hinton (born 6 December 1947)',
 'annotations': [{
   'user1': [{ 'token': 'Geoffrey', 'label': 'PERSON', 'iob': 1},
             { 'token': 'Everest', 'label': 'PERSON', 'iob': 1},
             { 'token': 'Hinton', 'label': 'PERSON', 'iob': 1},
             { 'token': 'born', 'label': 'null', 'iob': 2},
             { 'token': '6', 'label': 'CARDINAL', 'iob': 3},
             { 'token': 'December', 'label': 'DATE', 'iob': 3},
             ... ],
   'user2': [{ 'token': 'Geoffrey', 'label': 'null', 'iob': 1},
             { 'token': 'Everest', 'label': 'null', 'iob': 2},
             { 'token': 'Hinton', 'label': 'null', 'iob': 2},
             { 'token': 'born', 'label': 'null', 'iob': 2},
             { 'token': '6', 'label': 'DATE', 'iob': 3},
             { 'token': 'December', 'label': 'DATE', 'iob': 3},
             ... ]}]
```


Data Quality Control (In Progress)

```
>>> db.examples[0]
{'text': 'Geoffrey Everest Hinton (born 6 December 1947)
'annotations': [{
  'user1': [{'token': 'Geoffrey', 'label': 'PERSON', 'iob': 1},
    {'token': 'Everest', 'label': 'PERSON', 'iob': 1},
    {'token': 'Hinton', 'label': 'PERSON', 'iob': 1},
    {'token': 'born', 'label': 'null', 'iob': 2},
    {'token': '6', 'label': 'CARDINAL', 'iob': 3},
    {'token': 'December', 'label': 'DATE', 'iob': 3},
    ... ],
  'user2': [{'token': 'Geoffrey', 'label': 'null', 'iob': 1},
    {'token': 'Everest', 'label': 'null', 'iob': 2},
    {'token': 'Hinton', 'label': 'null', 'iob': 2},
    {'token': 'born', 'label': 'null', 'iob': 2},
    {'token': '6', 'label': 'DATE', 'iob': 3},
    {'token': 'December', 'label': 'DATE', 'iob': 3},
    ... ]}]
```

■ User 1

Geoffrey Everest Hinton **PERSON** (born **6** **CARDINAL** **December 1947** **DATE**) is a **British** **NORP** - Canadian
cognitive psychologist and computer scientist , most noted for his work on artificial neural networks .

■ User 2

Geoffrey Everest Hinton (born **6 December 1947** **DATE**) is a **British** **NORP** - Canadian cognitive
psychologist and computer scientist , most noted for his work on artificial neural networks .

Data Quality Control (In Progress)

```
>>> db.examples[0]
{'text': 'Geoffrey Everest Hinton (born 6 December 1947)
'annotations': [{
  'user1': [{'token': 'Geoffrey', 'label': 'PERSON', 'iob': 1},
    {'token': 'Everest', 'label': 'PERSON', 'iob': 1},
    {'token': 'Hinton', 'label': 'PERSON', 'iob': 1},
    {'token': 'born', 'label': 'null', 'iob': 2},
    {'token': '6', 'label': 'CARDINAL', 'iob': 3},
    {'token': 'December', 'label': 'DATE', 'iob': 3},
    ... ],
  'user2': [{'token': 'Geoffrey', 'label': 'null', 'iob': 2},
    {'token': 'Everest', 'label': 'null', 'iob': 2},
    {'token': 'Hinton', 'label': 'null', 'iob': 2},
    {'token': 'born', 'label': 'null', 'iob': 2},
    {'token': '6', 'label': 'DATE', 'iob': 3},
    {'token': 'December', 'label': 'DATE', 'iob': 3},
    ... ]}]
```

■ User 1

Geoffrey Everest Hinton **PERSON** (born **6** **CARDINAL** **December 1947** **DATE**) is a **British** **NORP** - Canadian cognitive psychologist and computer scientist , most noted for his work on artificial neural networks .

■ User 2

Geoffrey Everest Hinton (born **6 December 1947** **DATE**) is a **British** **NORP** - Canadian cognitive psychologist and computer scientist , most noted for his work on artificial neural networks .

■ Disagreement Visualization

■ Help needed

Data Quality Control (In Progress)

```
>>> db.examples[0]
{'text': 'Geoffrey Everest Hinton (born 6 December 1947)
'annotations': [{
  'user1': [{'token': 'Geoffrey', 'label': 'PERSON', 'iob': 1},
    {'token': 'Everest', 'label': 'PERSON', 'iob': 1},
    {'token': 'Hinton', 'label': 'PERSON', 'iob': 1},
    {'token': 'born', 'label': 'null', 'iob': 2},
    {'token': '6', 'label': 'CARDINAL', 'iob': 3},
    {'token': 'December', 'label': 'DATE', 'iob': 3},
    ... ],
  'user2': [{'token': 'Geoffrey', 'label': 'null', 'iob': 1},
    {'token': 'Everest', 'label': 'null', 'iob': 2},
    {'token': 'Hinton', 'label': 'null', 'iob': 2},
    {'token': 'born', 'label': 'null', 'iob': 2},
    {'token': '6', 'label': 'DATE', 'iob': 3},
    {'token': 'December', 'label': 'DATE', 'iob': 3},
    ... ]}]
```

■ User 1

Geoffrey Everest Hinton **PERSON** (born **6** **CARDINAL** **December 1947** **DATE**) is a **British** **NORP** - Canadian cognitive psychologist and computer scientist , most noted for his work on artificial neural networks .

■ User 2

Geoffrey Everest Hinton (born **6 December 1947** **DATE**) is a **British** **NORP** - Canadian cognitive psychologist and computer scientist , most noted for his work on artificial neural networks .

■ Disagreement Visualization

■ Help needed

■ Golden Standard

Geoffrey Everest Hinton **PERSON** (born **6 December 1947** **DATE**) is a **British** **NORP** - Canadian cognitive psychologist and computer scientist , most noted for his work on artificial neural networks .

Model Training

```
from transformers import DistilBertForTokenClassification, Trainer, TrainingArguments
training_args = TrainingArguments(
    num_train_epochs=3,          # total number of training epochs
    per_device_train_batch_size=16, # batch size per device during training
    per_device_eval_batch_size=64, # batch size for evaluation
    warmup_steps=500,           # number of warmup steps for learning rate scheduler
    weight_decay=0.01,          # strength of weight decay
)

model = DistilBertForTokenClassification.from_pretrained('distilbert-base-cased', num_labels=len(unique_tags))

trainer = Trainer(
    model=model,                 # the instantiated 🤗 Transformers model to be trained
    args=training_args,          # training arguments, defined above
    train_dataset=train_dataset,  # training dataset
    eval_dataset=val_dataset     # evaluation dataset
)

trainer.train()
```

Model Training

```
from transformers import DistilBertForTokenClassification, Trainer, TrainingArguments
training_args = TrainingArguments(
    num_train_epochs=3,          # total number of training epochs
    per_device_train_batch_size=16, # batch size per device during training
    per_device_eval_batch_size=64, # batch size for evaluation
    warmup_steps=500,           # number of warmup steps for learning rate scheduler
    weight_decay=0.01,          # strength of weight decay
)

model = DistilBertForTokenClassification.from_pretrained('distilbert-base-cased', num_labels=len(unique_tags))

trainer = Trainer(
    model=model,                 # the instantiated 🤗 Transformers model to be trained
    args=training_args,          # training arguments, defined above
    train_dataset=train_dataset, # training dataset
    eval_dataset=val_dataset     # evaluation dataset
)

trainer.train()
```

Model Training

```
from transformers import DistilBertForTokenClassification, Trainer, TrainingArguments

training_args = TrainingArguments(
    num_train_epochs=3,          # total number of training epochs
    per_device_train_batch_size=16, # batch size per device during training
    per_device_eval_batch_size=64, # batch size for evaluation
    warmup_steps=500,           # number of warmup steps for learning rate scheduler
    weight_decay=0.01,          # strength of weight decay
)

model = DistilBertForTokenClassification.from_pretrained('distilbert-base-cased', num_labels=len(unique_tags))

trainer = Trainer(
    model=model,                 # the instantiated 🤗 Transformers model to be trained
    args=training_args,         # training arguments, defined above
    train_dataset=train_dataset, # training dataset
    eval_dataset=val_dataset    # evaluation dataset
)

trainer.train()
```

Model Training

```
from transformers import DistilBertForTokenClassification, Trainer, TrainingArguments

training_args = TrainingArguments(
    num_train_epochs=3,          # total number of training epochs
    per_device_train_batch_size=16, # batch size per device during training
    per_device_eval_batch_size=64, # batch size for evaluation
    warmup_steps=500,           # number of warmup steps for learning rate scheduler
    weight_decay=0.01,          # strength of weight decay
)

model = DistilBertForTokenClassification.from_pretrained('distilbert-base-cased', num_labels=len(unique_tags))

trainer = Trainer(
    model=model,                  # the instantiated 🤗 Transformers model to be trained
    args=training_args,          # training arguments, defined above
    train_dataset=train_dataset,  # training dataset
    eval_dataset=val_dataset     # evaluation dataset
)

trainer.train()
```

Model Training

```
from transformers import DistilBertForTokenClassification, Trainer, TrainingArguments
training_args = TrainingArguments(
    num_train_epochs=3,          # total number of training epochs
    per_device_train_batch_size=16, # batch size per device during training
    per_device_eval_batch_size=64, # batch size for evaluation
    warmup_steps=500,           # number of warmup steps for learning rate scheduler
    weight_decay=0.01,          # strength of weight decay
)

model = DistilBertForTokenClassification.from_pretrained('distilbert-base-cased', num_labels=len(unique_tags))

trainer = Trainer(
    model=model,                 # the instantiated 🤗 Transformers model to be trained
    args=training_args,          # training arguments, defined above
    train_dataset=train_dataset, # training dataset
    eval_dataset=val_dataset     # evaluation dataset
)

trainer.train()
```


Model Training

```
from transformers import DistilBertForTokenClassification, Trainer, TrainingArguments

training_args = TrainingArguments(
    num_train_epochs=3,          # total number of training epochs
    per_device_train_batch_size=16, # batch size per device during training
    per_device_eval_batch_size=64, # batch size for evaluation
    warmup_steps=500,           # number of warmup steps for learning rate scheduler
    weight_decay=0.01,          # strength of weight decay
)

model = DistilBertForTokenClassification.from_pretrained('distilbert-base-cased', num_labels=len(unique_tags))

trainer = Trainer(
    model=model,                 # the instantiated 🤗 Transformers model to be trained
    args=training_args,         # training arguments, defined above
    train_dataset=train_dataset, # training dataset
    eval_dataset=val_dataset     # evaluation dataset
)

trainer.train()
```

Model Training

```
from transformers import DistilBertForTokenClassification, Trainer, TrainingArguments
training_args = TrainingArguments(
    num_train_epochs=3,          # total number of training epochs
    per_device_train_batch_size=16, # batch size per device during training
    per_device_eval_batch_size=64, # batch size for evaluation
    warmup_steps=500,           # number of warmup steps for learning rate scheduler
    weight_decay=0.01,          # strength of weight decay
)

model = DistilBertForTokenClassification.from_pretrained('distilbert-base-cased', num_labels=len(unique_tags))

trainer = Trainer(
    model=model,                 # the instantiated 🤗 Transformers model to be trained
    args=training_args,         # training arguments, defined above
    train_dataset=train_dataset, # training dataset
    eval_dataset=val_dataset     # evaluation dataset
)

trainer.train()
```

Model Selection using A/B Test (In progress)

Model Selection using A/B Test (In progress)

- Prediction from Trained Model 1

Geoffrey Everest Hinton **PERSON** (born **6** **CARDINAL** **December 1947** **DATE**) is a **British** **NORP** - Canadian cognitive psychologist and computer scientist , most noted for his work on artificial neural networks .

- Prediction from Trained Model 2

Geoffrey Everest Hinton (born **6 December 1947** **DATE**) is a **British** **NORP** - Canadian cognitive psychologist and computer scientist , most noted for his work on artificial neural networks .

Model Selection using A/B Test (In progress)

- Prediction from Trained Model 1

Geoffrey Everest Hinton **PERSON** (born **6** **CARDINAL** **December 1947** **DATE**) is a **British** **NORP** - Canadian cognitive psychologist and computer scientist , most noted for his work on artificial neural networks .

- Prediction from Trained Model 2

Geoffrey Everest Hinton (born **6 December 1947** **DATE**) is a **British** **NORP** - Canadian cognitive psychologist and computer scientist , most noted for his work on artificial neural networks .

- They both make one mistake and have the same accuracy. Which model to choose?
 - Let user make the decision.
 - E.g. A/B Test with 100 examples.
 - Model 1 has 68/100 "likes" and Model 2 has 55/100 "likes" -> Deploy model 1

Next Steps

- Functionality
 - Complete the data quality control interface
 - Model Selection with A/B testing
 - Data annotation with active learning
- Experiment
 - Benchmark the performance of different active learning algorithm on NER task
- Demo
 - Build a customized NER service for a specific domain (Medical Data)