MathWorks

MATLAB EXPO 2018

# 深度学习和时间序列分析

阮卡佳

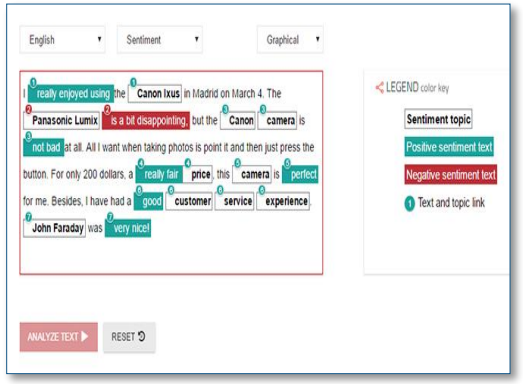
# 深度学习还是机器学习？

- 你有标签数据吗？
  - 如果没有，传统的机器学习可能是更合适
- 你了解你的数据吗？
  - 如果重要的特征提取需要专业领域知识，选择机器学习
- 你的数据规模是什么？
  - 深度学习通常需要大规模数据集
- 是否能够接受黑盒子模型？
  - 如果不能，传统的机器学习可能更合适
- 是否有 GPU 计算资源？
  - 深度学习需要大规模计算资源
- 你的模型期望精度是多少？
  - 机器学习可能会面临精度平台

# 深度学习在行动


语音识别

交易算法




情感分析


动作判断

## 我们从何起步？

- 应用类型
- 输入/输出类型
- 网络结构

# 你需要哪种应用？

- 应用类型
- 输入/输出类型
- 网络结构

A. 图片

B. 视频

C. 语音

D. 文字

E. 传感器

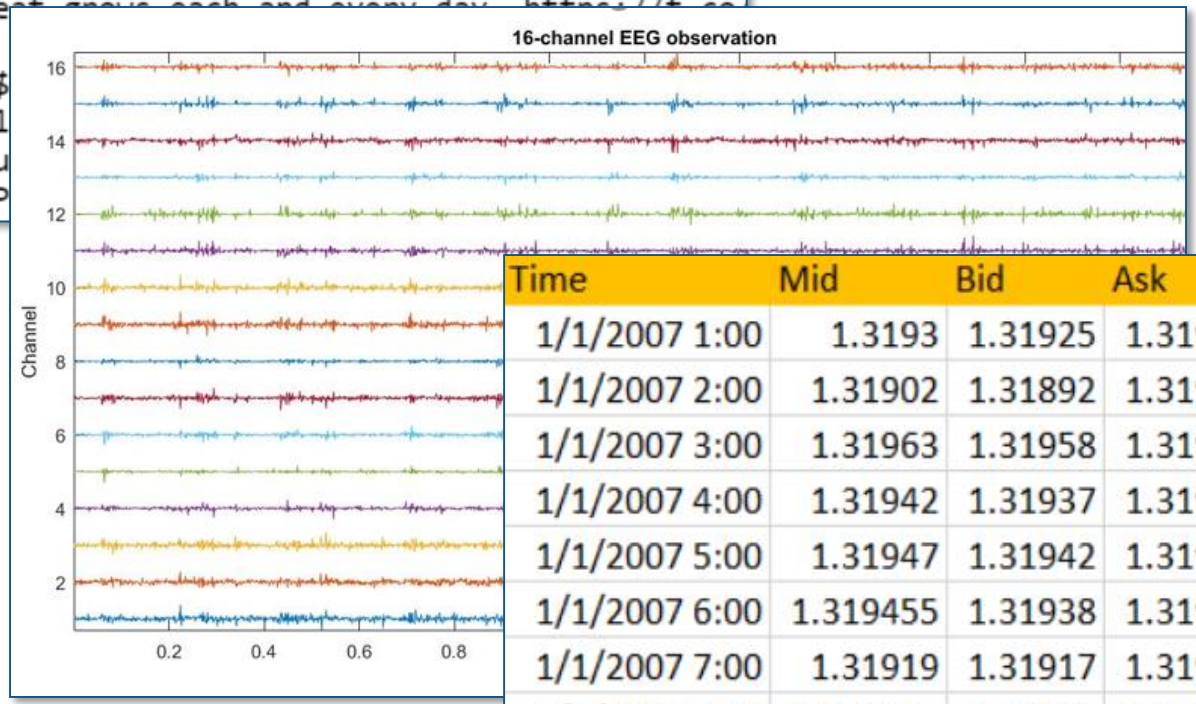
F. 以上都不是

没有图片，怎么办？

遇到序列数据或者时间序列数据，怎么办？

# 信号、文本和时间序列

```
ans = 508×1 string array
    "Walmart: "you wanna destroy Amazon?" Google: "bet" $WMT $GOOG
    "$WMT wants next level customer service w/highly personalized
    "Ironic prelude to $DIS buying $TWTR soon IMO $AAPL $GOOG $SPY
    "$AMZN the $WMT threat grows each and every day. https://t.co
    "MU Investments Co.
    "Ad $ are going to $
    "Big bullish unusual
    "REPORT: Apple to bu
    "RT @theflynews: REP
```

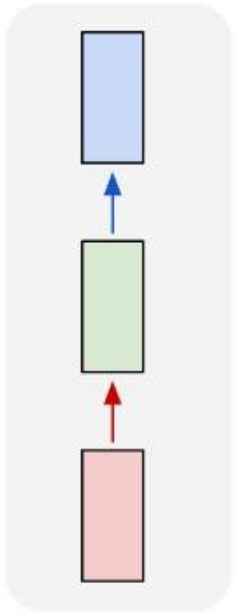



| Time | Mid | Bid | Ask | Return_5 | rsindex_5 | Action |
|---|---|---|---|---|---|---|
| 1/1/2007 1:00 | 1.3193 | 1.31925 | 1.31935 | 4.54807E-05 | 68.75 | sell |
| 1/1/2007 2:00 | 1.31902 | 1.31892 | 1.31912 | 3.79084E-05 | 55.31914894 | hold |
| 1/1/2007 3:00 | 1.31963 | 1.31958 | 1.31968 | 8.33636E-05 | 60 | sell |
| 1/1/2007 4:00 | 1.31942 | 1.31937 | 1.31947 | -9.85184E-05 | 30.3030303 | hold |
| 1/1/2007 5:00 | 1.31947 | 1.31942 | 1.31952 | 0.000144018 | 70.21276596 | sell |
| 1/1/2007 6:00 | 1.319455 | 1.31938 | 1.31953 | 8.71648E-05 | 68.25396825 | sell |
| 1/1/2007 7:00 | 1.31919 | 1.31917 | 1.31921 | -0.00012885 | 32.29166667 | hold |
| 1/1/2007 8:00 | 1.319235 | 1.31916 | 1.31931 | -3.79006E-06 | 49.20634921 | sell |
| 1/1/2007 9:00 | 1.31692 | 1.3169 | 1.31694 | 1.51872E-05 | 52.08333333 | hold |
| 1/1/2007 10:00 | 1.31702 | 1.31697 | 1.31707 | 0 | 50 | sell |

- 应用类型
- 输入/输出类型
- 网络结构

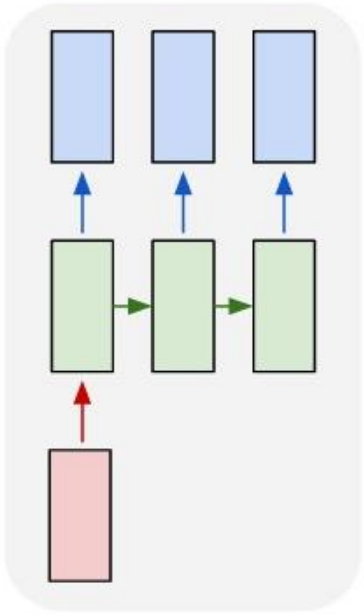
# Sequence 模型应用 ↔ 输入/输出类型

one to one



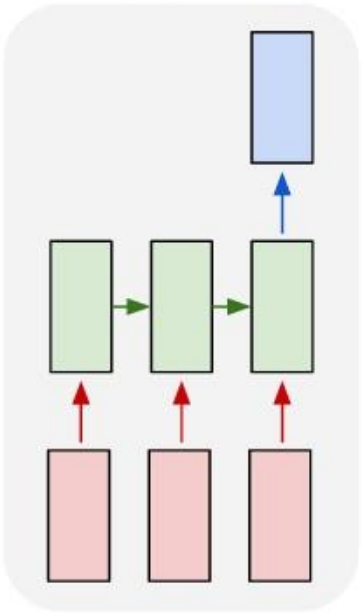
图像分类

one to sequence



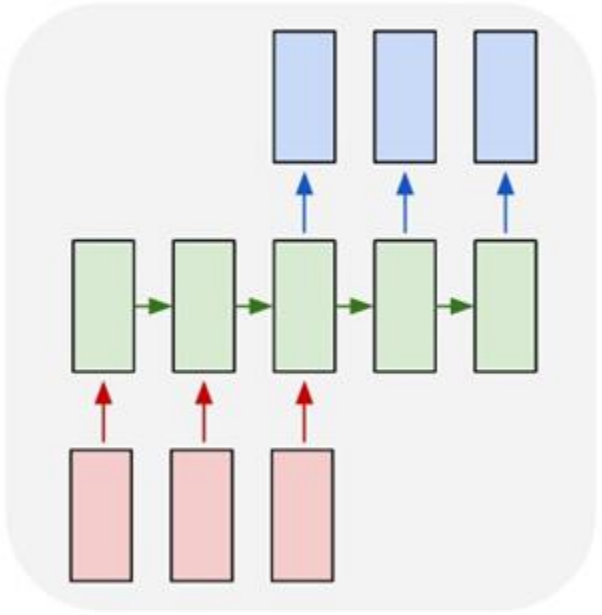
图片字幕

sequence to one



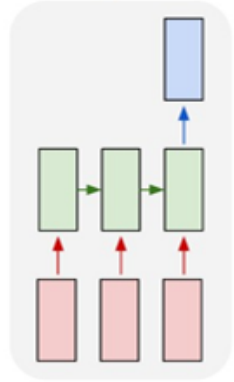
- 时间序列分类
- 时间序列回归（预测）
- 情绪分析
- 动作识别

sequence to sequence



- 语言翻译
- 自动补全（下个词组补全）
- 语音到文字翻译

# Sequence-to-One 类型



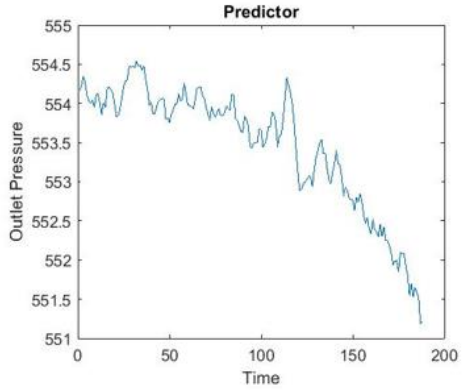
- 电影评论的情感分析（分类）

“This” “is” “the” “best” “movie” “ever” → positive

输入：单词序列（句子）

输出：标签（positve）或negative

- 发动机剩余使用时间（回归）

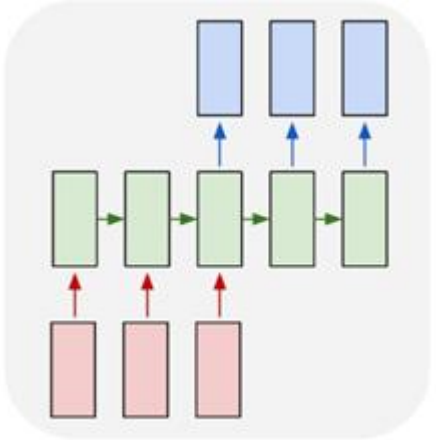



→ 130 （小时）

输入：传感器信号时间序列

输出：发动机失效的时间（标量）

# Sequence-to-Sequence 类型

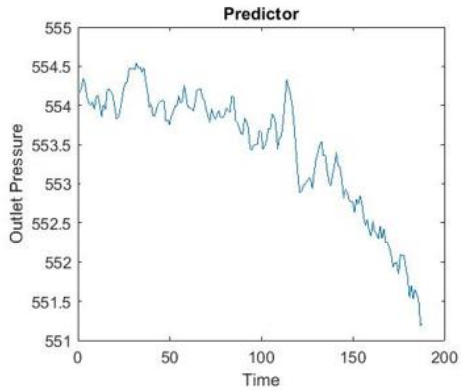


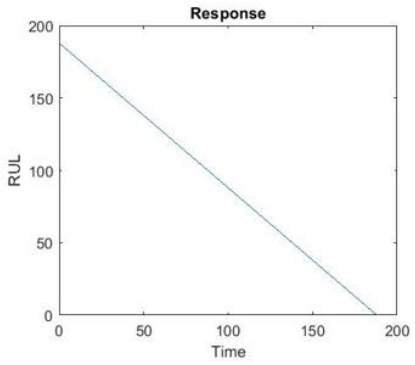
- 字母级语言建模（分类）

‘h’ ‘e’ ‘l’ ‘l’ → ‘e’ ‘l’ ‘l’ ‘o’

输入：字母序列　　输出：输入字母的下一个字母预测序列

- 发动机剩余使用时间（回归）




输入：传感器信号时间序列　　输出：每个时间点发动机失效的程度

# 我在中国长大。我说 _____ ？

# 我在中国长大……

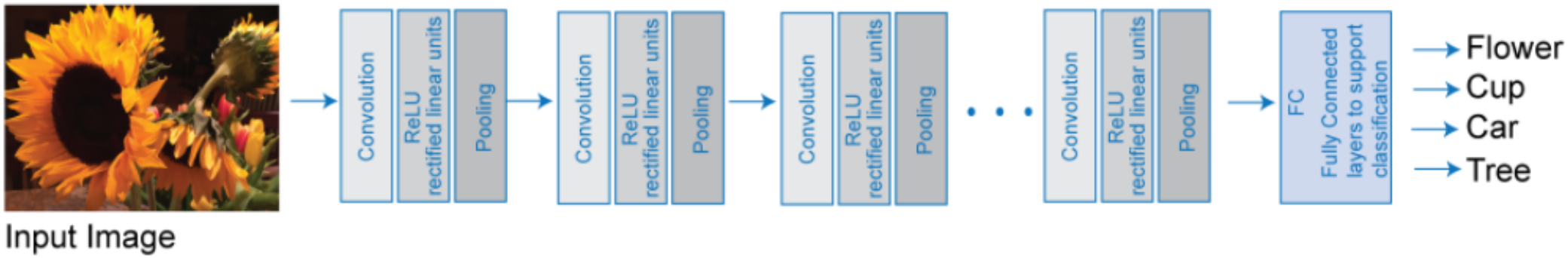
# [2000 字]

# ……我说 _____ ？

# 结构: 卷积神经网络 （CNN）

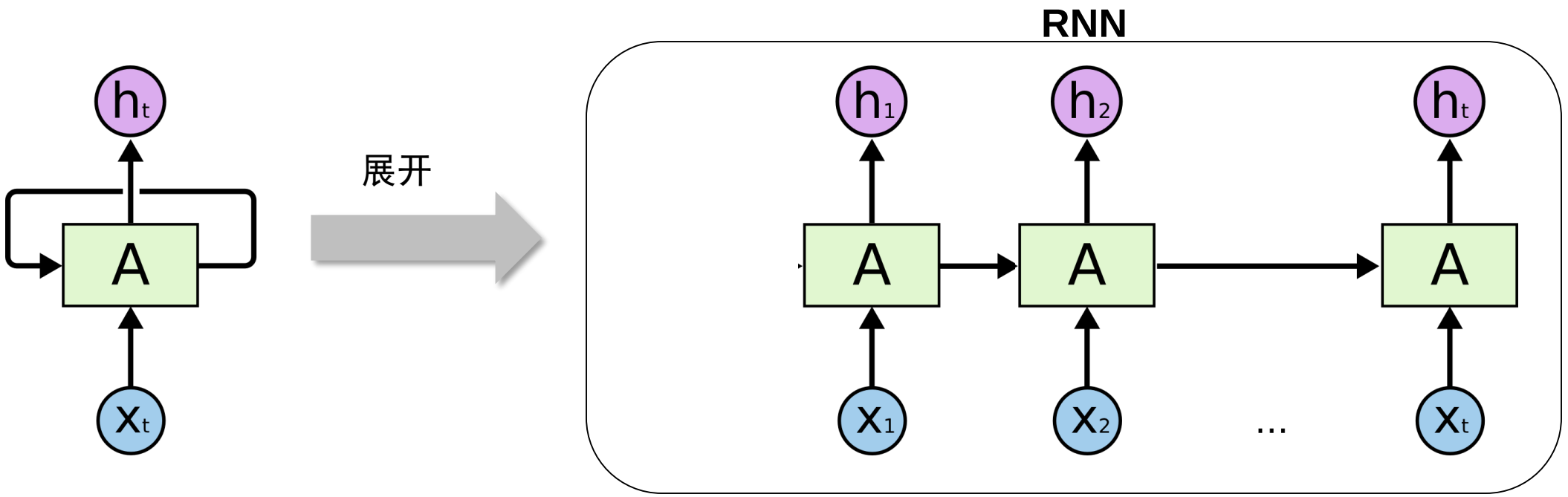
- 应用类型
- 输入/输出类型
- 网络结构

- CNN 是理想的图像和视频处理网络
- CNN 采用固定大小的输入并生成固定大小的输出
- 卷积将输入图像通过一组卷积滤波器，每一个都从图像中激活某些特征



# 结构：循环神经网络（RNN）

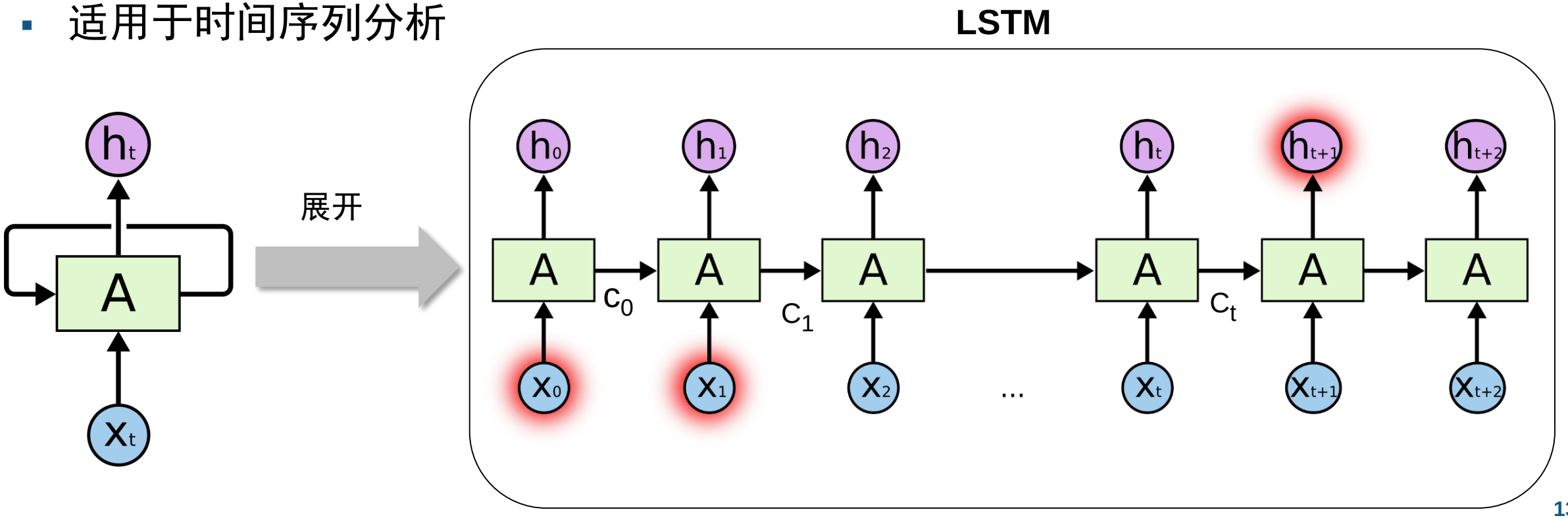
- RNN 是连接时间的神经网络
- RNN 可以处理任意输入/输出长度
- 使用输入之间的依赖关系



# 结构:长短期记忆网络（LSTM）

**lstmLayer()**

- LSTM 是递归神经网络（RNN）的延伸
- LSTM 也可以通过时间连接，它们保存了长期和短期的依赖关系
- 是理想的文本和序列数据分析网络
- 适用于时间序列分析



# 应用及深度学习模型


语音识别

sequence to sequence

**输入**: 语音信号

**输出:** 语句

**网络:** RNN/LSTM

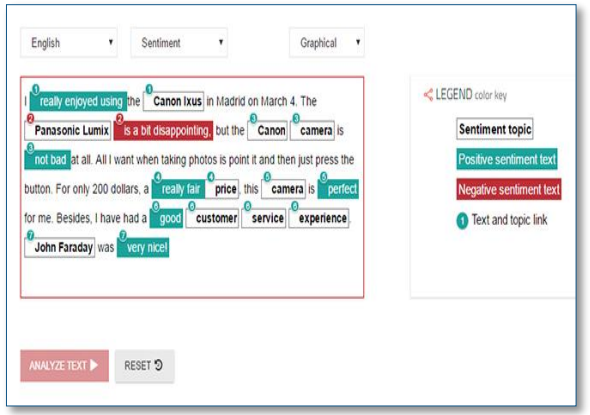
交易算法

sequence to one

**输入**: 时间序列数据

**输出: 买/卖**

**网络:** RNN/LSTM




情绪分析

sequence to one

**输入**: 文本

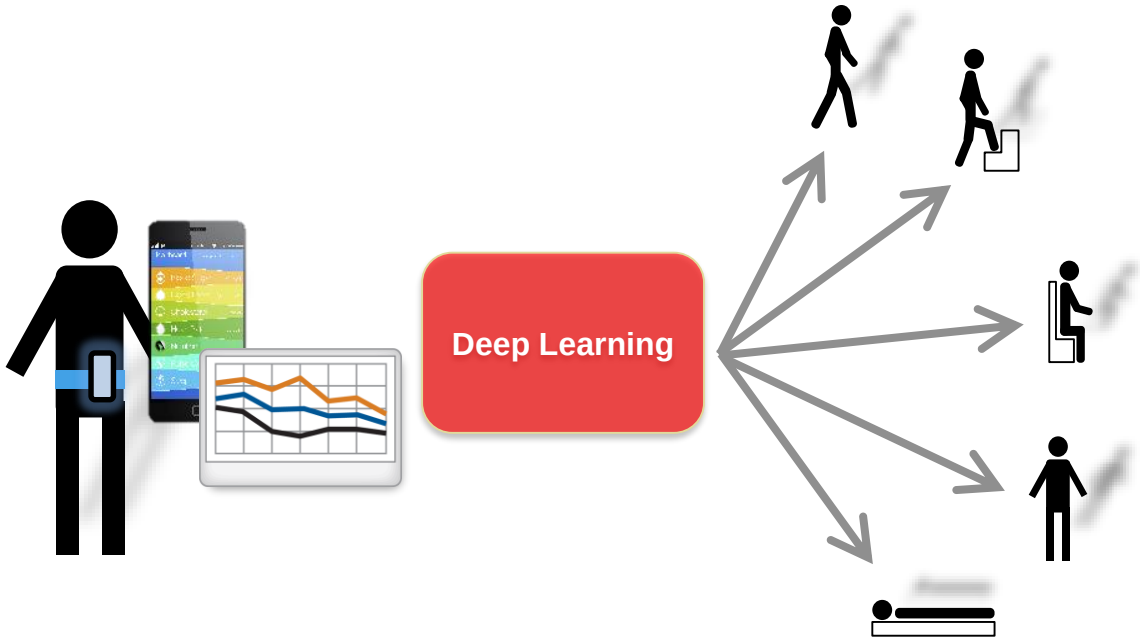
**输出: 词语**

**网络:** RNN/LSTM

# 深度学习进行时间序列分类

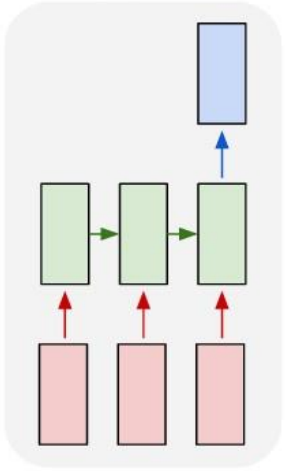
# 示例：人类动作识别（时间序列分类）

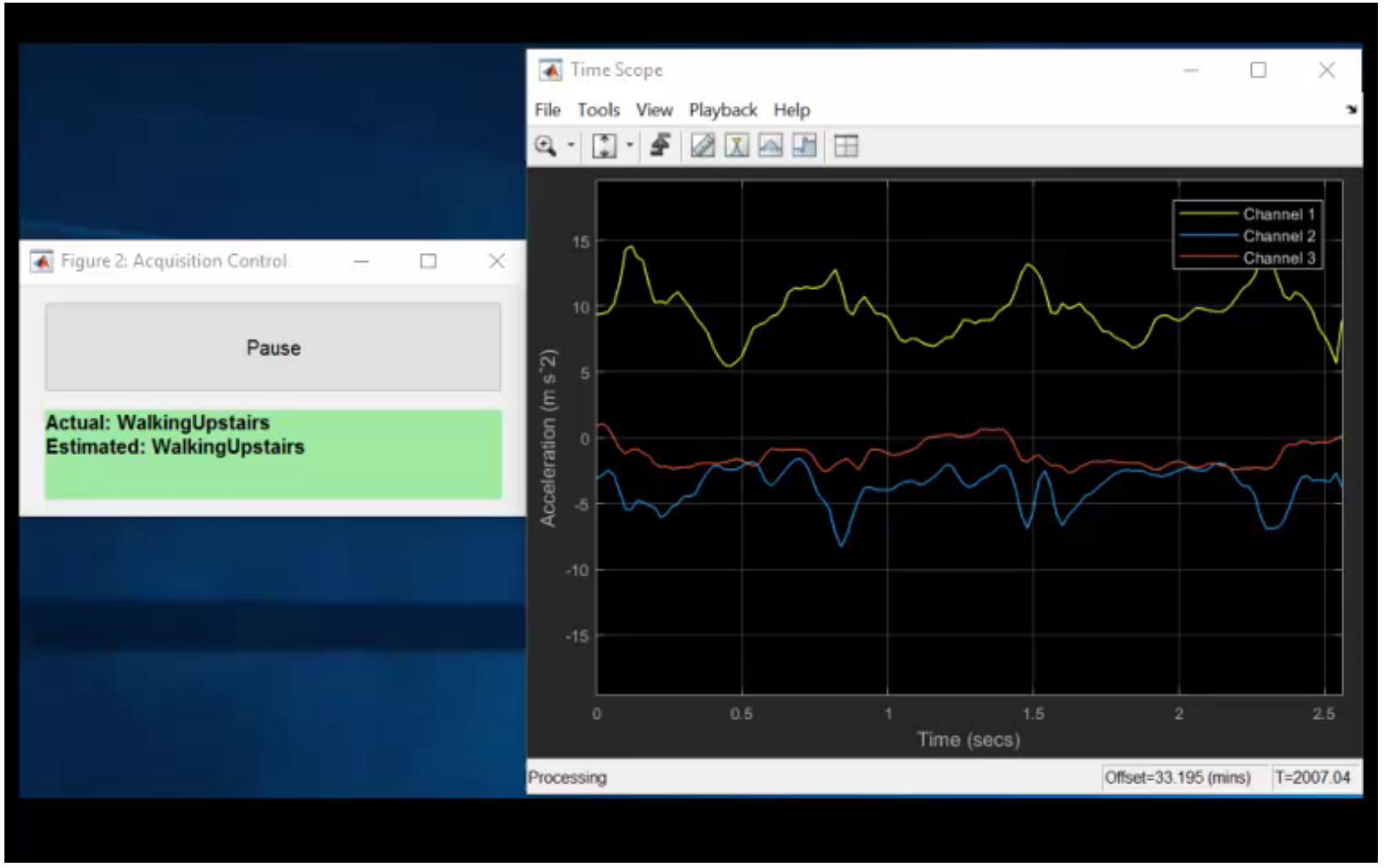
- 数据集是用智能手机捕获的加速度计和陀螺仪信号
- 数据是包含9个通道的时间序列的集合
- 进行六类行为分类：
  - Walking
  - Walking upstairs
  - Walking downstairs
  - Sitting
  - Standing
  - Laying



# 示例：人类动作识别（LSTM 网络）

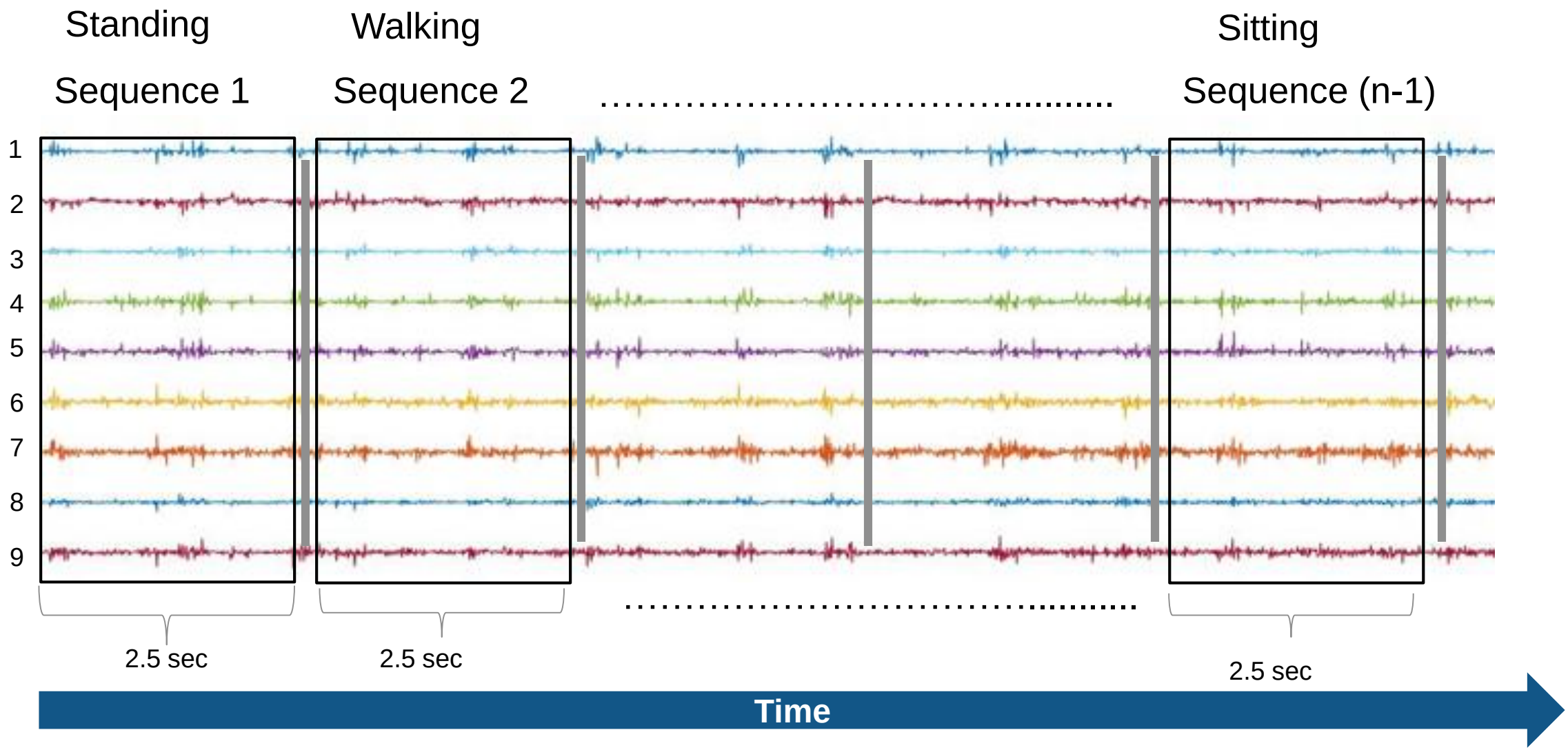
sequence to one



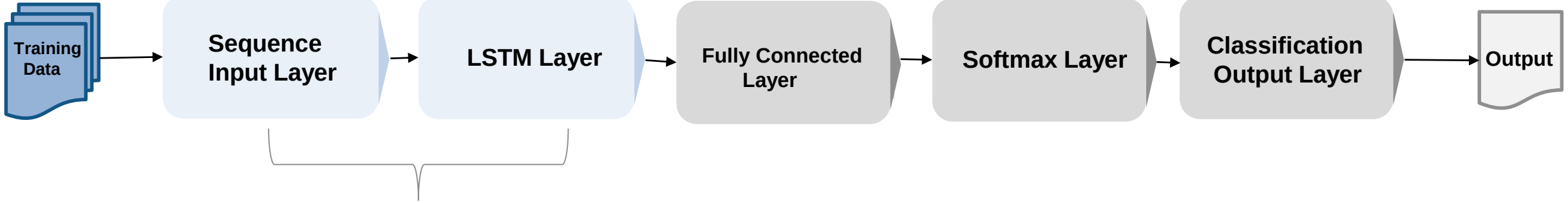



**使用 LSTM 进行信号分类**

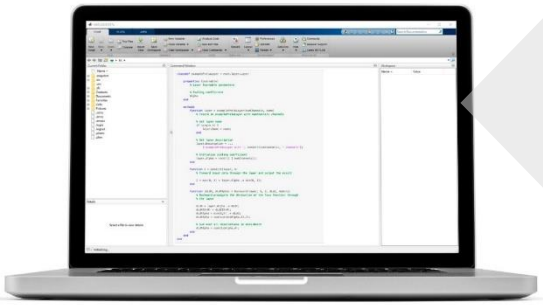
# 数据是怎样的?


Standing
Sequence 1
Walking
Sequence 2
Sitting
Sequence (n-1)
1
2
3
4
5
6
7
8
9
2.5 sec
2.5 sec
2.5 sec
Time

# 1) 创建深度学习 LSTM 网络




特征提取



```
inputSeq = 9;
no_classes = 6;

layers_default = [...
    sequenceInputLayer(inputSeq)
    lstmLayer(100, 'OutputMode', 'last')
    fullyConnectedLayer(no_classes)
    softmaxLayer
    classificationLayer];

options_default = trainingOptions('sgdm', 'Plots','training-progress', ...
    'ExecutionEnvironment','cpu')
```

**Seq-to-Seq Classification：**
'OutputMode' = 'Sequence'

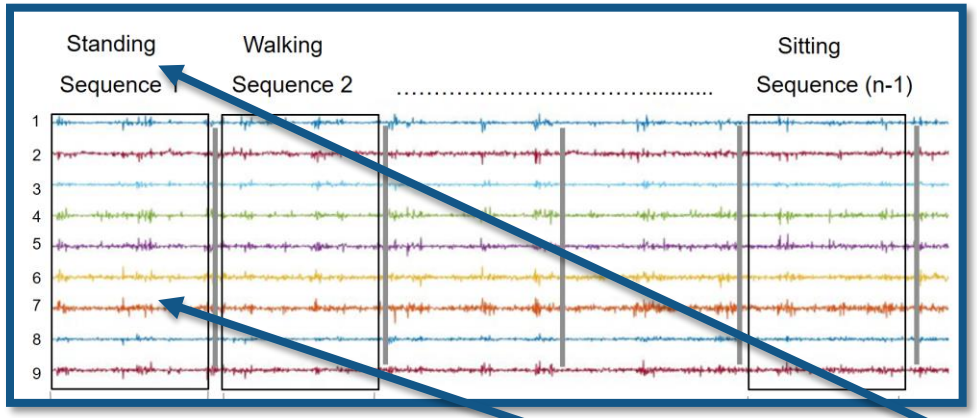
# 2) 训练参数设置

```
options = trainingOptions('sgdm',...
    'Momentum',BayesObject.XAtMinEstimatedObjective.Momentum,...
    'LearnRateSchedule','piecewise',...
    'InitialLearnRate',BayesObject.XAtMinEstimatedObjective.InitialLearnRate,...
    'LearnRateDropFactor',BayesObject.XAtMinEstimatedObjective.LearnRateDropFactor,...
    'LearnRateDropPeriod',BayesObject.XAtMinEstimatedObjective.LearnRateDropPeriod,...
    'MaxEpochs', BayesObject.XAtMinEstimatedObjective.MaxEpochs,...
    'MiniBatchSize',BayesObject.XAtMinEstimatedObjective.MiniBatchSize,...
    'L2Regularization', BayesObject.XAtMinEstimatedObjective.L2Regularization ,...
    'Plots','training-progress', ...
    'shuffle', 'once', ...
    'ExecutionEnvironment','cpu',...
    'OutputFcn',@(info)stopIfAccuracyNotImproving(info,150))
```

- 初始学习速率
  - 使用默认设置，然后进行超参数调整
- 使用贝叶斯优化 bayesopt 进行训练选项的超参数调整
  - Solver 名字, Mini Batch 大小, Max Epochs, Verbosity 等

# 3) 训练网络

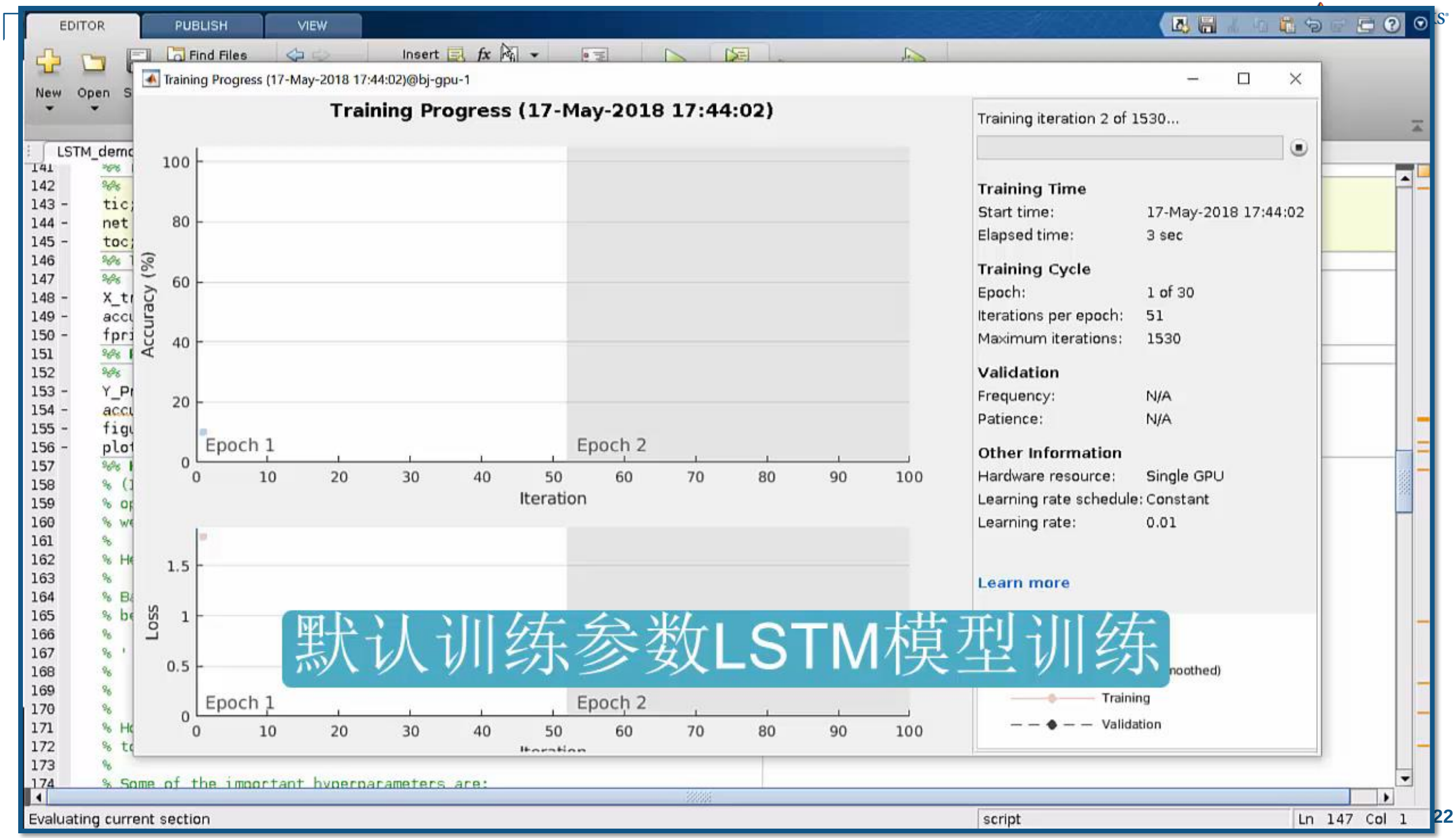
Bayesian Optimization




```
options = trainingOptions('sgdm',...
    'Plots','training-progress',...
    'InitialLearnRate',0.8,...
    'Verbose',false,...
    'MaxEpochs',40,...
    'MiniBatchSize',1000);
```

```
net = trainNetwork(Xtrain, Ytrain, layers, options);
```

```
% Dimension parameters
inputChannels = 9;
hiddenSize = 100;
numClasses = 6;

% Define layer architecture
layers = [ ...
    sequenceInputLayer(inputChannels)
    lstmLayer(hiddenSize)
    lstmLayer(hiddenSize, 'OutputMode', 'last')
    fullyConnectedLayer(numClasses)
    softmaxLayer()
    classificationLayer() ];
```


EDITOR
PUBLISH
VIEW
Training Progress (17-May-2018 17:44:02)@bj-gpu-1
Training Progress (17-May-2018 17:44:02)
Accuracy (%)
Iteration
Epoch 1
Epoch 2
Loss
Training iteration 2 of 1530...
Training Time
Start time: 17-May-2018 17:44:02
Elapsed time: 3 sec
Training Cycle
Epoch: 1 of 30
Iterations per epoch: 51
Maximum iterations: 1530
Validation
Frequency: N/A
Patience: N/A
Other Information
Hardware resource: Single GPU
Learning rate schedule: Constant
Learning rate: 0.01
Learn more
Training
Validation
默认训练参数LSTM模型训练
Evaluating current section
script
Ln 147 Col 1

# 关键要点

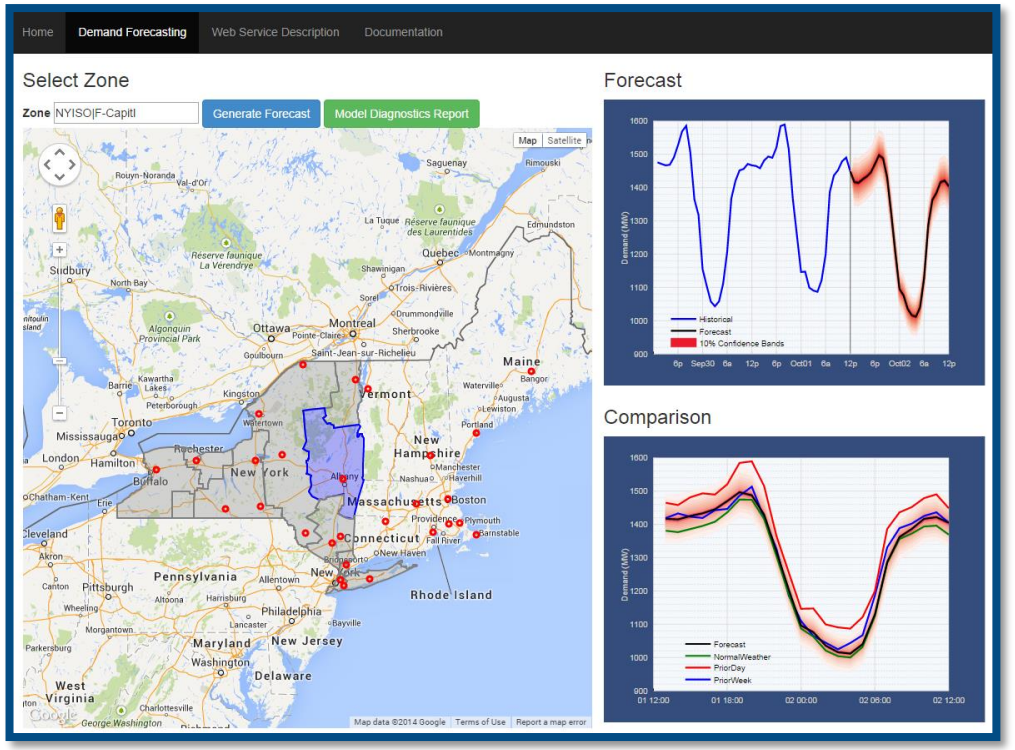
- 相对于机器学习而言，在获得相当的准确度的同时，所需努力要少得多
- 绕过特征提取阶段
- 超参数优化利用贝叶斯优化、进行超参数调制，提高了模型的精度

| | 算法 | 数据预处理工作 | 特征提取 | 是否需要超参数优化 | 精度 | 技能要求 | 总工作量 |
|---|---|---|---|---|---|---|---|
| 机器学习 | SVM (Quadratic) | High | Yes | Yes | 90.1 % | High | High |
| 深度学习 | LSTM | Low | No | Yes | 90.2 % | High | Medium |

# 深度学习进行时间序列预测

# 示例: LSTM 时间序列预测

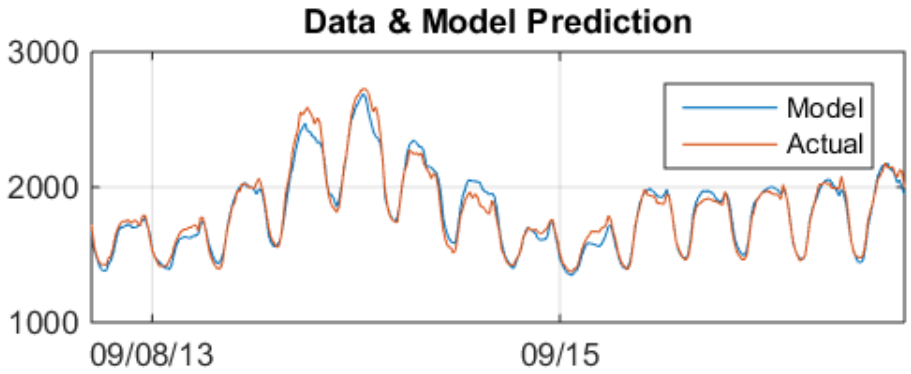
- 从历史的数值，预测下一个数值
- ‘**adam**’ 求解器：adaptive moment estimation
- 使用 **predictAndUpdateState** 函数预测时间步长，并在每次预测时更新网络状态。



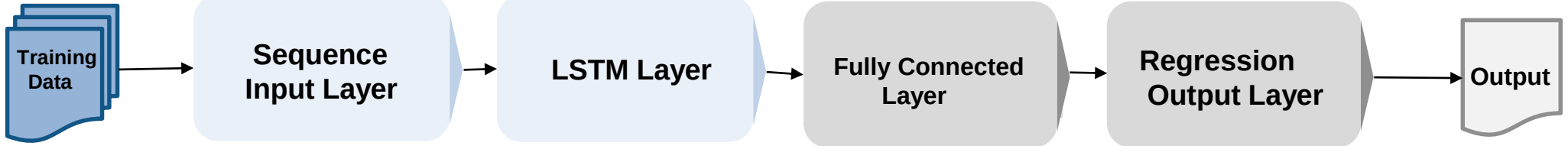

Predictors

Response

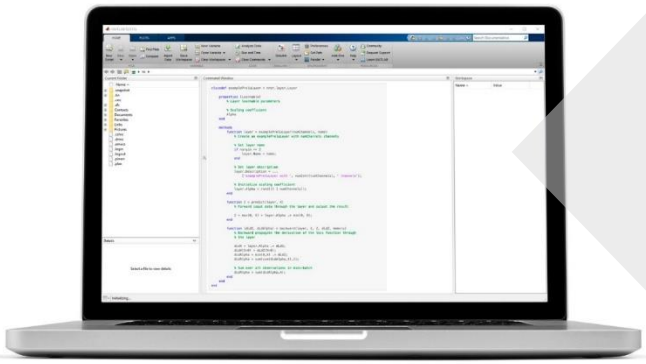
$$X_1, X_2, X_3, \ldots, X_T \rightarrow X_2, X_3, X_4, \ldots, X_{T+1}$$



# 时间序列回归




特征提取



```
inputSize = 12;
outputSize = 125;
numResponses = 1;
layers = [ ...
    sequenceInputLayer(inputSize)
    lstmLayer(outputSize,'OutputMode','sequence')
    fullyConnectedLayer(numResponses)
    regressionLayer];

options = trainingOptions('adam', ...
    'MaxEpochs',maxEpochs, ...
```

**Seq-to-Seq Regression：**
‘OutputMode’ = ‘sequence'

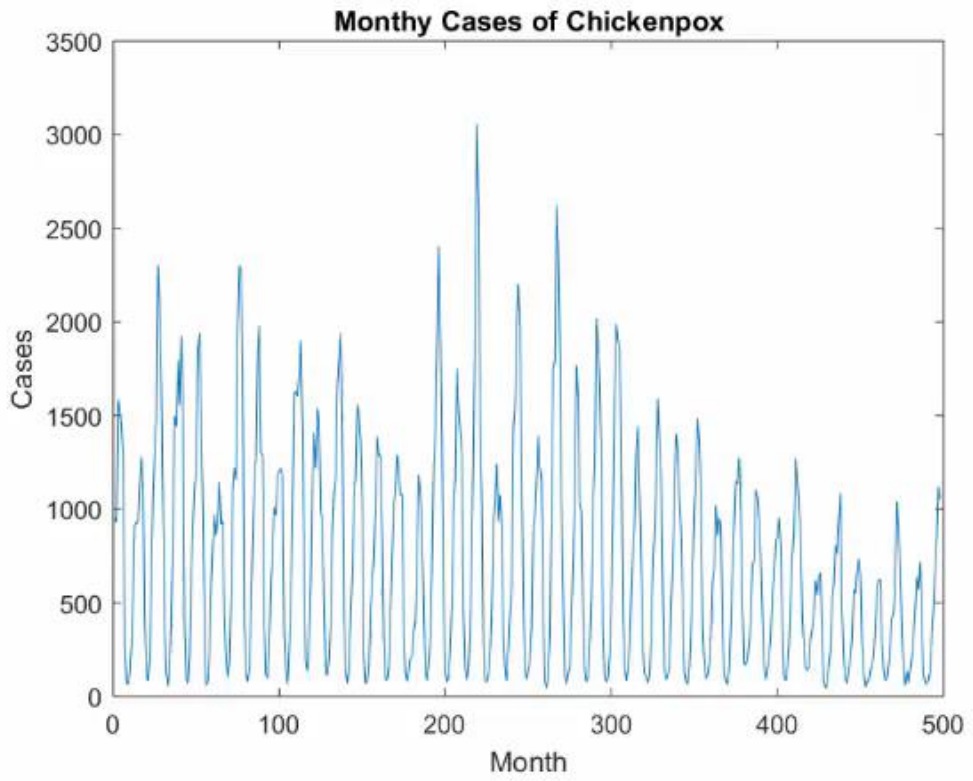
LIVE EDITOR | INSERT | VIEW

Single | Left/Right | Top/Bottom | Custom | Tabs Position | Shrink Tabs to Fit | Alphabetize | Line Numbers | Datatips | Full Screen | Clear all Output | Output Inline | Output on Right

TILES | DOCUMENT TABS | DISPLAY | OUTPUT | LAYOUT

TimeSeriesForecastingUsingDeepLearningExample.mlx *

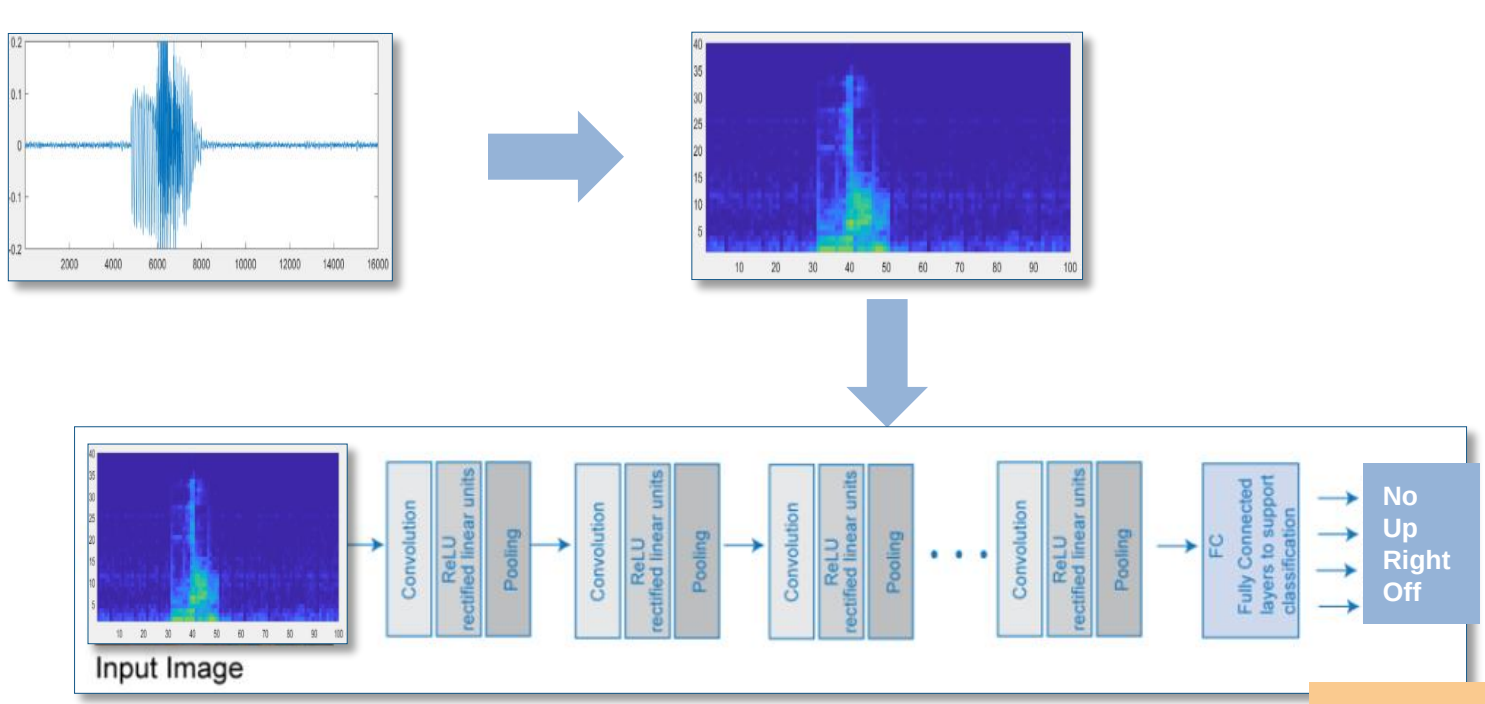
```
7  ylabel("Cases")
8  title("Monthy Cases of Chickenpox")
```

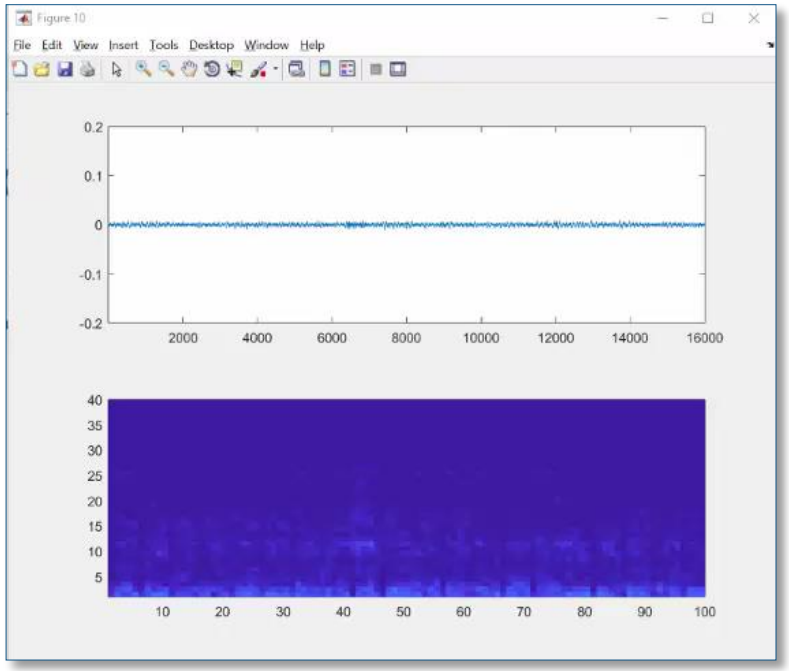


# 非图像应用的深度学习

# 语音识别

- 将时间序列转换为频谱图并将其用作CNN的输入




使用 CNN 进行语音识别

# 使用深度学习对文本数据进行分类

R2017b

- 文本数据的来源
  - 维护日志
  - 新闻/社交媒体
  - 客户调查
  - 现场报告
- 深度学习和并行处理
  - fastText 词嵌入预训练网络
  - 使用 LSTM 网络进行文本分类
  - 并行计算字频率

**Split Text into Individual Words**

```
documents = tokenizedDocument(repairNotes)
```

**Create a Bag-of-Words Model**

```
bag = bagOfWords(documents)
```

**Fit a Topic Model with 4 Topics**

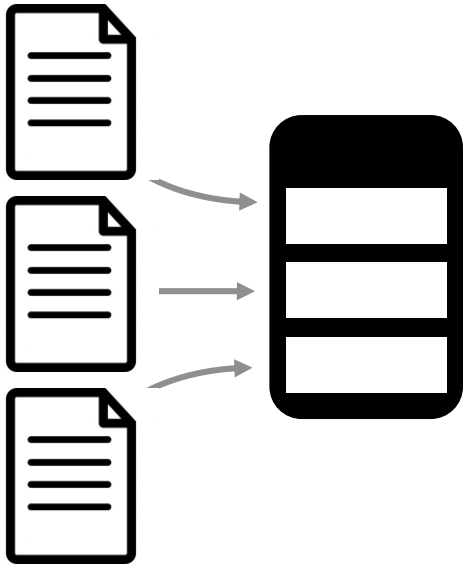
```
numTopics = 4;
mdl = fitlda(bag,numTopics)
```

```
  (36,1) coolant leak spinner light outwip
  (37,1) strob lights not workinghyd leak

bag =
bagOfWords with 913 words and 617 documents

    preventative    maintenance    service
               1              1          1
               0              0          0
               …

mdl =
  ldaModel with properties:

            NumTopics: 4
    TopicConcentration: 1.3262
    TopicProbabilities: [4×1 double]
     WordConcentration: 1
            Vocabulary: [1×913 string]
```

# Text Analytics Toolbox

数据获取与探索 → 数据预处理 → Develop Classification Models

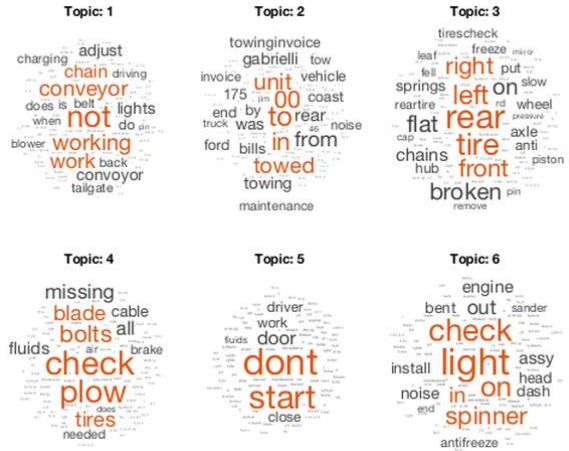
文本清理 → Convert to Numeric



Media reported two trees blown down along I-40 in the Old Fort area.

media report two tree blown down i40 old fort area

| | cat | dog | run | two |
|---|---|---|---|---|
| doc1 | 1 | 0 | 1 | 0 |
| doc2 | 1 | 1 | 0 | 1 |




- Word 文档
- PDF’s
- Text 文件
- HTML

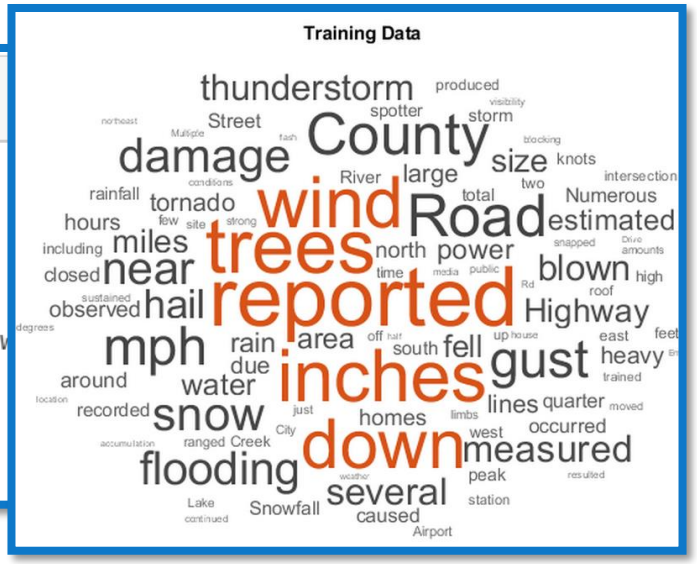
- 语言停止
- 词干提取
- 词语切分

- Bag of Words
- TF-IDF
- Word Embeddings

- LSTM

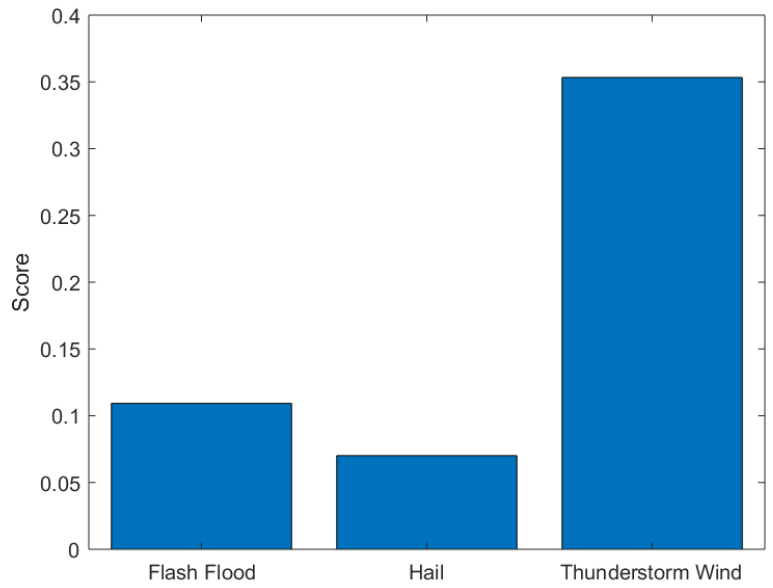
# 文本数据分类示例

```
documentsTrain(1:5)

ans =
  5×1 tokenizedDocument:

(1,1)  7 tokens: large tree down between plantersville and nettleton
(2,1) 37 tokens: one to two feet of deep standing water developed on a street on the w
(3,1) 13 tokens: nws columbia relayed a report of trees blown down along tom hall st
(4,1) 13 tokens: media reported two trees blown down along i40 in the old fort area
(5,1) 14 tokens: a few tree limbs greater than 6 inches down on hwy 18 in roseland
```




LSTM

```
net = trainNetwork(XTrain,YTrain,layers,options);

YPred = classify(net,XTest);
```



# 关键要点

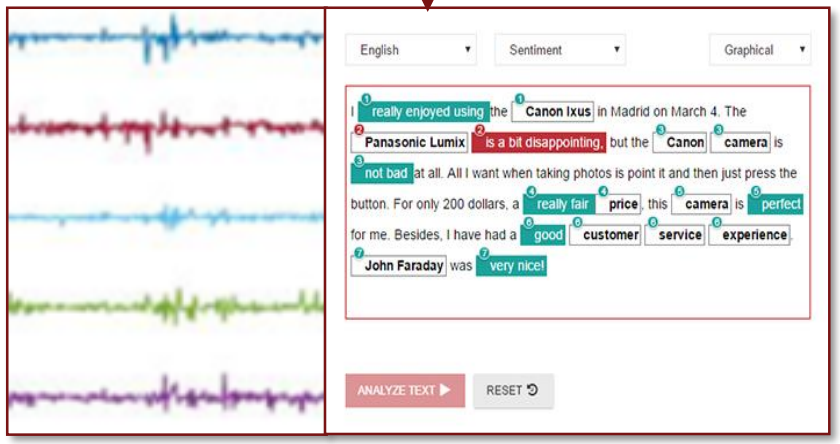
## 数据集类型

### 数值数据

| ID | WC_TA | RE_TA | EBIT_TA | MVE_BVTD | S_TA | Industry | Rating |
|---|---|---|---|---|---|---|---|
| 62394 | 0.013 | 0.104 | 0.036 | 0.447 | 0.142 | 3 | BB |
| 48608 | 0.232 | 0.335 | 0.062 | 1.969 | 0.281 | 8 | A |
| 42444 | 0.311 | 0.367 | 0.074 | 1.935 | 0.366 | 1 | A |
| 48631 | 0.194 | 0.263 | 0.062 | 1.017 | 0.228 | 4 | BBB |
| 43768 | 0.121 | 0.413 | 0.057 | 3.647 | 0.466 | 12 | AAA |
| 39255 | -0.117 | -0.799 | 0.01 | 0.179 | 0.082 | 4 | CCC |
| 62236 | 0.087 | 0.158 | 0.049 | 0.816 | 0.324 | 2 | BBB |
| 39354 | 0.005 | 0.181 | 0.034 | 2.597 | 0.388 | 7 | AA |
| 40326 | 0.47 | 0.752 | 0.07 | 11.596 | 1.12 | 8 | AAA |
| 51681 | 0.11 | 0.337 | 0.045 | 3.835 | 0.812 | 4 | AAA |

机器学习 or LSTM

### 时间序列、文本数据




LSTM or CNN

### 图像数据

CNN

# 谢谢