Semi-Supervised Learning - Real Life Examples

## 1. Email Classification with Limited Labels

### Problem

A company wants to classify incoming emails as spam or not spam but has labeled only 1% of their emails.

### Solution

Train a model with the labeled emails and apply pseudo-labeling to include high-confidence predictions on the unlabeled ones.

### Implementation

```python  
from sklearn.semi\_supervised import SelfTrainingClassifier  
model = SelfTrainingClassifier(LogisticRegression())  
model.fit(X\_combined, y\_combined)  
```

### Impact

Reduced manual labeling cost while improving email filtering accuracy.

### Conclusion

Semi-supervised learning boosted performance without relying entirely on expensive human-annotated data.

## 2. Medical Diagnosis from X-rays with Few Expert Labels

### Problem

Radiologists label only a small set of X-ray images due to high cost and time.

### Solution

Use semi-supervised learning to learn from the limited labeled and many unlabeled X-ray images.

### Implementation

Combine CNN-based deep learning with pseudo-labeling and consistency regularization.

### Impact

Achieved similar accuracy to fully labeled datasets with only 20% labeled data.

### Conclusion

Greatly reduced labeling overhead in medical imaging with scalable accuracy.

## 3. Voice Recognition with Partial Transcripts

### Problem

Only some audio files have corresponding transcripts for training a speech-to-text model.

### Solution

Use labeled audio to train an initial model and generate pseudo-transcripts for unlabeled audio.

### Implementation

Models like DeepSpeech or Wav2Vec with semi-supervised fine-tuning.

### Impact

Expanded dataset from 10 hours labeled to 100+ hours usable data.

### Conclusion

SS learning enabled richer speech models in low-resource scenarios.

## 4. Product Categorization in E-Commerce

### Problem

Millions of products are uploaded, but only a small portion are manually categorized.

### Solution

Use labeled examples to propagate labels to similar products via label spreading.

### Implementation

```python  
from sklearn.semi\_supervised import LabelSpreading  
model = LabelSpreading()  
model.fit(X\_combined, y\_combined)  
```

### Impact

Automated product taxonomy improved search and recommendation accuracy.

### Conclusion

Efficiently expanded product classification coverage using existing data.

## 5. Fraud Detection with Incomplete Annotations

### Problem

Banks often know about only a small set of confirmed fraudulent transactions.

### Solution

Train a semi-supervised model to generalize fraud patterns from limited confirmed cases.

### Implementation

Combine anomaly detection with self-training on partially labeled financial data.

### Impact

Detected hidden fraud with fewer false positives than a purely supervised system.

### Conclusion

SS learning improved fraud detection when confirmed fraud labels were scarce.