NLP From Scratch

Solving the cold start problem for natural language processing

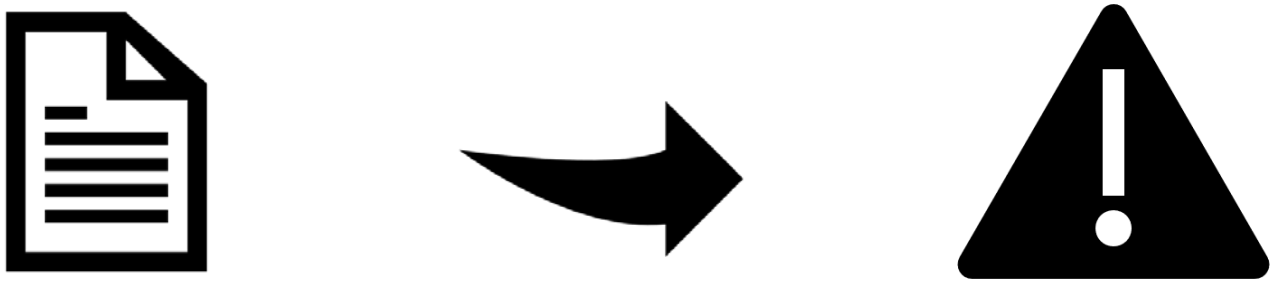
San Francisco, CA

Norris Heintzelman

Michael Johnson

(obligatory picture of F-35)

How can we extract supply chain risk from open source unstructured data?

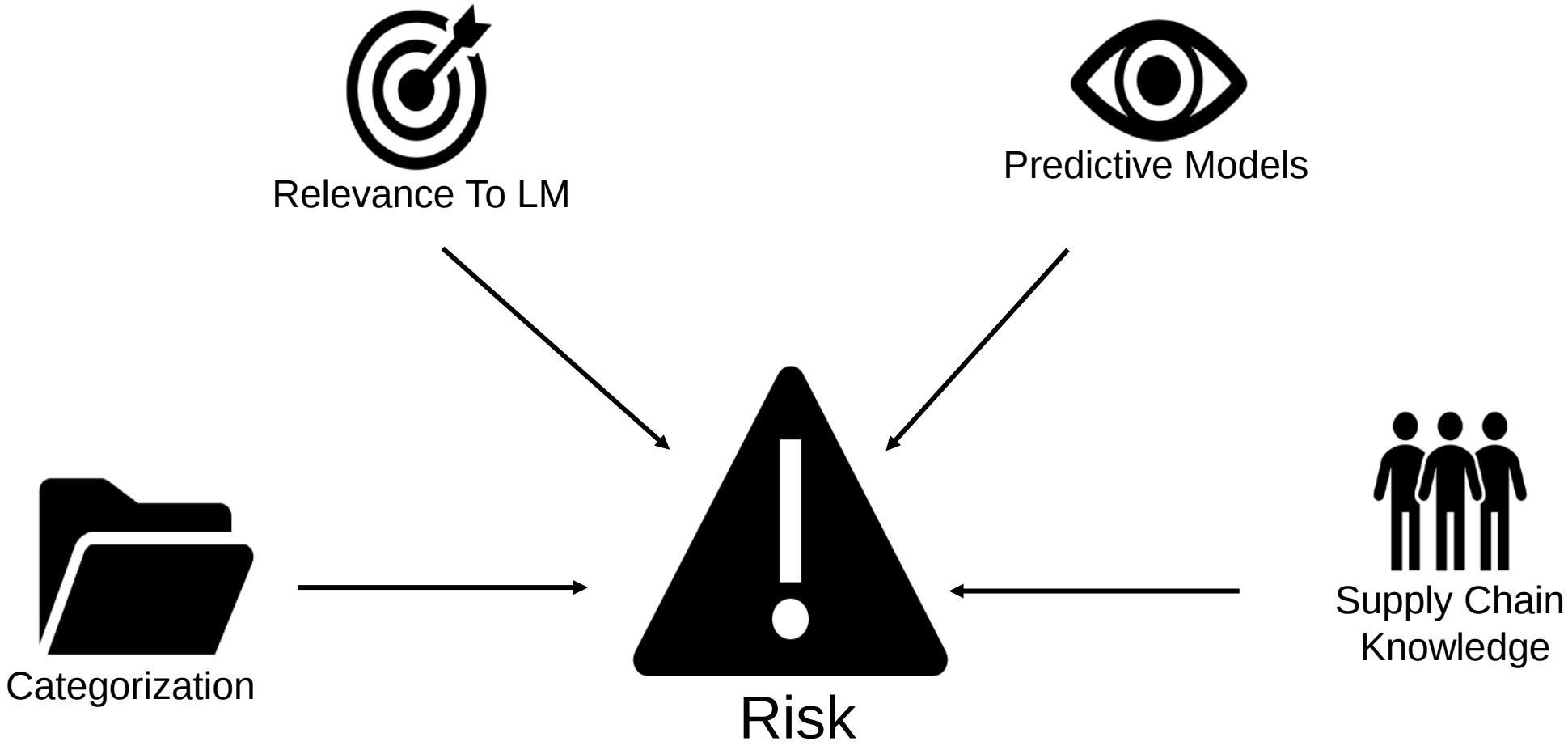


Exclusive: UTC set to win EU approval for $23 billion Rockwell Collins deal

The deal, announced in September last year, would create a new player in the top echelon of suppliers to Boeing, Airbus, Bombardier, and other plane makers

- Contract negotiations
- Part number integrity
- Business Stability

How Do You Define Risk?

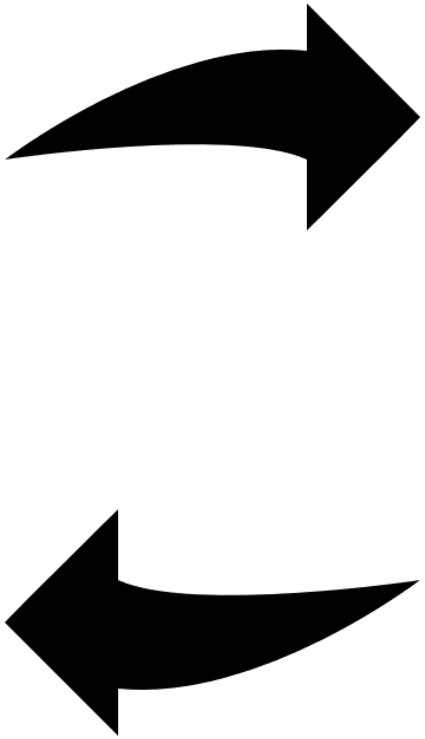


Typical NLP approaches require lots of human labeled training data

Hybrid Systems

Human

‘Truth’, Intuition



Machine

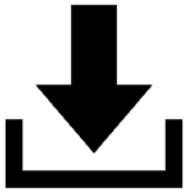
Speed, Consistency

Cold Start Problem

How do you train a machine learning model with no training data?

Transfer Learning + Active Learning

Hybrid Rules/Machine Learning



Use Pre Trained Models

Creative Training Sets

Weak Supervision

Rule/Heuristic

- ✓ High precision
- ✓ Easily written/evaluated by humans
- ❖ Rigid/Fragile

VS

Machine Learning

- ✓ Can Generalize
- ✓ Probabilistic
- ❖ Requires hand labeled data

Weak Supervision

Rule/Heuristic

Feed rules-based labels to ML Model

Machine Learning

ML Model will just memorize rules

Weak Supervision

Rule/Heuristic

Apply Rules to Titles

UTC set to win EU approval for $23 billion Rockwell Collins deal

The deal, announced in September last year, would create a new player in the top echelon of suppliers to Boeing, Airbus, Bombardier, and other plane makers…

Feed text to ML Model

Machine Learning

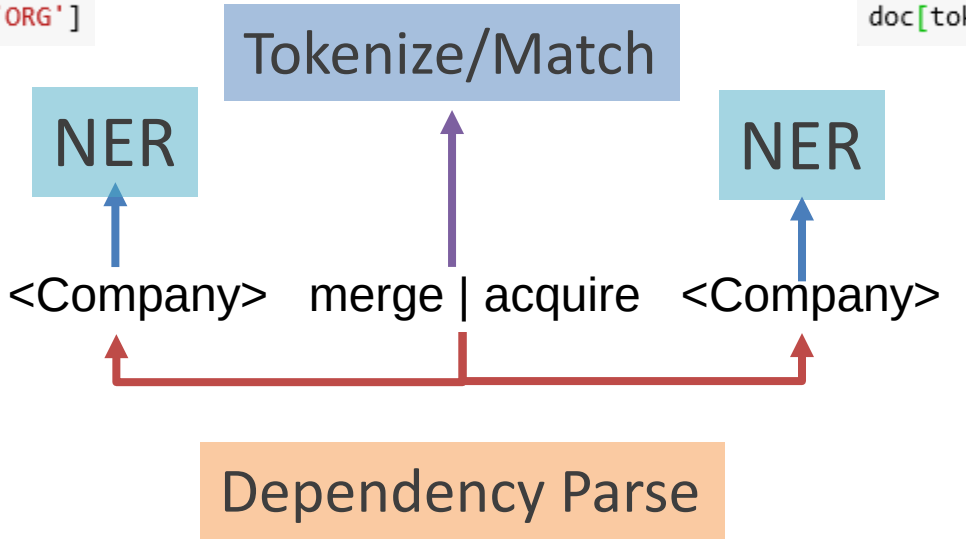
ML Model will learn flexible rules

How do you write a rule?

Example Rule: Merger and Acquisition Using spaCy

```
matcher.add("MandA",None,
            [{'LEMMA': 'acquire'}],
            [{'LEMMA': 'merge'}],
            [{'LEMMA': 'acquisition'}])
```

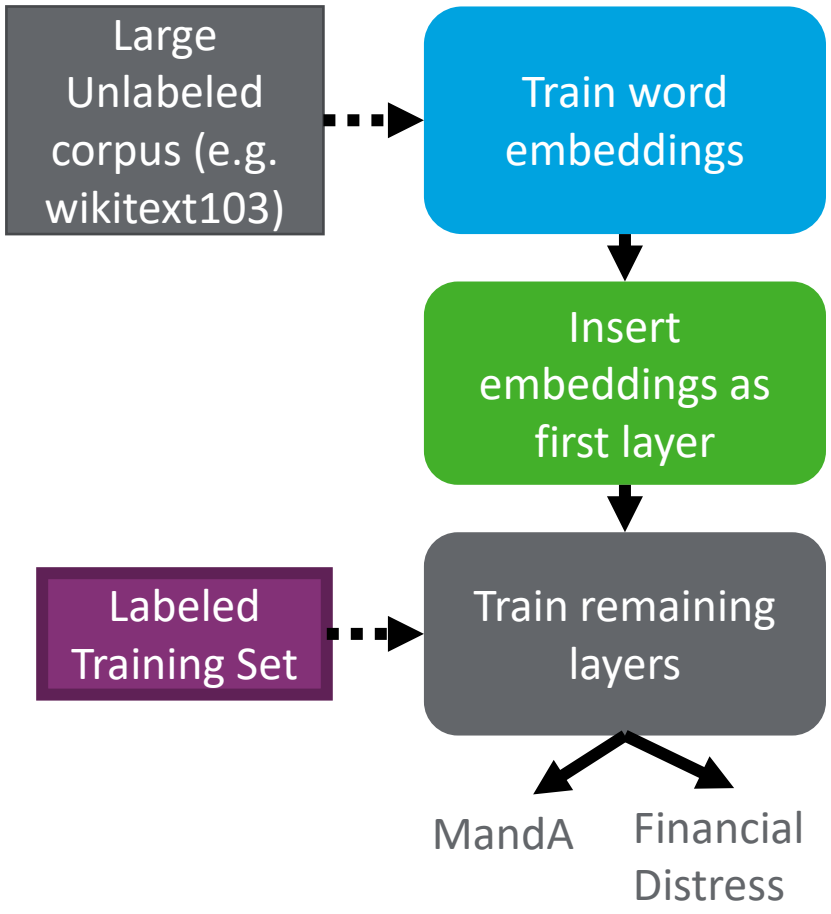
```
doc[token].ent_type_ in ['PERSON','ORG']
```



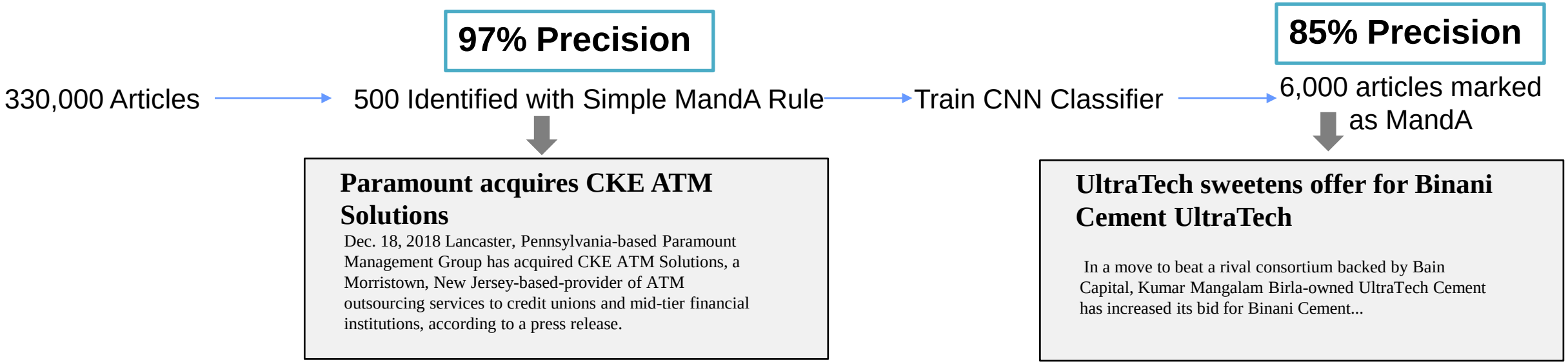

```
doc[token].ent_type_ in ['PERSON','ORG']
```

```
for word in MorA.rights:
    if(word.ent_type_) in['PERSON','ORG']:
        dependentCompany = Word
```

Example Model: WV Pre-Training + CNN



Weakly Supervised Learning: Performance



Training considerations:

- ✓ Carefully control sampling of majority (negative) class
- ✓ Pre-train your word vectors
  - ✓ If you have a large enough corpus (~50m tokens) use that
  - ✓ If not, use pretrained (fasttext, glove, word2vec)

Works well when:

- ✓ You have lots of unlabeled data
- ✓ Your class target class is a minority
- ✓ You can write a high precision rule

Can we apply a similar strategy to NER?

Custom Named Entity Recognition (NER) Model

- Training usually lots of humans marking text

Based in Santa Maria, 7 Days Plumbing Repairs Plus is a plumbing contractor that provides water purification system installation, leak detection and other services.

Based in Santa Maria, 7 Days Plumbing Repairs Plus is a plumbing contractor that provides water purification system installation, leak detection and other services.

Date

Location

Organization

Product/Service

Custom NER Model using SpaCy

Based in Santa Maria, 7 Days Plumbing Repairs Plus is a plumbing contractor that provides water purification system installation, leak detection and other services.

SpaCy Date
SpaCy Location

Custom NER Model using SpaCy

Based in Santa Maria, 7 Days Plumbing Repairs Plus is a plumbing contractor that provides water purification system installation, leak detection and other services.

SpaCy Date
SpaCy Location
Company Name Rules

Custom NER Model using SpaCy

Based in Santa Maria, 7 Days Plumbing Repairs Plus is a plumbing contractor that provides water purification system installation, leak detection and other services.

SpaCy Date
SpaCy Location
Company Name Rules
Other Engine Organization

Custom NER Model using SpaCy

Based in Santa Maria, 7 Days Plumbing Repairs Plus is a plumbing contractor that provides water purification system installation, leak detection and other services.

SpaCy Date
SpaCy Location
Company Name Rules
Other Engine Organization
Products/Service Rules

Custom NER Model using SpaCy

Based in Santa Maria, 7 Days Plumbing Repairs Plus is a plumbing contractor that provides water purification system installation, leak detection and other services.

SpaCy Date

SpaCy Location

Company Name Rules

Other Engine Organization

Products/Service Rules

- **But we need non-overlapping labels (AKA Annotations)**
- **So we assign confidence values, and in an overlap, pick the highest confidence**

– Company Name Rules	High Confidence
– Products/Service Rules	High Confidence
– SpaCy Date	Medium Confidence
– SpaCy Location	Medium Confidence
– Other Engine Organization	Low Confidence

- **Note: the assignment of confidences to particular situations could be quite sophisticated**

Custom NER Model using SpaCy

Based in Santa Maria, 7 Days Plumbing Repairs Plus is a plumbing contractor that provides water purification system installation, leak detection and other services.

Date

Location

Organization

Product/Service

- **Now we just need several thousand more such sentences and some labor to review, correct, add variability**

Real Results:

“Rockwell Collins (COL) surpassed the Zacks Consensus Estimate in three of the trailing four quarters, with an average beat of 2.61%.”

Spacy English Large Out of Box Model

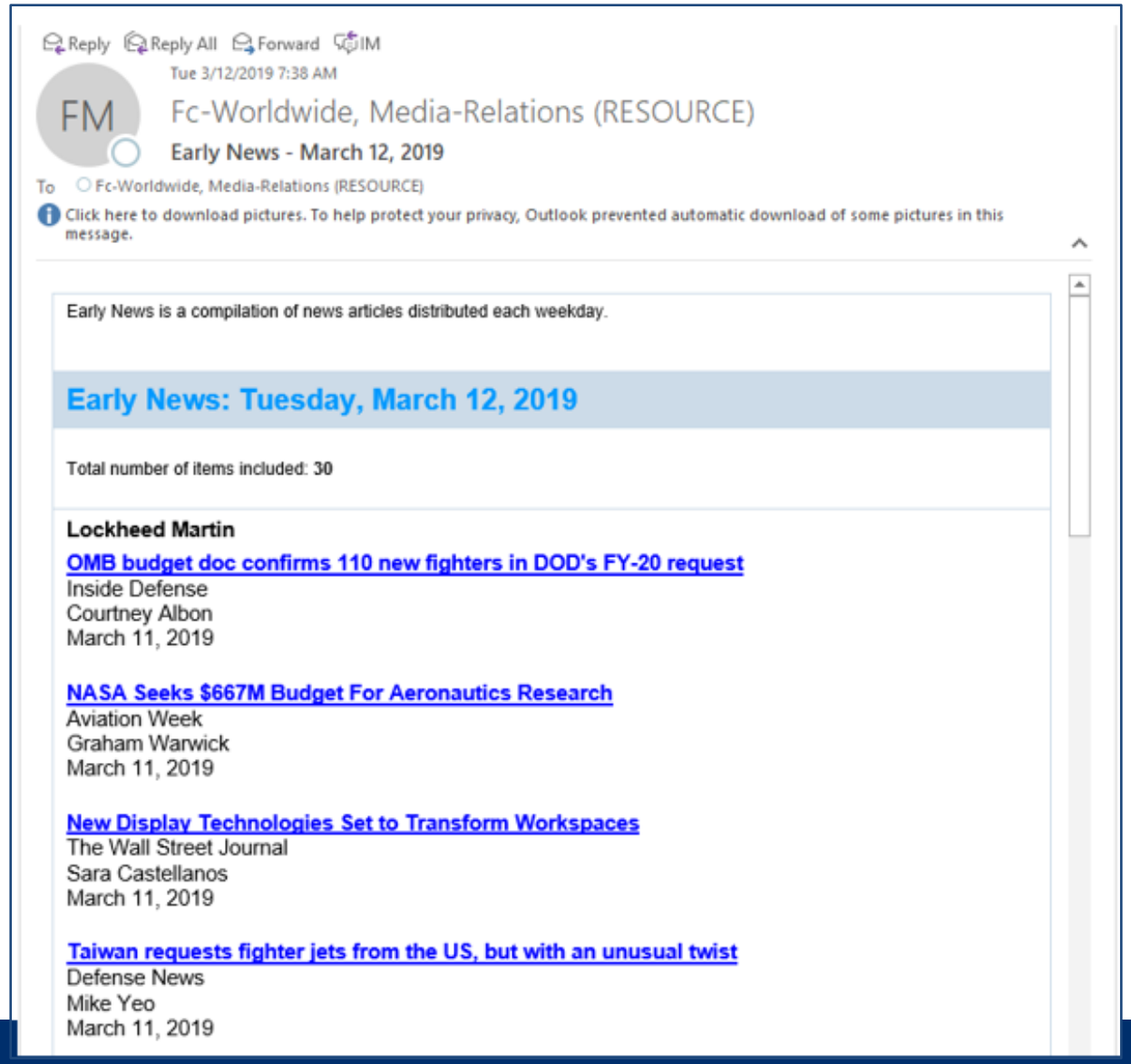
Rockwell Collins	PERSON
COL	ORG
the Zacks Consensus Estimate	ORG
three	CARDINAL
four quarters	DATE
2.61%	PERCENT

New Spacy Model trained on Spacy Eng Large Tags plus Name Rules

Rockwell Collins	ORG
COL	ORG
the Zacks Consensus Estimate	ORG
three	CARDINAL
four quarters	DATE
2.61%	PERCENT

What if training data already exists, but not where you expect it?

Creative Training Sets



Reply Reply All Forward IM

Tue 3/12/2019 7:38 AM

FM

Fc-Worldwide, Media-Relations (RESOURCE)

Early News - March 12, 2019

To Fc-Worldwide, Media-Relations (RESOURCE)

Click here to download pictures. To help protect your privacy, Outlook prevented automatic download of some pictures in this message.

Early News is a compilation of news articles distributed each weekday.

Early News: Tuesday, March 12, 2019

Total number of items included: 30

Lockheed Martin

OMB budget doc confirms 110 new fighters in DOD's FY-20 request
Inside Defense
Courtney Albon
March 11, 2019

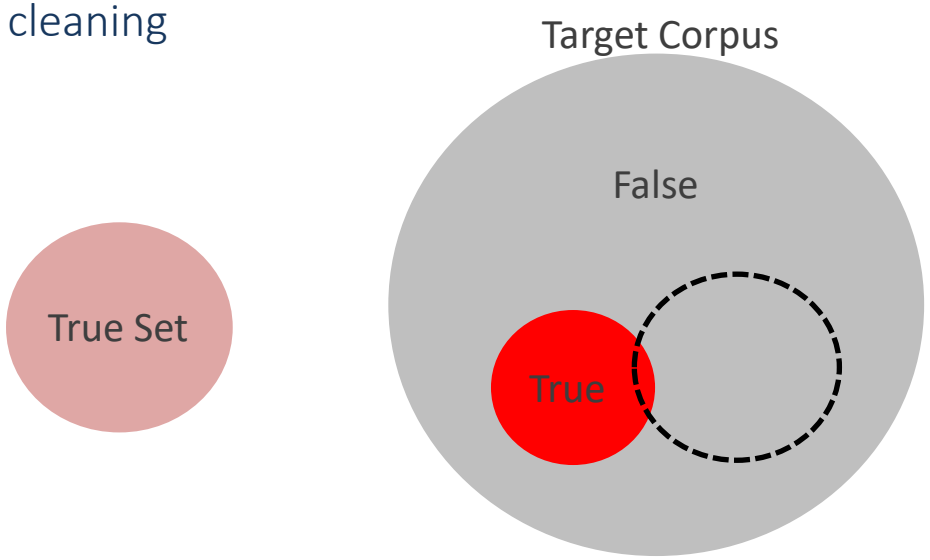
NASA Seeks $667M Budget For Aeronautics Research
Aviation Week
Graham Warwick
March 11, 2019

New Display Technologies Set to Transform Workspaces
The Wall Street Journal
Sara Castellanos
March 11, 2019

Taiwan requests fighter jets from the US, but with an unusual twist
Defense News
Mike Yeo
March 11, 2019

- **83k hand picked news stories from 2000-2018**
- **Challenges:**
  - Only positive examples!
  - Generalize model

- **How do you train w/only positive examples**
  - Joint WV training
  - Balance sampling
  - Data cleaning



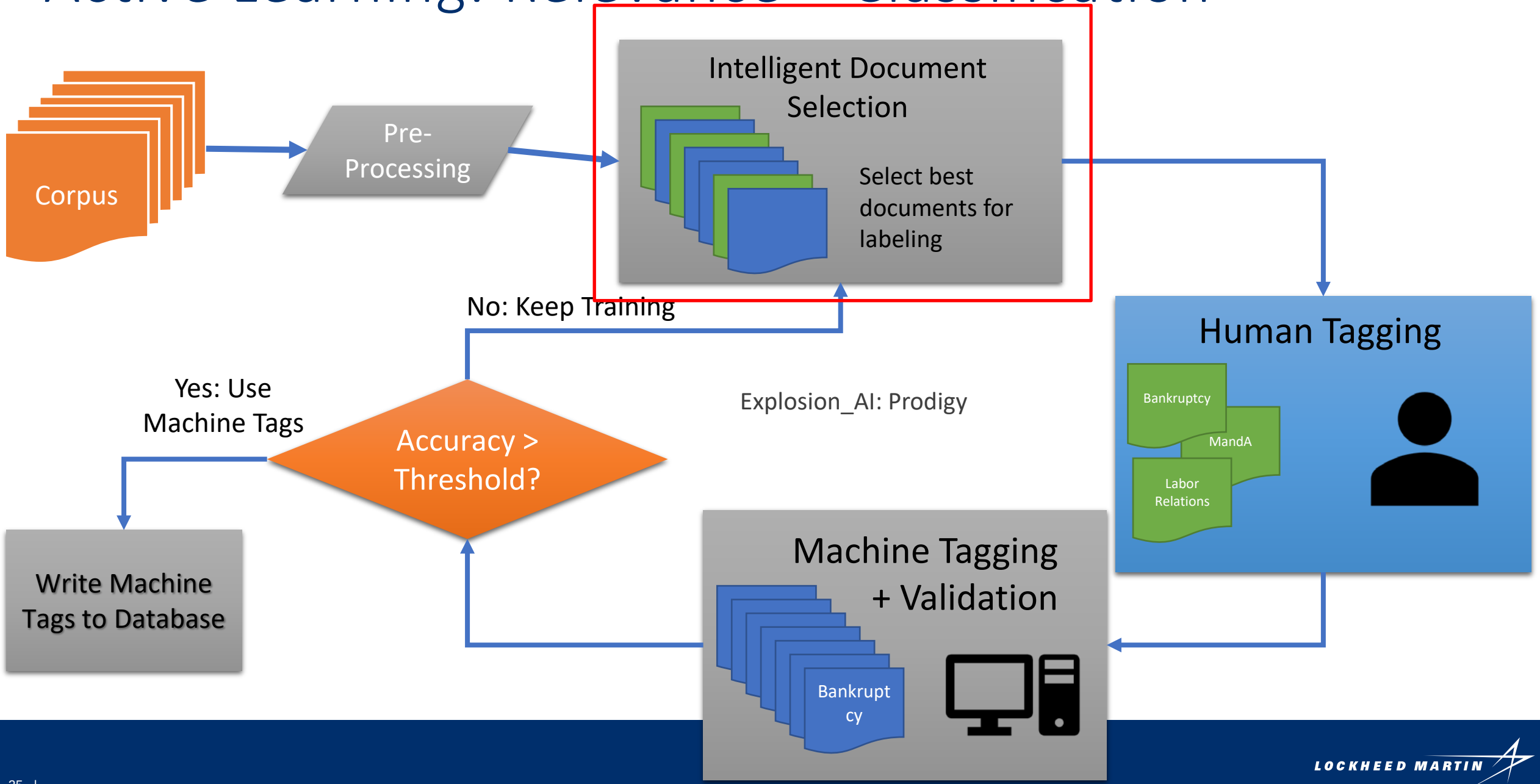
LOCKHEED MARTIN

Models Learn Obvious Features

- **Sikorsky S-97 Raider** **vs** **F-35**

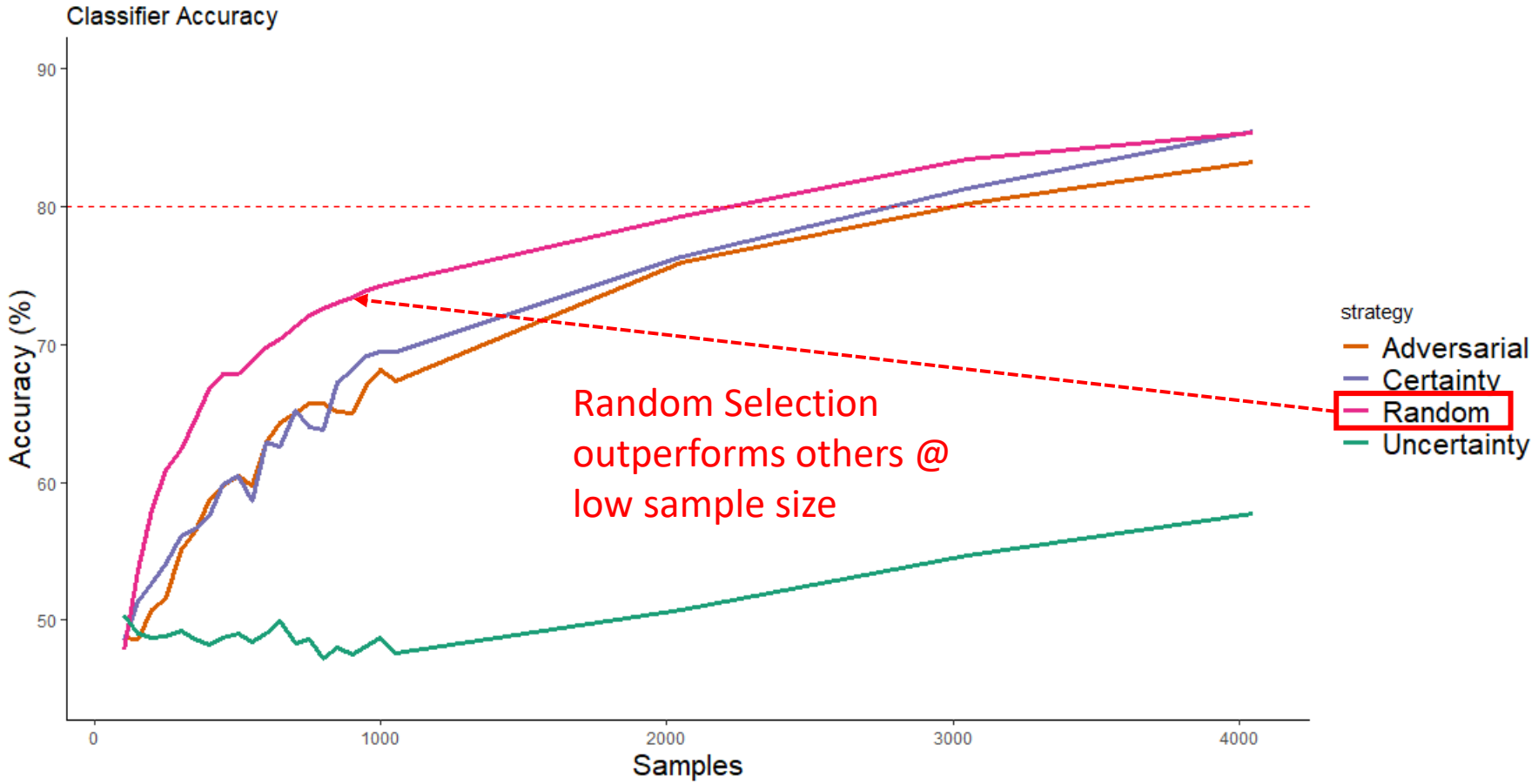
How do you collect ground truth?

Active Learning: Relevance + Classification

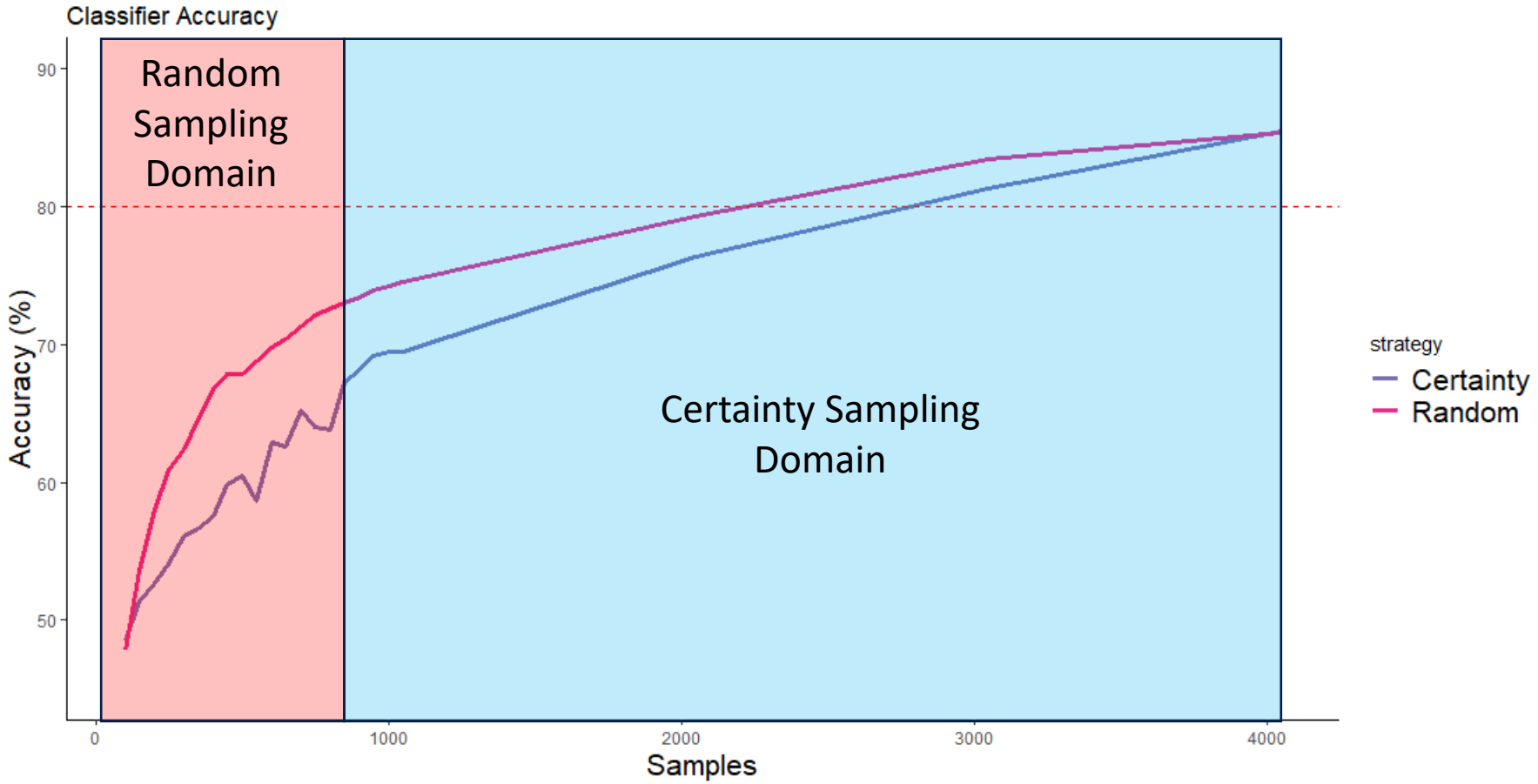


LOCKHEED MARTIN

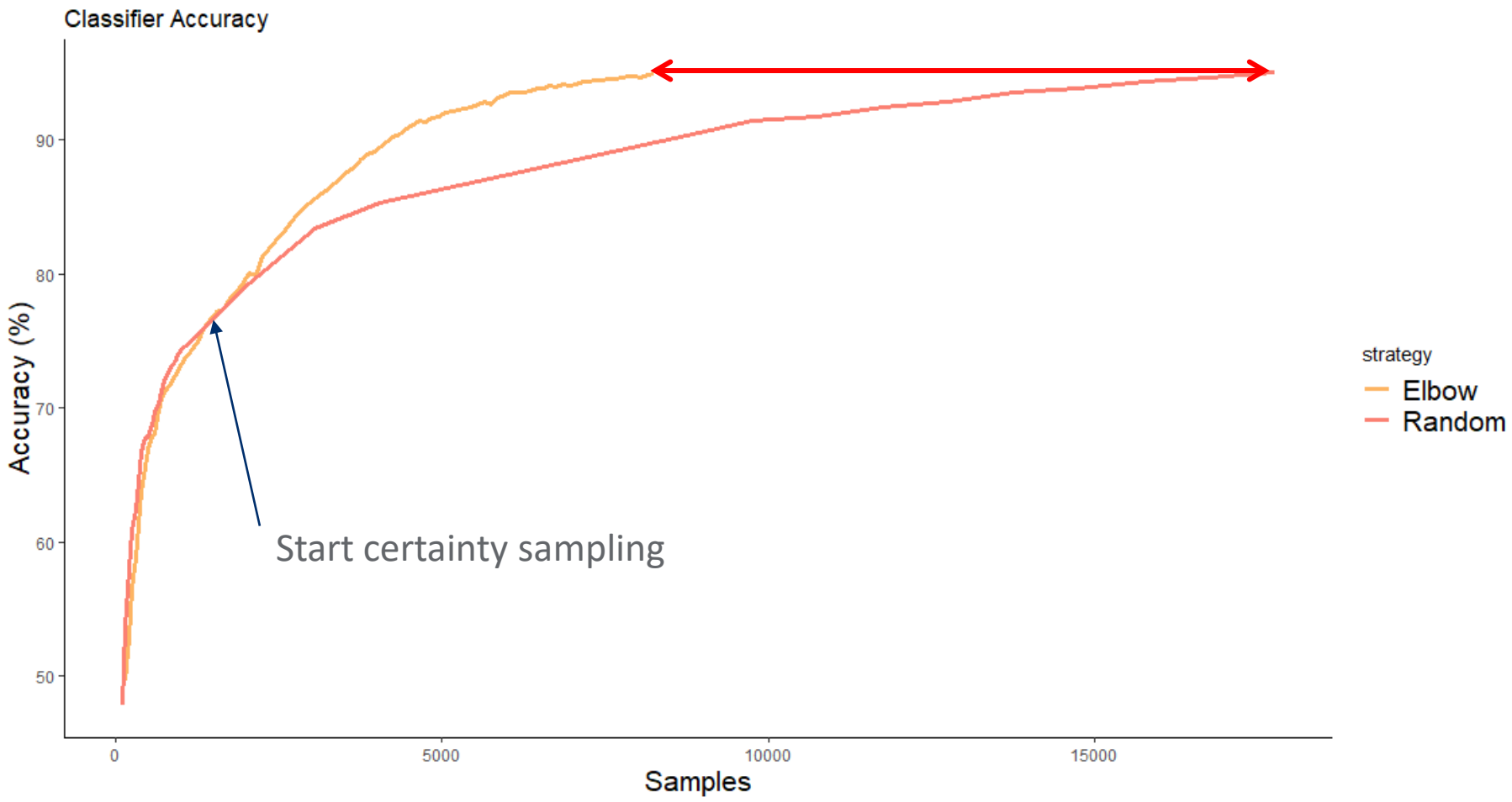
Adversarial Sampling



‘Elbow’ Sampling



Elbow Sampling Results



Elbow sampling requires **2X fewer samples** to achieve full accuracy

Capture Human Ground Truth as Part of Process

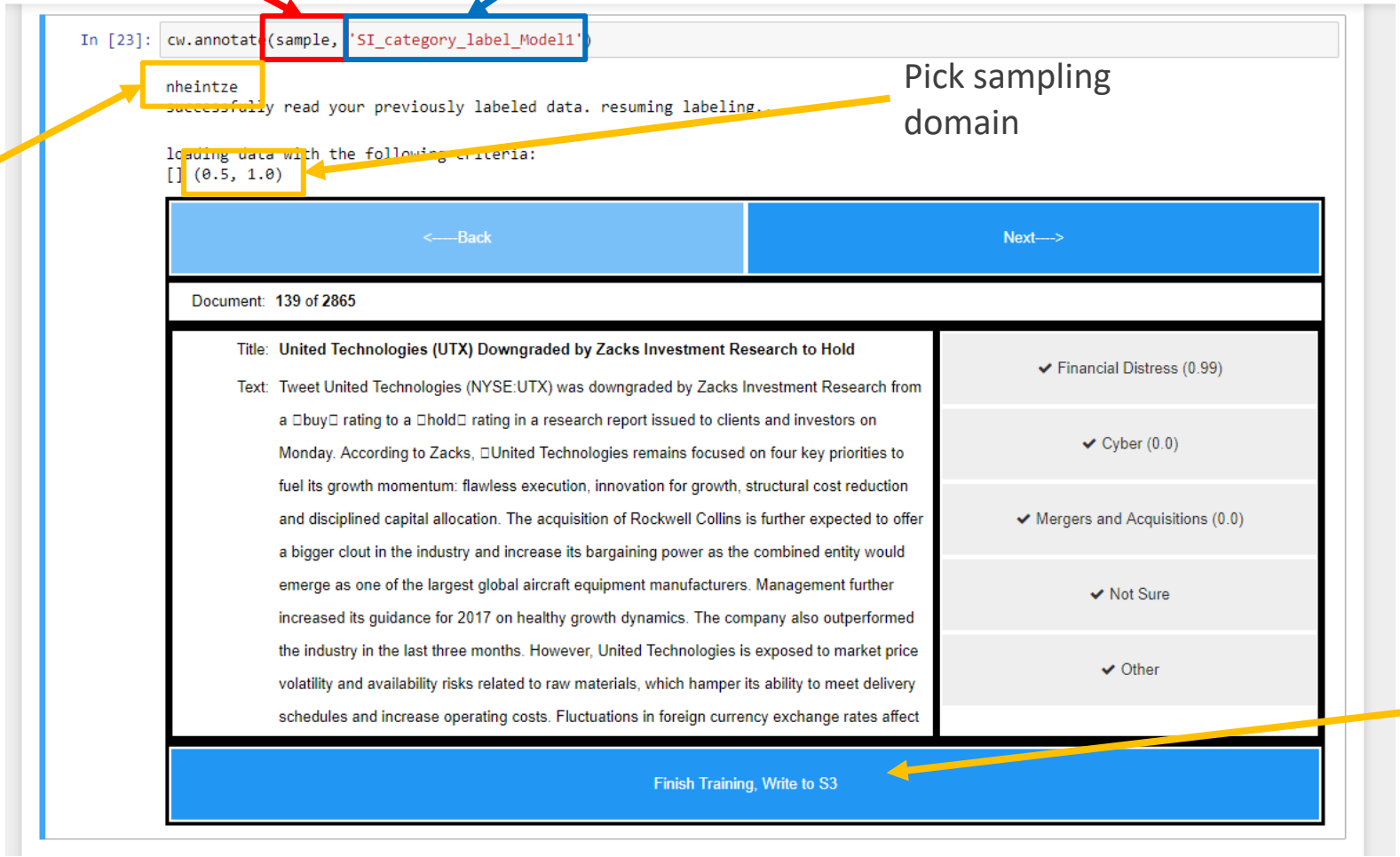
Reads pandas data frame

Loads configuration/previous labels

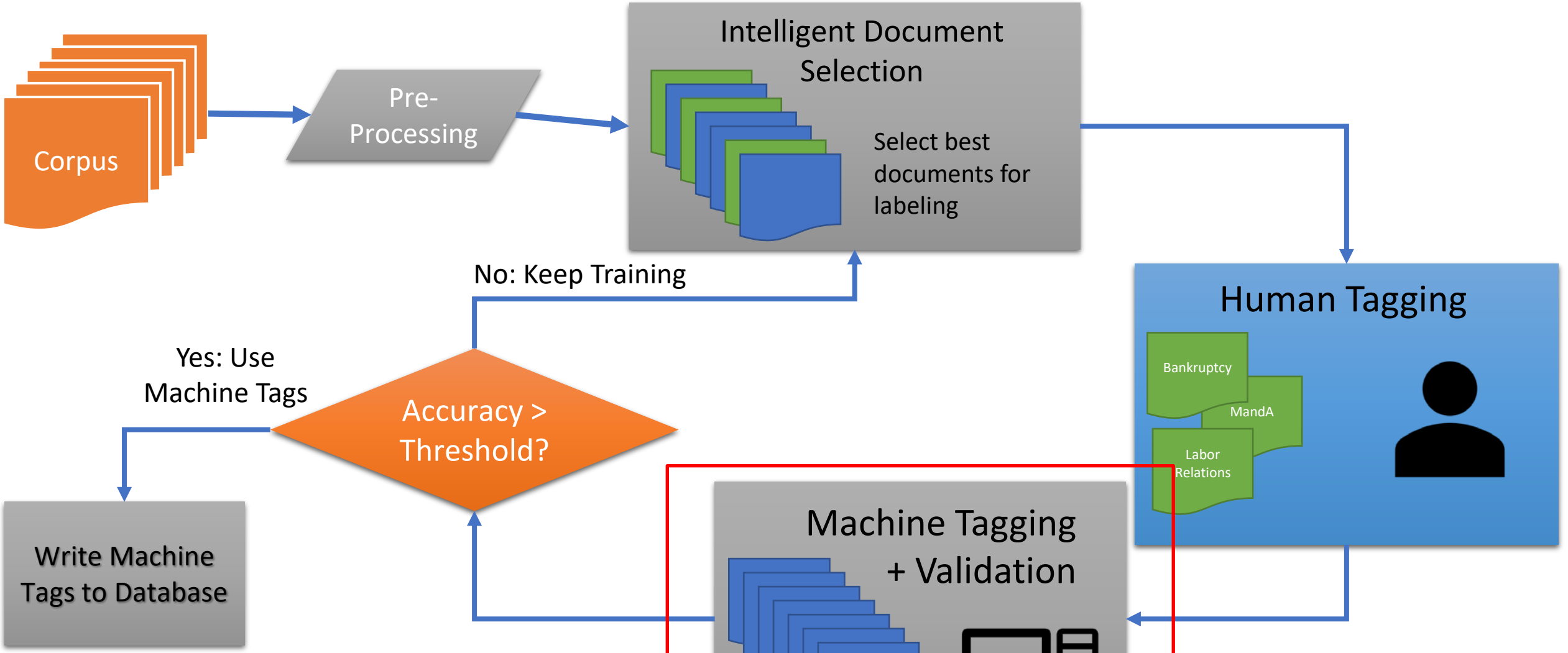
Logs each user's labels

Pick sampling domain

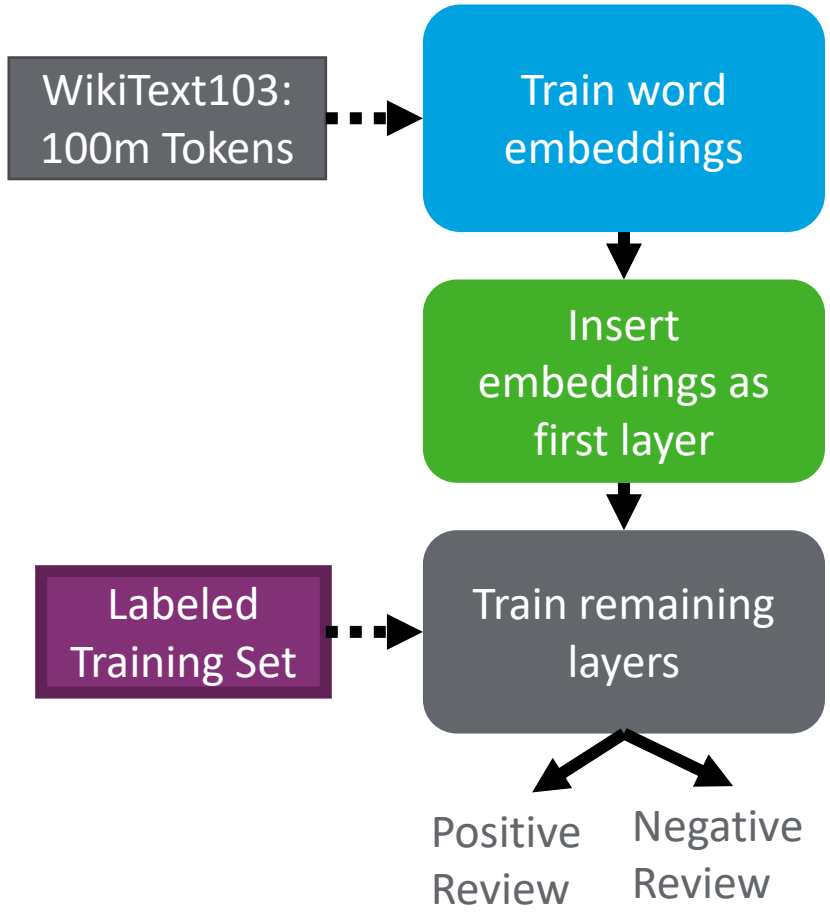
Write back to S3



Active Learning: Relevance + Classification



Transfer Learning for NLP 1.0: Pre Trained Word Vectors



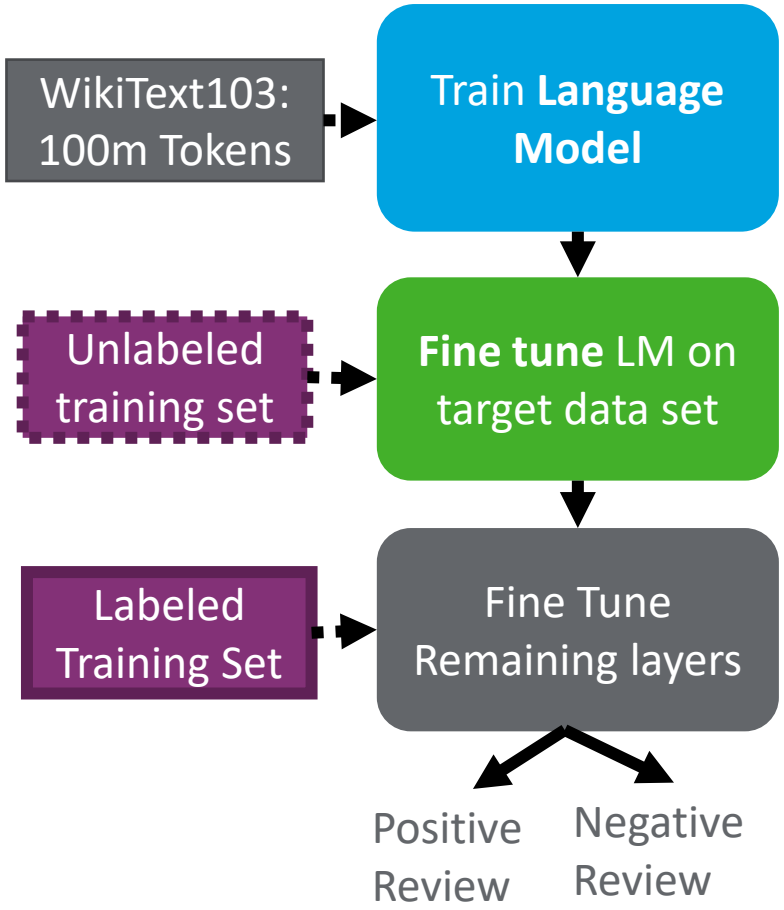
CNN + Word2Vec Issues

- Approach only bootstraps first layer of the network.
  - Loses complex structure captured in language
- Word vectors are context independent. Words with multiple meanings are captured generically:

Take it to the **bank**.
He slammed into the **bank.**
He took a **bank** shot.

Can we do better?

Transfer Learning 2.0: Language Model Pre Training



… with only 100 labeled examples, it matches the performance of training from scratch on 100x more data…

Universal Language Model Fine-tuning for Text Classification
Jeremy Howard, Sebastian Ruder

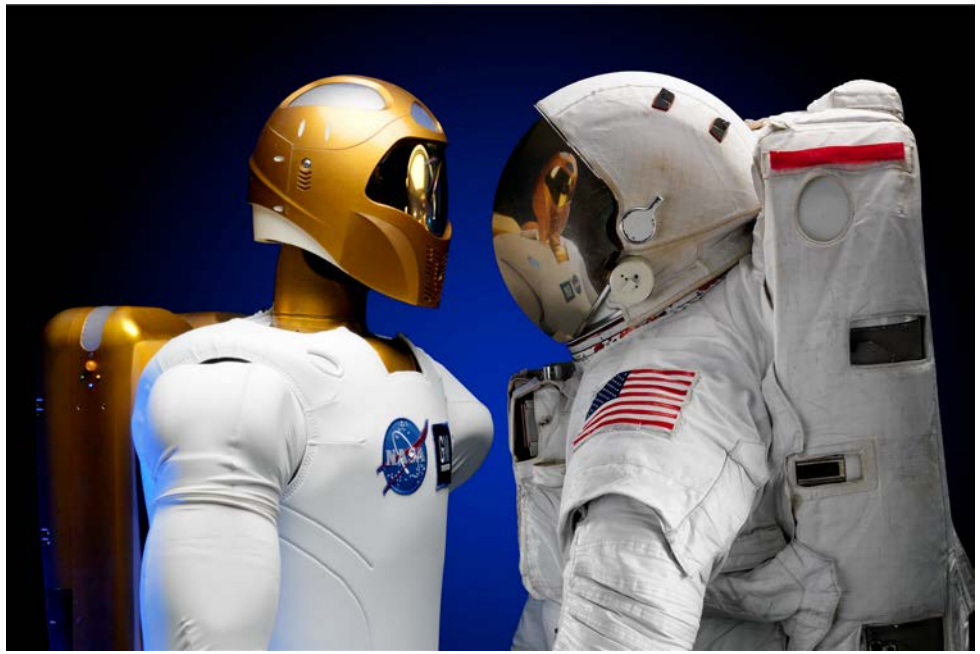
Our Experience W/ Language Modeling

- CNN Architecture is robust across many domains
  - Good place to start
  - Fast
- Consider language model pre training if:
  - Your domain is sufficiently similar to pre trained model (news, blogs, social media posts, etc) OR
  - You have sufficient in-domain documents to pre train language model
- Changes are afoot: transfer learning for NLP is here
  - ULMFit: Language Model Fine Tuning
  - ELMo: Contextualized word representations
  - BERT: Bidirectional embeddings with attention

Approach	GPU	Training Time
Doc2Vec + SVM	None	10Min
CNN + Pre Trained W2Vec	Tesla V100	10 Min
AWD-LSTM	Tesla V100	10-15 Hours

Conclusions

- There is no substitute for (at least some) ground truth!
- Break down ML problems into accomplishable steps
- Custom models are difficult to tune maintain
  - Cleaning is more likely to improve performance than tuning
  - Non deterministic
- Buy vs Build:
  - Snorkel
    - Weak supervision
  - Prodigy
- Does your training set already exist?




LOCKHEED MARTIN